



Melioration learning in games with constant and frequency-dependent pay-offs

Thomas Brenner, Ulrich Witt*

*Max-Planck-Institute for Research into Economic Systems, Evolutionary Economics Unit,
Kahlaische Strasse 10, D-07745 Jena, Germany*

Received 6 March 2001; accepted 11 September 2001

The paper is dedicated to the commemoration of R.J. Herrnstein, a great proponent of interdisciplinary research. The authors would like to thank M. Erlei, J. Irving-Lessmann, R. Joosten, R. Selten, and G. von Wangenheim for helpful comments on earlier drafts.

Abstract

The paper explores the implications of melioration learning—an empirically significant variant of reinforcement learning—for game theory. We show that in games with invariable pay-offs melioration learning converges to Nash equilibria in a way similar to the replicator dynamics. Since melioration learning is known to deviate from optimizing behavior when an action's rewards decrease with increasing relative frequency of that action, we also investigate an example of a game with frequency-dependent pay-offs. Interactive melioration learning is then still appropriately described by the replicator dynamics, but it indeed deviates from rational choice behavior in such a game. © 2002 Elsevier Science B.V. All rights reserved.

JEL classification: C72; D62; D83; Q20

Keywords: Learning; Melioration; Reinforcement learning; Matching law; Replicator dynamics; Evolutionary game theory; Games with variable pay-offs; Social traps; Littering game

1. Introduction

In many cases individual learning in the economy takes place simultaneously and interactively so that the analytical framework of repeated games may be used in which players adapt their strategies to the private experience they make. It is not clear, however, to what extent the players fully recognize the interaction dynamics and, thus, which behavioral assumptions are appropriate for the analysis of the repeated game. Doubts may be raised

* Corresponding author. Tel.: +49-3641-686820/686801; fax: +49-3641-686868.
E-mail addresses: witt@mpiew-jena.mpg.de, brenner@mpiew-jena.mpg.de (U. Witt).

as to whether the strong version of rational behavioral adaptation represented by Bayesian learning is empirically relevant here (Selten, 1991). Other concepts have therefore been considered, in particular, concepts derived from evolutionary game theory, which were originally conceived for explaining the evolution of genetically fixed behavioral traits in socially interacting populations under the influence of natural selection (see e.g. Samuelson, 1993; Binmore et al., 1995; Weibull, 1995; Vega-Redondo, 1996). However, unless the dynamics of learning can be shown to be equivalent to the dynamics of genetic adaptation, the relevance of these concepts for the economic context, where most behavior adaptation that occurs is due to learning, is not clear either.

If one looks at the dynamics of learning in non-strategic situations (with low or no cognitive participation), a large number of theories to explain behavior have been proposed. The one best confirmed in innumerable experiments both with humans and animals (cf. Hergenhahn and Olson, 1997), is probably the theory of reinforcement learning. Its explanatory power decreases somehow if the subjects recognize the frequency-dependence of their pay-offs (Herrnstein et al., 1993). Roth and Erev (1995), Erev and Roth (1998), nonetheless, found that reinforcement learning describes human behavior in various interaction experiments very well. To investigate implications of reinforcement learning—as we do in the present paper—may therefore be considered of interest in its own.

The theory of reinforcement learning comes in two different variants, the ‘stimulus sampling’ approach (Bush and Mosteller, 1955; Estes and Suppes, 1974) and the ‘meliorating behavior’ approach of Herrnstein (Herrnstein and Loveland, 1975; Vaughan and Herrnstein, 1987; Herrnstein and Prelec, 1991; see also Loewenstein and Prelec, 1992). In both cases, the probabilities assigned to the possible behaviors change in response to experienced reinforcement, i.e. the outcome of the chosen actions. What is different in the two variants is the specification of the probability updating. While Bush and Mosteller (1955) consider only the outcome of the last action, the melioration hypothesis focuses on the average utility realized in the last few choices of the alternative actions. Vaughan and Herrnstein (1987) assume that the relative frequencies of choosing alternatives are then adjusted in such a way that, eventually, their average utilities (if chosen at all) are equal, a condition known in the psychological literature as the “matching law” (Herrnstein, 1970, not to be confused with ‘probability matching’ discussed by Simon, 1957).

In recent efforts, the Bush–Mosteller model of learning has been adapted to a game-theoretic setting and could indeed be shown to imply dynamic patterns equivalent to those of genetic adaptation (Börgers and Sarin, 1997 and Brenner, 1997). Other adaptive learning models, similar in some respects to melioration learning, have also been studied in recent years (Ellison, 1993; Young, 1993; Samuelson, 1994; Sanchirico, 1996; Dawid, 1997, and Offerman and Sonnemans, 1998). Yet, a systematic investigation of the melioration learning hypothesis in a game-theoretic setting is still missing. It will be developed in the present paper as follows. In Section 2 we specify the melioration hypothesis more closely and apply it to learning in a 2×2 game. In Section 3, we explore the convergence properties of the interactive learning process with Nash equilibria as a benchmark for rational behavior and the ‘replicator dynamics’ as the abstract characterization of genetic adaptation (Hofbauer and Sigmund, 1988, Chapter 24). With Section 4 we turn to an extension which is motivated by a remarkable implication of behavior following the matching law: its

dramatic deviation from utility maximization in “distributed choice” settings, i.e. in the case of average utilities decreasing with the relative frequency with which a strategy is chosen (Herrnstein and Prelec, 1991). In the game-theoretic framework this suggests considering a class of games in which pay-offs depend on the relative frequencies with which strategies have previously been chosen. We, therefore, introduce the “littering game” as an example and analyze the implications of meliorating behavior in that game. Section 5 offers the conclusions.

2. Melioration learning in a 2×2 game set-up

Consider those low profile economic decision problems that occur repeatedly in everyday life where the decision maker takes little or no effort to reflect on the alternative to choose in each single case. Take, as an example, expenditures on food items and their time distribution or the choice of leisure time activities and their time distribution. In these frequently recurring situations people tend to pick one or another alternative, switching back and forth between the a alternatives they have.¹ Hence, in such a setting there is a probability $p(i, t)$, $i = 1, 2, \dots, a$, for each possible stage game action i to be chosen by the agent at time t . The initial values $p(i, 1)$ may be arbitrarily chosen. As more specific information is lacking, they may be assumed equally distributed. Learning then means that the probability distribution over the stage game actions may change systematically in response to the experienced consequences (pay-offs) of the alternative actions. At any time t during the learning process the condition

$$\sum_{i=1}^a p(i, t) = 1 \quad (1)$$

must be satisfied. In the case of only two possible stage game actions, i.e. $a = 2$, this means that the value of $p(2, t)$ can be deduced from $p(1, t)$ and, for convenience, we can write $p(t) = p(1, t)$.

