Accepted Manuscript

Applying the matching law as micro-foundation of social phenomena

Johannes Zschache

PII: S0049-089X(17)30200-4

DOI: 10.1016/j.ssresearch.2018.03.010

Reference: YSSRE 2154

To appear in: *Social Science Research*

Received Date: 6 March 2017

Revised Date: 19 March 2018

Accepted Date: 20 March 2018

Please cite this article as: Zschache, J., Applying the matching law as micro-foundation of social phenomena, *Social Science Research* (2018), doi: 10.1016/j.ssresearch.2018.03.010.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Applying the matching law as micro-foundation of social phenomena

Johannes Zschache

Institute of Sociology, Leipzig University

Beethovenstraße 15, 04107 Leipzig, Germany

zschache@sozio.uni-leipzig.de

March 30, 2018

word count: 10976

Acknowledgements: This article is based on material that has formed part of my doctoral thesis. The thesis is available at http://nbn-resolving.de/urn:nbn:de:bsz:15-qucosa-216771. Advice given by Thomas Voss and Andreas Tutić has been of great help in the development of this work. Three anonymous reviewer provided valuable comments.

Software: The Simulations were run on NetLogo (Wilensky, 1999). The source code can be found at https://github.com/JZschache/NetLogo-ql/blob/master/models/volunteers.nlogo. It requires an extension of NetLogo that must be manually installed: https://github.com/JZschache/NetLogo-ql.

ACCEPTED MANUSCRIPT

Abstract

Social phenomena are suggested to be explained by the matching law - an empirical regularity of individual behaviour. While a considerable amount of psychological research on this law exists, only a few sociological applications can be found. This paper points to the problems that come with its usage as micro-foundation of social behaviour and provides solutions. In particular, a model of melioration learning enables the derivation of social phenomena from the matching law. The proposed approach is illustrated by the application of the learning model to the volunteer's dilemma. In contrast to game-theoretical solutions, the matching law leads to more intuitive results in case of the asymmetric dilemma. The relationship between the matching law and utility maximisation is discussed by its integration into economic consumer theory.

Keywords: social theory; methodological individualism; melioration learning; reinforcement learning; volunteer's dilemma; consumer theory

1 Introduction

When adopting the explanatory framework of methodological individualism, behavioural assumptions must be specified at the micro-level. Different rules of decision-making are applicable. For instance, the assumption of rationality has been repeatedly used in sociological research (e.g. Olson, 1965; Becker, 1981; Coleman, 1990). In many contexts, this assumption facilitates the analysis of social behaviour. However, theories of rational choice have been criticised, especially for the requirement of extensive information and cognitive skills on the part of the actors (e.g. Simon, 1955; Gigerenzer et al., 1999; Bendor, 2001).

ACCEPTED MANUSCRIPT

Next to theories of rational choice, behavioural psychology has provided a basis of social theories. Well-known sociologists such as George C. Homans (1961), Richard M. Emerson (1972), and many others (Burgess and Bushell, 1969; Hamblin and Kunkel, 1977) employed principles from behavioural psychology in the study of social phenomena. While some ideas remain in contemporary theories of social exchange (Molm, 2006, p. 29) and backward-looking rationality (Macy and Flache, 2009, pp. 250-251), systematic treatments of these principles are presently missing.

In particular, Homans (1974, pp. 21-22) suggested that the **matching law** (Herrnstein, 1997) could serve as foundation of individual behaviour in sociological theories. As elaborated in section 2, the matching law describes a relatively simple empirical regularity of individual decision-making. It has been observed in a variety of psychological experiments (Herrnstein, 1961; de Villiers and Herrnstein, 1976; Baum, 1979; Pierce and Epling, 1983; McDowell, 2005) and social situations (e.g. Conger and Killeen, 1974; Hamblin, 1977, 1979; Sunahara and Pierce, 1982; McDowell, 1988; Borrero et al., 2007).

Besides a few exceptions (Gray and von Broembsen, 1976; Gray et al., 1982), the matching law has not been used to derive macro-level phenomena. This paper is an attempt to fill this gap and investigates the possibility of applying the matching law as micro-foundation of social theory. As further explained in section 3, the main problem appears when making the transition from the individual to the social level. Since the matching law describes aggregated individual behaviour, an application as micro-foundation requires an additional mechanism of decision-making that allows predictions at particular points in time.

Multiple mechanisms exist that both imply the matching law on the individual level and allow for the derivation of testable predictions on the social level.

ACCEPTED MANUSCRIPT

In section 4, a simple model of **melioration learning** is argued to fulfil these requirements. Even though this model is not entirely realistic in reflecting human behaviour, it serves as starting point when analysing interactive situations and shows what can or cannot be explained without applying more complex models.

Furthermore, a link to the extensive research on reinforcement learning methods in the computer sciences (Sutton and Barto, 1998; Wiering and van Otterlo, 2012) is accomplished by implementing melioration as ε-greedy selection with Q-learning (Watkins, 1989). This algorithm is one of the most basic methods of estimating reward functions if the underlying probabilities of reinforcement are unknown (van Otterlo and Wiering, 2012, p. 31).

Section 5 illustrates the proposed solution by applying the model to the repeatedly played volunteer's dilemma. Given the basic (symmetric) version of this dilemma, the results correspond to the predictions of the Nash equilibrium. But in its asymmetric version, the matching law leads to more intuitive results than the mixed Nash equilibrium.

In section 6, the melioration model and the application of the matching law in social situations are further discussed. Finally, by following earlier work of Rachlin et al. (1976), the matching law is integrated into economic consumer theory in section 7. This approach avoids the definition of the matching law as tautology and enables a comparison to the classic economic solution of utility maximisation.

In sum, this paper argues for the usage of a simple and falsifiable model of individual decision-making in social theory. It integrates psychological with sociological research by the implementation of a reinforcement learning algorithm from the computer sciences. In simple situations, the present model of melioration learning implies the matching law as empirical regularity of individual behaviour.

By means of computer simulations, it can also be applied to many social situations of strategic interdependence. It is demonstrated that, when applied to a particular problem, it generates testable predictions that more accurately correspond to empirical findings than other theories.

2 The matching law

Past psychological research has found that individual behaviour often conforms to a simple equation called the matching law (e.g. Baum, 1979; Herrnstein, 1997; McDowell, 2013a). This law is supposed to hold in any situation of repeated decision-making if the choices are regularly accompanied by some form of reinforcement. A reinforcement is any kind of consequence that is deemed valuable by the individual. More specifically, if an actor repeatedly decides between two alternatives, the matching law predicts a matching of the ratio of choices to the ratio of reinforcements in the long run:

$$\frac{k_1}{k_2} = \frac{s_1}{s_2}, \tag{1}$$

where k_1, k_2 denote the absolute frequencies of choices and s_1, s_2 the corresponding absolute frequencies of reinforcement.

This strict version of the matching law requires that reinforcements are approximately equal in value. An illustrative example is the penalty kick situation of (European) football games. The kicker can be assumed to choose the left or right side of the goal. The value of a reinforcement, which is a scored