In their inspection of learning in a two-actions case Vaughan and Herrnstein (1987) consider an isolated individual. For our purposes we have to adapt Vaughan and Herrnstein’s time continuous adjustment formalism for updating the probability distribution to discrete time (T denotes the time span from one choice to the next). Focusing, as they do, on the difference between the average outcomes of the alternative choices we obtain

$$p(t + T) = p(t) + F(\Phi(1, t) - \Phi(2, t)) \quad (2)$$

where $\Phi(i, t)$ is the average outcome (pay-off) which the individual experiences after choosing action i at time t . F is a strictly monotone increasing function that passes through the origin. In order to ensure that $p(t) \in [0, 1]$ for all t , the dynamics given by Eq. (2) must somehow be constrained to that interval. One way of doing this—which will be pursued here—is to multiply the second term on the RHS of Eq. (2) by the factor $p(t)(1 - p(t))$. This implies a speed of learning converging to zero when $p(t)$ reaches 0 or 1 and, in addition with

¹ Herrnstein and Prelec (1991) call this ‘distributed choice’ behavior and consider low or no cognitive participation an important prerequisite of melioration learning.

Assumption 4, $p(t) \in [0, 1]$ for all t . We, thus, transform Eq. (2) and represent Herrnstein's melioration learning hypothesis in the form of

Hypothesis 1. In the two-actions case, melioration learning is a change of the probabilities of action as given by

$$p(t + T) = p(t) + p(t)(1 - p(t))F(\Phi(1, t) - \Phi(2, t)). \quad (3)$$

Remark. Eq. (3) is a specification of the general definition of learning as a probability distribution over actions changing systematically in response to experienced consequences.

Since we are interested here in interactive learning, we have to derive the implications of Hypothesis 1 (Eq. (3)) from a game-theoretic setting. To this end, we assume the conventional set up of a repeated 2×2 game:

Assumption 1. A symmetric game is played repeatedly. In each single stage game players are drawn at random from a large population and are matched in pairs. Without recognizing earlier opponents they simultaneously choose one of two possible actions.

Note that under Assumption 1, Eq. (3) describes a stochastic process. Although for each individual the state of learning is well defined at each time by Eq. (3), it depends on the pay-offs obtained in former interactions which are a result of random matching and random choices. This means that $\Phi(1, t)$ and $\Phi(2, t)$ are stochastic variables.

Since melioration learning is supposed to occur in choice situations not well-reflected on by the agents, it appears appropriate for the game-theoretic context to assume that the players do not recognize each single pay-off they themselves obtain or those of their opponents. Likewise, they can hardly be assumed to consider all the strategies their opponents might choose and their respective implications. Hence, the learning dynamics depend exclusively on the averages of previously realized own pay-offs as far as they enter the formation of the averages. (Realized pay-offs as well as their averages are, of course, determined by the strategies which the opponents happened to play in the past and by the stage game pay-off matrix denoted by $\Pi(i, j)$). For the formation of the average pay-off $\Phi(i, t)$ the number of past outcomes of action i that are taken into account is decisive. Since, in general, memory is bounded, only a narrowly limited number of past experiences is considered by each individual. We assume:

Assumption 2. The average pay-off $\Phi(i, t)$ refers to the last k_1 occasions of choosing action 1 and the last k_2 occasions of choosing action 2. k_1 and k_2 are finite.

By Assumption 2, the average pay-off $\Phi(i, t)$ depends on what is still memorized. Let $k_{11}(t)$ denote the number of past occasions accounted for in the formation of the averages at time t in which a player chose stage game action 1 while her/his opponent played stage game action 1. In the same manner we define $k_{12}(t)$, $k_{21}(t)$, and $k_{22}(t)$. By Assumption 2 we have $k_{11}(t) + k_{12}(t) = k_1(t)$ and $k_{21}(t) + k_{22}(t) = k_2(t)$. Consequently, the average pay-offs are given by

$$\Phi(1, t) = \frac{k_{11}(t)}{k_1(t)}\Pi(1, 1) + \frac{k_{12}(t)}{k_1(t)}\Pi(1, 2) \quad (4)$$

and

$$\Phi(2, t) = \frac{k_{21}(t)}{k_2(t)} \Pi(2, 1) + \frac{k_{22}(t)}{k_2(t)} \Pi(2, 2). \tag{5}$$

For expository convenience we assume with respect to the time structure of the individual adjustments:²

Assumption 3. The proportion $r(t)$ of players in the population who choose action 1 at time t changes sufficiently slowly so that it can be treated as constant during the time interval—determined by $k_{11}(t)$, $k_{12}(t)$, and T —to which the formation of the average pay-offs refers.

Finally, we need an assumption specifying function $F(\cdot)$ in Eq. (3):

Assumption 4. F is a linear function given by

$$F(\Pi) = \alpha \Pi \tag{6}$$

with $0 < \alpha < 1/(\max_{i,j,k,l}(\Pi_i, j) - \Pi(k, l))$.

By Assumption 4 it is ensured that $p(t)$ remains in the unit simplex. On the basis of these assumptions we can prove:

Lemma 1. *Given Assumptions 1–4, Hypothesis 1 implies*

$$p(t + T) = p(t) + \alpha p(t)(1 - p(t)) \times \left(\frac{k_{11}(t)}{k_1} \Pi(1, 1) + \frac{k_{12}(t)}{k_1} \Pi(1, 2) - \frac{k_{21}(t)}{k_2} \Pi(2, 1) - \frac{k_{22}(t)}{k_2} \Pi(2, 2) \right) \tag{7}$$

where each combination of $k_{11}(t)$, $k_{12}(t) = k_1 - k_{11}(t)$, $k_{21}(t)$, and $k_{22}(t) = k_2 - k_{21}(t)$ occurs with the probability

$$P(k_{11}(t), k_{21}(t) | k_1, k_2) = \binom{k_1}{k_{11}(t)} \binom{k_2}{k_{21}(t)} r(t)^{k_{11}(t)+k_{21}(t)} (1 - r(t))^{k_1+k_2-k_{11}(t)-k_{21}(t)}. \tag{8}$$

Remark. Eq. (7) shows that the dynamics of melioration learning are not strictly determined by the probability distribution of the alternative stage game actions in the population which is given by $r(t)$. Rather, the learning process involves an additional stochastic element resulting from the particular sequence of individually realized pay-offs. The individual sequence of pay-offs enters Eq. (7) in the form of the frequencies $k_{11}(t)$ and $k_{12}(t)$, which range from 0 to k_1 and k_2 , respectively. The probability for a certain history characterized by $k_{11}(t)$ and $k_{12}(t)$ is given by Eq. (8).

² Assumption 3 refers to the speed of the learning process relative to the length of the memorized time period. If either the speed of learning, given by α , is low, or the memory of the players is short, Assumption 3 is justified.

Let us now turn to the implications at the population level. Because of symmetry (Assumption 1) all players learn in an identical manner. This means that in Eq. (7), which gives the individual learning dynamics, the function F is the same for all agents. Nevertheless, the players' behavior at a time t , given by $p(t)$, is likely to differ, because the players' learning processes depend on the particular realization of the stochastic matching process they have individually experienced in the past. In contrast to the reinforcement model of Börgers and Sarin (1997) we thus, explicitly account for heterogeneity in learning experience and behavior. The behavior of the whole population of players is adequately described by a density $f(p, t)$ which, for each possible behavior $p(t) = p$, denotes the likelihood of finding a player behaving that way in the population at time t . The overall probability $r(t)$ of meeting an opponent who plays action 1 at time t is then given by

$$r(t) = \int pf(p, t) dp. \tag{9}$$

This is to say that the dynamics of $r(t)$ depend on the dynamics of $f(p, t)$ which, in turn, depend on the learning algorithm specified in Eq. (7). The dynamics of interactive melioration learning at the population level can be described by the following theorem.

Theorem 1. *The probability $r(t)$ of a player chosen at random from the population playing action 1 at time t changes over time according to*

$$r(t + T) = r(t) + \alpha [r(t)(1 - r(t)) - v(t)] [r(t) (\Pi(1, 1) - \Pi(2, 1)) + (1 - r(t)) (\Pi(1, 2) - \Pi(2, 2))] \tag{10}$$

where

$$v(t) = \int (p - r(t))^2 f(p, t) dp. \tag{11}$$

The proof of Theorem 1 is given in Appendix A.

3. Dynamic properties of melioration learning

The implications of melioration learning (Hypothesis 1) for the interactive learning case can now be demonstrated by investigating the convergence properties of the resulting process (10). A first question that can be raised is whether the stable states of interactive melioration learning correspond to Nash equilibria. To answer the question, we calculate the stationary states of the melioration learning process on the population level:

Lemma 2. *The difference Eq. (10) has up to three stationary solutions:*

- if $\Pi(1, 1) + \Pi(2, 2) - \Pi(1, 2) - \Pi(2, 1) \neq 0$ and
 - either $\Pi(1, 1) + \Pi(2, 2) - \Pi(1, 2) - \Pi(2, 1) > \Pi(2, 2) - \Pi(1, 2) > 0$
 - or $\Pi(1, 1) + \Pi(2, 2) - \Pi(1, 2) - \Pi(2, 1) < \Pi(2, 2) - \Pi(1, 2) < 0$,

Table 1
Conditions for stable behavior in a 2×2 game

State	Condition for ESS met by stable state of melioration learning	Nash equilibrium condition
$r_l = 0$	$\Pi(2, 2) \geq \Pi(1, 2)$ or $\Pi(2, 2) = \Pi(1, 2)$ and $\Pi(2, 1) > \Pi(1, 1)$	$\Pi(2, 2) > \Pi(1, 2)$
r_0 (given by Eq. (12))	$\Pi(1, 1) + \Pi(2, 2) - \Pi(1, 2) - \Pi(2, 1) < 0$	$\Pi(1, 1) + \Pi(2, 2) - \Pi(1, 2) - \Pi(2, 1) < 0$
$r_r = 1$	$\Pi(1, 1) \geq \Pi(2, 1)$ or $\Pi(1, 1) = \Pi(2, 1)$ and $\Pi(1, 2) > \Pi(2, 2)$	$\Pi(1, 1) > \Pi(2, 1)$

the following three stationary solutions exist:

$$r_0 = \frac{\Pi(2, 2) - \Pi(1, 2)}{\Pi(1, 1) + \Pi(2, 2) - \Pi(1, 2) - \Pi(2, 1)} \quad (12)$$

$r_l = 0$, and $r_r = 1$,

- if $\Pi(1, 1) + \Pi(2, 2) - \Pi(1, 2) - \Pi(2, 1) = 0$ or either $\Pi(1, 1) + \Pi(2, 2) - \Pi(1, 2) - \Pi(2, 1) > 0 > \Pi(2, 2) - \Pi(1, 2)$ or $\Pi(1, 1) + \Pi(2, 2) - \Pi(1, 2) - \Pi(2, 1) < 0 < \Pi(2, 2) - \Pi(1, 2)$, r_l and r_r are the only possibly stable stationary solutions.

In the first case r_0 is asymptotically stable if $\Pi(1, 1) + \Pi(2, 2) - \Pi(1, 2) - \Pi(2, 1) = 0$, while for both cases it holds that r_l is asymptotically stable if either $\Pi(2, 2) > \Pi(1, 2)$ or $\Pi(2, 2) = \Pi(1, 2)$ and $\Pi(2, 1) > \Pi(1, 1)$, and r_r is asymptotically stable if either $\Pi(1, 1) > \Pi(2, 1)$ or $\Pi(1, 1) = \Pi(2, 1)$ and $\Pi(1, 2) > \Pi(2, 2)$.

The proof of Lemma 2 is given in Appendix A.

Remark. Each asymptotically stable state of interactive melioration learning (represented by a stationary state of Eq. (10) and the corresponding stability conditions given in Lemma 2) is also an evolutionarily stable strategy (ESS)³ and a Nash equilibrium, see Table 1.

As mentioned in the introduction, evolutionary game theory has been able to show that the dynamics of interactive genetic adaptation follow a replicator dynamic. A second question that may therefore be raised here is whether interactive melioration learning also implies a replicator dynamic. The answer is as follows:

Theorem 2. In 2×2 games with properties as in Assumption 1, interactive melioration learning changes the behavior of the players always in the same direction as the replicator dynamics do, yet, the speed of behavior changes differs.

The proof of Theorem 2 is given in Appendix A.

³ Thus, interactive melioration learning with the specification of the function F according to Assumption 4 converges to an ESS in generic symmetric 2×2 -games. If two ESSs exist, the initial conditions determine to which of the equilibria the learning process converges.

Remark. The speed of learning is determined in Eq. (10) by $v(t)$, the variance of behavior within the population of players. The greater this variance, i.e. the more the players differ with respect to their behavior, the slower is the learning within the population as a whole. In the limiting cases of either $p = 0$ or $p = 1$ for all players, $v(t)$ reaches a maximum and equals $r(t)(1-r(t))$ which means that the behavior of the players does not change anymore. Such a state is a limit point. It is not a locally stable state (cf. Young and Foster, 1991 for a definition of local stability) because the learning process does not converge to this very state for all states in the immediate neighborhood. Nonetheless, interactive melioration learning may converge towards such a state. Hence, melioration learning does not necessarily attain the asymptotically stable states to which the replicator dynamics converge. Furthermore, Theorem 2 has been proven given that the function F is linear (Assumption 4). For non-linear functions F the replicator dynamics might no longer be a good approximation of the dynamic patterns resulting from interactive melioration learning.

4. Melioration under frequency-dependent pay-offs—the littering game

In the previous section it could be proved that, under fairly weak assumptions, the stable stationary states of melioration learning are Nash equilibria in a 2×2 game as long as pay-offs are invariable. The result supports what has been expected, namely that, under the latter condition, behavior following the matching law does not crucially deviate from optimizing behavior. The situation changes dramatically if the average outcome of each action varies with the frequency with which the action has been chosen in the past (Herrnstein, 1991). Full rationality would then require the individuals to account for that frequency-dependency effect when they make their choices, while behavior following the matching law ignores the effect. Herrnstein and Prelec (1991) report that, in experiments in which they arranged average outcomes of choices to vary inversely with the relative frequency of the respective choices, most of their test persons behaved according to the matching law. By doing so, they attained clearly suboptimal results. In order to fully assess the implications of melioration learning for game theory it seems necessary, therefore, to extend our analysis to games in which the pay-offs are frequency-dependent in the sense just mentioned.

Games with changing pay-offs have not been one of the major research topics in game theory so far (see Joosten et al., 1994 for an exception). Several aspects of such games have been neglected. One such aspect is, to our knowledge, the consideration of learning processes in games with changing pay-offs. This neglect is in stark contrast to the empirical significance such games have. This is particularly true for repeated games in which average pay-offs decrease with the relative frequencies of the strategies chosen, i.e. for situations where repeated interactions run the risk of winding up in “social traps” (cf. Cross and Guyer, 1980). For example, in many instances of social dilemmas the disastrous consequences—e.g. the tragedy of the commons—seem to emerge, and intensify, only after repeated defective choices have been made.⁴ Conversely, they seem to vanish only gradually, if at all, when

⁴ Cf. the classical fishing game in Lancaster, 1973. More broadly speaking many situations seem to qualify as examples here in which an observer would be inclined to exclaim: “Why didn’t they look around, realize what they were doing, and stop before it was too late?” as in the sad story of Easter Island, see Diamond, 1995.

Table 2
Pay-off matrix of the littering game

Row-player	Column-player	
	Littering	Not littering
Littering	$6-8q(t), 6-8q(t)$	$4-8q(t), 6-8q(t)$
Not littering	$4-8q(t), 6-8q(t)$	$4-8q(t), 4-8q(t)$

the defective strategy is abandoned. In order to convey the basic idea and to discuss its implications in an exemplary fashion in a simple 2×2 game setting, consider a repeated game which we will call “littering game”. In each of its iterations a stage game is played which, in normal form, is described in Table 2.

The logic of the game is as follows. Each stage, the players independently and repeatedly choose between two actions. One is associated with littering the environment in some sense, while the other is not. The crucial point is that for each of the players the immediate pay-off in the stage game is at any time greater if they choose the littering action over the non-littering action, but that continued littering develops a disastrous impact on the future pay-offs: the more the players have littered in the past the lower their pay-offs will be in the future. This is expressed by the term $q(t)$, defined as the average percentage of littering in the rounds before time t weighted with an exponentially decreasing factor. In the numerical example in Table 1, the stage game pay-offs of the players for littering and not littering both decrease with $q(t)$ linearly and, as $q(t)$ grows, eventually turn negative. Thus, the game is an example of a classical social trap.⁵

If, after extensive littering, the littering action is abandoned the environment is imagined to slowly recover. This means that the negative effect of littering on the stage game pay-offs needs time to disappear. More precisely we assume:

Assumption 5. The dynamics of $q(t)$ are given by

$$q(t + T) = (1 - \eta)q(t) + \eta l(t) \quad (13)$$

where η is an impact or speed parameter, $0 < \eta < 1$, and $l(t)$ is the relative share of players who litter at time t .

This assumption implies that the impact of littering decreases exponentially after littering has been abandoned.⁶ Because of its temporal structure, this game is different from the standard repeated prisoner’s dilemma game. First, although the behavior of each player has an impact on the pay-offs of all players in a way similar to social dilemma games, the impact

⁵ Notice that the stage game pay-off of each player at any time t does not depend on the behavior of the other player at time t . This simplifying assumption eases our exposition and calculations below. It certainly represents a special case of a dominant strategy which may, however, be justified as long as the players’ behavior in each iteration of the game has only a marginal effect on the accumulating disaster. What is decisive is the dependence of the stage game pay-offs at time t on the strategies played by *both* players before time t .

⁶ In the case of two players the proportion $l(t)$ is either 0, 0.5, or 1. The same proportion $l(t)$ of littering for a long time causes the variable $q(t)$ to converge to the value of $l(t)$.

occurs with a time delay. Second, should the impact of $q(t)$ on the own pay-offs be recognized, there is an additional intra-personal dilemma which, in face of the temptation of the greater immediate reward, is expressed by the well-known ‘beware-of-the-consequences’ maxim.⁷

To bring all these features to bear within the “distributed choices” framework and the simple 2×2 game we have to relax Assumption 1 and assume instead:

Assumption 1’. Two players repeatedly play the littering game by choosing one of two possible actions simultaneously with stage game pay-offs as given in Table 2.

Similarly to the analysis of melioration learning in the repeated 2×2 game with invariable pay-offs, each player I ($I \in \{1, 2\}$) is characterized by her/his probabilities $p_I(t)$ and $(1 - p_I(t))$ of choosing action 1 or 2—littering or not littering respectively—where the choice at time t again depends on the average pay-offs $\Phi_I(1, t)$ and $\Phi_I(2, t)$ determined according to Assumption 2. In line with Hypothesis 1 and analogously to Eq. (3), the probabilities $p_I(t)$ change according to the first-order difference equation

$$p_I(t + T) = p_I(t) + \alpha p_I(t)(1 - p_I(t))(\Phi_I(1, t) - \Phi_I(2, t)). \tag{14}$$

The average pay-offs $\Phi_I(1, t)$ and $\Phi_I(2, t)$ are independent of the probability distributions $p_J(t)$ of other players J but depend on the variable $q(t)$ which varies over time according to Eq. (13). To analyze the dynamics in Eq. (14) we have to introduce the following analogy to Assumption 3.

Assumption 3’. The value of $q(t)$ changes sufficiently slowly so that it can be treated as constant during the time interval—determined by $k_{11}(t)$, $k_{12}(t)$, and T —to which the formation of the average pay-offs refers.

We now can prove the following result.

Theorem 3. Given Assumptions 1’, 2 and 3’,

$$\lim_{t \rightarrow \infty} p_I(t) = 1 \quad \forall I \in \{1, 2\}$$

i.e. interactive melioration learning in the littering game converges to a stationary state in which both players choose the littering action in each stage game.

The proof of Theorem 3 is given in Appendix A.

Remark. It may be worth noting that the proof of Theorem 3 does not depend on any of the parameters.

In order to assess the implications of melioration learning in this social trap game, a comparison with rational choice behavior would be illuminating. It is not clear, however,

⁷ The Latin verse “*principiis obsta, sero medicina paratur*” (withstand the beginnings—no cure will be available) from Ovid’s *Remedia Amoris* has become a famous quote indicating that the perils of the slippery slope epitomized by the intra-personal dilemma are already known to the old Romans.

what kind of rational behavior is suitable for such a comparison. By Assumption 1', two players repeatedly interact with each other so that contingent action would be feasible to them. Yet, for the low profile routine decision problems of everyday life on which the 'distributed choice' interpretation of the melioration hypothesis focuses highbrowed contingent action may not necessarily appear an appropriate level of responding 'rationally'. Let us therefore investigate two cases: one in which the players do not behave contingent on the actions of others and one in which they do, i.e. where their actions depend on all previous moves.

In the first case contingent action (as e.g. tit-for-tat like strategies) is thus ignored. Players are assumed to account for the influence which their own current choices exert on the future development of their pay-offs only. This implies that the players recognize the 'beware-of-the-consequences' dilemma of the littering game against which they can individually choose an optimal strategy depending on their time preference. For a discounting of future pay-offs by the factor \hat{a} we can then prove:

Theorem 4a. *Rational players who maximize the discounted stream of expected pay-offs in the littering game without resorting to contingent action choose to litter in each stage game as long as*

$$\beta < \frac{1}{1 + \eta}. \quad (15)$$

are indifferent between the two actions in each stage game if $\beta = 1/(1 + \eta)$, and choose not to litter in any stage game otherwise.

The proof of Theorem 4a is given in Appendix A.

Remark. In this version of rational behavior, optimal strategic choice in the littering game depends on two parameters: η , the 'speed of adjustment' (see Eq. (13)) and β , the discount factor.

Let us now turn to the second case involving contingent action.

Theorem 4b. *For two rational players who maximize the discounted stream of expected pay-offs and who resort to contingent action, the following solutions are obtained:*

- (1) *for $\beta < 1/(1 + 3\eta)$, the repeated game has one Nash equilibrium in which both players litter at all stages.*
- (2) *for $1/(1 + 3\eta) < \beta < 1/(1 + \eta)$, any strategy with both players either littering or not littering, i.e. doing the same in each stage of the repeated game, can be "stabilized".⁸*

⁸ Consider a pair of repeated-game strategies that is not a Nash equilibrium. These repeated game strategies may be augmented by some kind of punishing strategy such that grim strategies result. These grim strategies are such that both players use the original repeated game strategies as long as the opponent does the same, and both players use the punishing repeated game strategy in case that the opponent deviates. If the original repeated game strategies lead to a higher pay-off than the threat-point of the punishing strategy for both players, the grim strategies represent a Nash equilibrium. In this sense, the original strategies can be "stabilized".

(3) for $\beta > 1/(1 + \eta)$, sequences of stage game actions in which the players never litter or litter only every s -th period can be “stabilized” if

$$s > \frac{\ln[(1 - 2\beta + \beta^2 - \eta\beta + 3\eta\beta^2)/(2\eta\beta)]}{\ln \beta}. \tag{16}$$

The proof of Theorem 4b is given in Appendix A.

Remark. The results of Theorems 4a and 4b are identical if $\beta < 1/(1 + 3\eta)$ is satisfied. For $\beta > 1/(1 + 3\eta)$ the use of contingent action enlarges the set of equilibria. If $1/(1 + \eta) > \beta > 1/(1 + 3\eta)$, a situation can emerge in which both players are never littering. Thus, for $1/(1 + \eta) > \beta > 1/(1 + 3\eta)$ contingent action indeed makes a great difference compared to what Theorem 4a predicts when contingent action is absent.

Rationality thus dictates to choose littering only if the ability of nature to recover is sufficiently weak compared with the discounting factor, i.e. $\beta < 1/(1 + 3\eta)$. This clearly contrasts with the consequences of interactive melioration learning which leads to littering independent of the values of the parameters η and β .

How are melioration learning and the replicator dynamics related in the littering game setting? For each single player, melioration learning has been described by Eq. (14). The probability distribution for the behavior of a population of players changes according the same equation. Under Assumption 4 we get

$$p(t + T) = p(t) + 2\alpha p(t)(1 - p(t)) \tag{17}$$

if $p(t)$ is taken to indicate the (average) likelihood of encountering an opponent who chooses littering at time t . As stated above, replicator dynamics mean that the average likelihood $p(t)$ changes according to

$$p(t + T) - p(t) = \mu(t)p(t)(\Phi_1(t) - \langle \Phi(t) \rangle) \tag{18}$$

where $\Phi_1(t)$ is the average stage game pay-off of a player who chooses littering at time t , i.e. $\Phi_1(t) = 6 - 8q(t)$, and $\langle \Phi(t) \rangle$, is the average stage game pay-off in the population of players at time t , i.e. $\langle \Phi(t) \rangle = p(t)(6 - 8q(t)) + (1 - p(t))(4 - 8q(t))$. Inserting equations for $\Phi_1(t)$ and $\langle \Phi(t) \rangle$, into Eq. (17) we obtain

$$p(t + T) - p(t) = 2\mu(t)p(t)(1 - p(t)). \tag{19}$$

It may thus be concluded that the replicator dynamics adequately describe melioration learning also in games with frequency-dependent pay-offs.

5. Conclusions

An attempt has been made in this paper to analyze the implications of the melioration learning hypothesis within a game-theoretic setting. Melioration learning—a variant of reinforcement learning suggested by the psychologist R.J. Herrnstein—converges to the matching law which describes a kind of behavior for which a huge amount of experimental

evidence has been collected in psychology (Williams, 1988). In the present investigation, two different classes of repeated games have been explored: games with invariable stage game pay-offs and those with frequency-dependent stage game pay-offs. For the former class we have shown that melioration learning converges to Nash equilibria and does so in a way similar to the replicator dynamics. For the latter class of games, melioration learning has been found to deviate from optimal behavior in a non-interactive setting, if the pay-offs decrease when the relative frequency with which an action is chosen increases (Herrnstein and Prelec, 1991). As an example of a repeated game with frequency-dependent stage game pay-offs the “littering game” has been introduced. For this game, it has been shown that interactive melioration learning is still appropriately described by the replicator dynamics, but that it can systematically deviate from what may be considered rational behavior in such a game.

From a more general point of view, we have been able to demonstrate that the tools provided by evolutionary game theory may be applied to also describe the implications of this empirically significant kind of reinforcement learning in interactive settings. A characteristic feature of all forms of reinforcement learning is the low cognitive participation of the individuals when making choices. It is perhaps no wonder, then, that in situations in which the pay-offs are frequency-dependent—as in social trap games—reinforcement learning promises no good. If, for whatever reasons, the players fail to recognize the long-run effects of those choices which are more rewarding in the short-run, and if they simply follow the melioration principle (or the replicator dynamics) in adapting their behavior, then they are doomed to wind up in what is, in terms of rational conduct, an inefficient or even harmful state of affairs.

Appendix A

Proof of Lemma 1. Eq. (7) follows from inserting Eqs. (4) and (5) into Eq. (3). According to Assumption 3, $r(t)$ is held constant over the time interval considered for calculating the average pay-offs. Since the $k_{ij}(t)$ are stochastic variables which depend on the particular sequence of events in the past, the probability that there are $k_{i1}(t)$ occasions on which the players chose action i and their opponents played action 1 is given by a binomial distribution

$$P(k_{i1}(t) = k) = \binom{k_i(t)}{k} r(t)^k (1 - r(t))^{k_i(t) - k} \quad (\text{A.1})$$

with $k_{12}(t) = k_1 - k_{11}(t)$ and $k_{22}(t) = k_2 - k_{21}(t)$. The probability for a set $k_{11}(t)$, $k_{12}(t)$, $k_{21}(t)$, and $k_{22}(t)$ to occur is then given by the multiplication of the probability for $k_{11}(t)$ and the probability for $k_{21}(t)$ as given in Eq. (8). \square

Proof of Theorem 1. Consider all players using action 1 with probability $p(t)$ at time $t + T$. Since the probability of choosing action 1 can only change by certain values according to Eq. (7), those players must have chosen action 1 with probability $p(t)$ at time t where $p(t + T)$ is given by Eq. (7). However, only a proportion $P(k_{11}, k_{21}/k_1, k_2)$ of those players who have

chosen action 1 with probability $p(t)$ at time t do change their behavior correspondingly. If we take the number of players who may change their behavior in such a way and multiply it with the corresponding probability, we obtain for $f(p(t + T), t + T)$

$$f(p(t + T), t + T) = \sum_{k_{11}, k_{21}} f(p(t), t) \frac{dp(t)}{dp(t + T)} P(k_{11}, k_{21} | k_1, k_2). \tag{A.2}$$

Furthermore, $r(t + T)$ can be written as a function of $f(p(t + T), t + T)$:

$$r(t + T) = \int p(t + T) f(p(t + T), t + T) dp(t + T). \tag{A.3}$$

Inserting Eq. (A.2) into Eq. (A.3) we obtain

$$r(t + T) = \sum_{k_{11}, k_{21}} P(k_{11}, k_{21} | k_1, k_2) \int p(t + T) f(p(t), t) dp(t). \tag{A.4}$$

By inserting Eq. (7) into Eq. (A.4) the integral can be solved and the sum can be calculated with the result

$$r(t + T) = r(t) + \alpha [r(t)(1 - r(t)) - v(t)] [r(t) (\Pi(1, 1) - \Pi(2, 1)) + (1 - r(t)) (\Pi(1, 2) - \Pi(2, 2))] \tag{A.5}$$

where

$$v(t) = \int (p - r(t))^2 f(p, t) dp. \quad \square \tag{A.6}$$

Proof of Lemma 2. In a stationary solution $r(t + T) = r(t)$. In Eq. (10) this is the case if either

$$r(t)\Pi(1, 1) + (1 - r(t))\Pi(1, 2) - r(t)\Pi(2, 1) - (1 - r(t))\Pi(2, 2) = 0 \tag{A.7}$$

or

$$r(t) (1 - r(t)) - v(t) = 0. \tag{A.8}$$

Rearranging Eq. (A.7) we get Eq. (12) if $\Pi(1, 1) + \Pi(2, 2) - \Pi(1, 2) - \Pi(2, 1) \neq 0$. If $\Pi(1, 1) + \Pi(2, 2) - \Pi(1, 2) - \Pi(2, 1) = 0$ then Eq. (A.7) leads to $\Pi(1, 2) = \Pi(2, 2)$ implying also $\Pi(1, 1) = \Pi(2, 1)$. We exclude the case in which $\Pi(1, 2) = \Pi(2, 2)$ and $\Pi(1, 1) = \Pi(2, 1)$ hold simultaneously here because in this case the stage game pay-offs of the player would not depend on the actions of the opponent so that the interaction between them has no effect. As a consequence, Eq. (A.7) cannot be satisfied for $\Pi(1, 1) + \Pi(2, 2) - \Pi(1, 2) - \Pi(2, 1) = 0$ so that there is no stationary state r_0 in this case. Furthermore, the value of r_0 might be greater than 1 or less than 0. In this case the stationary state r_0 is not relevant for the dynamics of the learning process. Thus, r_0 represents a stationary state within the relevant range of r only, if either $\Pi(1, 1) + \Pi(2, 2) - \Pi(1, 2) - \Pi(2, 1) > \Pi(2, 2) - \Pi(1, 2) > 0$ or $\Pi(1, 1) + \Pi(2, 2) - \Pi(1, 2) - \Pi(2, 1) < \Pi(2, 2) - \Pi(1, 2) < 0$.

Inserting $r(t) = r_0 + \varepsilon$ into Eq. (10) and using Eq. (12) to transform the last factor we obtain

$$r(t + T) = r_0 + \varepsilon + \alpha [(r_0 + \varepsilon)(1 - r_0 - \varepsilon) - v(t)] \varepsilon (\Pi(1, 1) - \Pi(1, 2) - \Pi(2, 1) + \Pi(2, 2)). \tag{A.9}$$

From Eq. (A.9) it follows that the stationary state r_0 is stable if and only if $\Pi(1, 1) + \Pi(2, 2) - \Pi(1, 2) - \Pi(2, 1) < 0$ is satisfied, because

$$r(t) (1 - r(t)) - v(t) \geq 0 \tag{A.10}$$

holds for all $r(t)$ as will be proven in the following. $v(t)$ is defined by Eq. (11) that can be transformed to

$$v(t) = \int p^2 f(p, t) dp - 2r(t) \int pf(p, t) dp + r(t)^2 \int f(p, t) dp. \tag{A.11}$$

Thus,

$$v(t) = \int p^2 f(p, t) dp - r(t)^2 \leq \int pf(p, t) dp - r(t)^2 = r(t) - r(t)^2 \tag{A.12}$$

which proves Eq. (A.10). Eq. (A.8) is satisfied if $v(t) = r(t) - r(t)^2$ which, according to Eq. (A.11), holds if

$$\int p^2 f(p, t) dp = \int pf(p, t) dp. \tag{A.13}$$

Eq. (A.13) holds if and only if $f(p, t) = 0$ for all $p \neq 0, 1$. Hence, Eq. (A.7) is satisfied if and only if for each player either $p(t) = 0$ or $p(t) = 1$. Since the dynamics of $p(t)$ are given for each player by the same Eq. (7), $p(t)$ either increases or decreases for all individuals in the average, or it stays the same. Hence, a situation in which $p(t) = 0$ for some players and $p(t) = 1$ for others cannot be a stable state. The only possibly stable states that satisfy Eq. (A.13), and hence, Eq. (A.8), are those where either $p(t) = 0$ or $p(t) = 1$ for all players. It follows that $r_l = 0$ and $r_r = 1$ are the possible stable stationary states of Eq. (10) that satisfy Eq. (A.8). Inserting $r(t) = \varepsilon$ into Eq. (10) we obtain

$$r(t + T) = r(t) + \alpha [\varepsilon(1 - \varepsilon) - v(t)] [(\Pi(1, 2) - \Pi(2, 2)) + \varepsilon (\Pi(1, 1) - \Pi(1, 2) - \Pi(2, 1) + \Pi(2, 2))]. \tag{A.14}$$

By Eq. (A.14), r_l is stable if and only if either $\Pi(2, 2) > \Pi(1,2)$ or $(\Pi(2, 2) = \Pi(1, 2)$ and $\Pi(2, 1) > \Pi(1, 1))$. Similarly, we obtain that r_r is stable if and only if $\Pi(1, 1) > \Pi(2, 1)$ or $(\Pi(1, 1) = \Pi(2, 1)$ and $\Pi(1, 2) > \Pi(2, 2))$. \square

Proof of Theorem 2. In the case of two strategies the replicator dynamics are defined as (see e.g. Taylor and Jonker, 1978)

$$r(t + T) - r(t) = \mu(t)r(t)(\Phi_1(t) - \langle \Phi(t) \rangle). \tag{A.15}$$

where $\mu(t)$ represents a measure for selection pressure, $\Phi_1(t)$ is the average pay-off of an agent who chooses action 1 at time t , and $\langle \Phi(t) \rangle$ is the average pay-off realized by the entire

population of agents at time t . In a 2×2 game with the population playing action 1 with a probability of $r(t)$ the average pay-off for a player who takes action 1 is

$$\Phi_1(t) = r(t)\Pi(1, 1) + (1 - r(t))\Pi(1, 2). \tag{A.16}$$

The average pay-off of the whole population is given by

$$\langle \Phi(t) \rangle = r(t)^2\Pi(1, 1) + r(t)(1 - r(t))(\Pi(1, 2) + \Pi(2, 1)) + (1 - r(t))^2\Pi(2, 2). \tag{A.17}$$

Inserting Eqs. (A.16) and (A.17) into Eq. (A.15) we obtain

$$\frac{dr(t)}{dt} = \mu(t)r(t)(1 - r(t))(r(t)\Pi(1, 1) + (1 - r(t))\Pi(1, 2) - r(t)\Pi(2, 1) - (1 - r(t))\Pi(2, 2)). \tag{A.18}$$

Eq. (A.18) is, except for the term $v(t)$, identical to Eq. (10). Both depend on the pay-offs $\Pi(i, j)$ in the same way and, according to Eq. (A.10), the factor $r(t)(1 - r(t)) - v(t)$ is always positive or equal to zero, which also holds for $r(t)(1 - r(t))$, so that the direction of change is the same for melioration learning as for the process implied by the replicator dynamics. Only the speed of this change differs between the two processes. □

Proof of Theorem 3. In the formation of the average pay-off of player I , the last k_{I1} occasions in which (s)he has taken action 1 are accounted for. Let all these occasions form a set M of points in time such that player I remembers all $t \in M$ where (s)he chose action 1. Given Assumption 1' and Assumption 2, the average pay-off $\Phi_I(1, t)$ is then given by

$$\Phi_I(1, t) = 6 - \frac{8}{k_{I1}} \sum_{t' \in M} q(t'). \tag{A.19}$$

Under Assumption 3', Eq. (A.19) can be approximated by

$$\Phi_I(1, t) = 6 - 8q(t). \tag{A.20}$$

Analogously, $\Phi_I(2, t)$ can be approximated by

$$\Phi_I(2, t) = 4 - 8q(t). \tag{A.21}$$

Inserting Eqs. (A.20) and (A.21) into Eq. (14) we obtain

$$p_I(t + T) = p_I(t) + p_I(t)(1 - p_I(t))F(2). \tag{A.22}$$

The right hand side of Eq. (A.22) is strictly positive for $p_I(t) \neq 0, 1$, since $F(2)$ is positive. Thus, if $p_I(0) > 0$, $p_I(t + T) > p_I(t)$ holds for all t as long as $p_I(t) \neq 1$. Hence, $p_I(t)$ increases with time for all $p_I(t) \neq 0, 1$ and is bounded by $p_I(t) = 1$ so that $p_I(t)$ converges to 1. □

Proof of theorem 4a. s_1 and s_2 denoting the repeated game strategies used by the two players, the β -discounted reward $\gamma_\beta(s_1, s_2, t)$ at stage t of the repeated game is given by

$$\gamma_\beta(s_1, s_2, t) = (1 - \beta) \sum_{\tau=(t/T)}^{\infty} (6 - 2c_1(\tau) - 8q(\tau))\beta^{\tau-(t/T)}$$

with the action $c_1(\tau)$ of the players at time τ ($c_1(\tau) = 0$ denotes littering and $c_1(\tau) = 1$ not littering) according to the strategies s_1 . With the help of Eq. (13) and after some transformations we obtain

$$\begin{aligned} \gamma_\beta(s_1, s_2, t) = & 6 - (1 - \beta) \sum_{\tau=(t/T)}^{\infty} [\beta^{\tau-t/T} 2c_1(\tau)] - 8q(t) \frac{1 - \beta}{1 - \beta + \beta\eta} \\ & - 8\eta \frac{(1 - \beta)\beta}{1 - \beta + \beta\eta} \sum_{\tau=(t/T)}^{\infty} [l(\tau)\beta^{\tau-t/T}]. \end{aligned} \tag{A.23}$$

At time t a player has the option of littering or not littering. By making their choices, the players either add or do not add the value $1/2$ to $l(t)$. The stage game pay-off at time t increases by 2 if the players are littering instead of not littering. Thus, with $c_L(t) = 0$ and $c_N(t) = 1$ at time t the difference in the β -discounted reward between littering at time t ($s_1 = s_L$) and not littering at time t ($s_1 = s_N$) is given by

$$\Delta\gamma_\beta(s_L|s_N, s_2, t) = \gamma_\beta(s_L, s_2, t) - \gamma_\beta(s_N, s_2, t) = 2(1 - \beta) - \frac{4\eta(1 - \beta)\beta}{1 - \beta + \beta\eta}. \tag{A.24}$$

Hence, a rational player prefers littering over not littering if

$$\frac{2\eta\beta}{1 - \beta + \beta\eta} < 1 \tag{A.25}$$

which corresponds to Eq. (15). (S)he is indifferent between the two actions if

$$\frac{2\eta\beta}{1 - \beta + \beta\eta} = 1 \tag{A.26}$$

and prefers not to litter if

$$\frac{2\eta\beta}{1 - \beta + \beta\eta} > 1. \tag{A.27}$$

□

Proof of Theorem 4b. With reference to the folk theorem we determine the threat-point first. The obvious choice for punishing an opponent in the present game is the littering action. According to Theorem 4a, a player who is being punished will be littering if $\beta < 1/(1 + \eta)$ and will not be littering if $\beta > 1/(1 + \eta)$. The threat-point, therefore, follows from Eq. (A.23) for $\beta < 1/(1 + \eta)$ as

$$\gamma_{\beta, \min}(s_1, s_2, t) = 6 - 8q(t) \frac{1 - \beta}{1 - \beta + \beta\eta} - 8\eta \frac{\beta}{1 - \beta + \beta\eta} \tag{A.28}$$

with s_1, s_2 again denoting the repeated game strategies, while for $\beta < 1/(1 + \eta)$ the threat-point is

$$\gamma_{\beta, \min}(s_1, s_2, t) = 4 - 8q(t) \frac{1 - \beta}{1 - \beta + \beta\eta} - 4\eta \frac{\beta}{1 - \beta + \beta\eta}. \tag{A.29}$$

All sequences of stage game actions that lead to a higher reward than $\lambda_{\beta, \min}(s_1, s_2, t)$ for both players can be stabilized by a grim strategy that threatens to punish any deviation from those sequences by choosing littering in successive stage games. Hence, we have to consider sequences of actions which induce for both players rewards higher than $\gamma_{\beta, \min}(s_1, s_2, t)$. To keep complexity down we consider those sequences of stage game actions where, at each point in time, both players choose the same action. Two cases have to be treated separately.

Assume first that condition $\beta < 1/(1 + \eta)$ is satisfied. In this case the threat-point results from both players choosing littering at all times as follows. Let there be a sequence of stage game actions in which both players are always littering except at one point in time τ in which they simultaneously are not littering in a stage game. The resulting change in rewards for such a repeated game strategy change is given by

$$\Delta\gamma_{\beta}(t) = -2(1 - \beta)\beta^{(\tau-t)/T} + \frac{8\eta(1 - \beta)\beta^{(\tau-t)/T+1}}{1 - \beta + \beta\eta}. \tag{A.30}$$

$\Delta\gamma_{\beta}(t)$ in Eq. (A.30) is positive if $\beta > 1/(1 + 3\eta)$, and it is negative if $\beta < 1/(1 + 3\eta)$. The same holds for any additional stage game in which the players switch from littering to not-littering. Thus, if $1/(1 + \eta) > \beta > 1/(1 + 3\eta)$ is satisfied, all strategies where both players choose the same action at any time lead to a reward that is higher than the threat-point and is therefore individually rational and feasible. If $\beta < 1/(1 + 3\eta)$, then $\Delta\gamma_{\beta}(t) < 0$. Thus, if both players refrain from littering in the same way they decrease their rewards. Hence, in that case the only strategy combination with both players acting identically that can be stabilized is littering in each stage game.

Now imagine the condition $\beta > 1/(1 + \eta)$ is satisfied. The threat-point as given in Eq. (A.29) then results from a situation in which one player is never littering in any single stage game while the opponent is always littering. Assume both players are never littering in any stage game ($s_1 = s_2 = s_0$). They then obtain a reward of

$$\gamma_{\beta}(s_0, s_0, t) = 4 - 8q(t) \frac{1 - \beta}{1 - \beta + \beta\eta} \tag{A.31}$$

which is greater than the threat-point. Therefore, a repeated game strategy where both players are never littering can be stabilized. Furthermore, if, starting at time t , both players are littering in every s -th stage game and are not littering in all other stage games ($s_1 = s_2 = s_s$), they obtain a reward given by

$$\gamma_{\beta}(s_s, s_s, t) = 4 + \frac{2(1 - \beta)}{1 - \beta^s} - 8q(t) \frac{1 - \beta}{1 - \beta + \beta\eta} - 8\eta \frac{(1 - \beta)\beta}{(1 - \beta + \beta\eta)(1 - \beta^s)}. \tag{A.32}$$

To stabilize such a behavior the reward (A.32) has to be greater than the threat-point, i.e.

$$\frac{2(1-\beta)}{1-\beta^s} - 8\eta \frac{(1-\beta)\beta}{(1-\beta+\beta\eta)(1-\beta^s)} > -4\eta \frac{\beta}{1-\beta+\beta\eta}. \quad (\text{A.33})$$

Multiplying both sides of Eq. (A.33) with $(1-\beta+\beta\eta)(1-\beta^s)$ we obtain

$$1 - 2\beta + \beta^2 - \eta\beta + 3\eta\beta^2 > 2\beta^{1+s}\eta. \quad (\text{A.34})$$

Finally, Eq. (A.34) can be transformed into

$$s > \frac{\ln[(1 - 2\beta + \beta^2 - \eta\beta + 3\eta\beta^2)/(2\eta\beta)]}{\ln \beta}. \quad (\text{A.35})$$

If Condition (A.35) is satisfied, a strategy combination where each player is littering in every s -th stage game can be stabilized. \square

References

- Binmore, K., Samuelson, L., Vaughan, R., 1995. Musical chairs: modelling noisy evolution. *Games and Economic Behavior* 11, 1–35.
- Börgers, T., Sarin, R., 1997. Learning through reinforcement and replicator dynamics. *Journal of Economic Theory* 77, 1–16.
- Brenner, T., 1997. Reinforcement Learning in a 2×2 Games and the Concept of Reinforcably Stable Strategies. *Papers on Economics & Evolution* no. 9703. Max-Planck-Institute, Jena.
- Bush, R.R., Mosteller, F., 1955. *Stochastic Models for Learning*. Wiley, New York.
- Cross, J.G., Guyer, M.J., 1980. *Social Traps*. University of Michigan Press, Ann Arbor.
- Dawid, H., 1997. Learning of equilibria by a population with minimal information. *Journal of Economic Behavior & Organization*. 32, 1–18.
- Diamond, J., Easter's End. *Discover*. August 1995, pp. 62–69.
- Ellison, G., 1993. Learning, local interaction, and coordination. *Econometrica* 61, 1047–1071.
- 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, 848–881.
- Estes, W.K., Suppes, P., 1974. Foundations of stimulus sampling theory. In: Krantz, D.H., Luce, R.D., Atkinson, R.C., Suppes, P. (Eds.), *Contemporary Developments in Mathematical Psychology, Vol. I, Learning, Memory, and Thinking*. San Francisco, pp. 163–183.
- Hergenhahn, B.R., Olson, M.H., 1997. *An Introduction to Theories of Learning*, 5th Edition. Prentice-Hall, Upper Saddle River.
- Herrnstein, R.J., 1970. On the law of effect. *Journal of Experimental Analysis of Behavior* 13, 243–266.
- Herrnstein, R.J., 1991. Experiments on stable suboptimality in individual behavior. *American Economic Review (Papers and Proceedings)* 81, 360–364.
- Herrnstein, R.J., Loveland, D.H., 1975. Maximizing and matching on concurrent ratio schedules. *Journal of the Experimental Analysis of Behavior* 24, 107–116.
- Herrnstein, R.J., Prelec, D., 1991. Melioration: a theory of distributed choice. *Journal of Economic Perspectives* 5, 137–156.
- Herrnstein, R.J., Loewenstein, G.F., Prelec, D., Vaughan Jr., W., 1993. Utility maximization and melioration: internalities in individual choice. *Journal of Behavioral Decision Making* 6, 149–185.
- Hofbauer, J., Sigmund, K., 1988. *The Theory of Evolution and Dynamical Systems*. Cambridge University Press, Cambridge.
- Joosten, R., Peters, H., Thuijsman, F., 1994. Games with changing payoffs. In: Silverberg, G., Soete, L. (Eds.), *The Economics of Growth and Technical Change*. Aldershot, Edward Elgar, pp. 244–257.
- Lancaster, K., 1973. The dynamic inefficiency of capitalism. *Journal of Political Economy* 81, 1098–1109.

- Loewenstein, G., Prelec, D., 1992. Anomalies in intertemporal choice: evidence and an interpretation. *The Quarterly Journal of Economics* 107, 573–597.
- Offerman, T., Sonnemans, J., 1998. Learning by experience and learning by imitating successful others. *Journal of Economic Behavior & Organization* 34, 559–575.
- Roth, A., Erev, I., 1995. Learning in extensive form games: experimental data and simple dynamic models in the intermediate run. *Games and Economic Behavior* 6, 164–212.
- Samuelson, L., 1993. Recent advantages in evolutionary economics: comments. *Economic Letters* 42, 313–319.
- Samuelson, L., 1994. Stochastic stability in games with alternative best replies. *Journal of Economic Theory* 64, 35–65.
- Sanchirico, C.W., 1996. A probabilistic model of learning in games. *Econometrica* 64, 1375–1393.
- Selten, R., 1991. Evolution, learning, and economic behavior. *Games and Economic Behavior* 3, 3–24.
- Simon, H.A., 1957. *Models of Man: Social and Rational*. Wiley, New York.
- Taylor, P.D., Jonker, L.B., 1978. Evolutionary stable strategies and game dynamics. *Mathematical Biosciences* 40, 145–156.
- Vaughan, W., Herrnstein, R.J., 1987. Stability, melioration, and natural selection. In: Green, L., Kagel, J.H. (Eds.), *Advances in Behavioral Economics*, Vol. 1. Ablex, Norwood.
- Vega-Redondo, F., 1996. *Evolution, Games, and Economic Behavior*. Oxford University Press, Oxford.
- Weibull, J.W., 1995. *Evolutionary Game Theory*. The MIT Press, Cambridge.
- Williams, B.A., 1988. Reinforcement, choice, and response strength. In: Atkinson, R.C., Herrnstein, R.J., Lindzey, G., Luce, R.D. (Eds.), *Stevens' Handbook of Experimental Psychology*, Vol. 2. Wiley, New York, pp. 167–244.
- Young, P., Foster, D., 1991. Cooperation in the short and in the long run. *Games and Economic Behavior* 3, 145–156.
- Young, P., 1993. The evolution of conventions. *Econometrica* 61, 57–84.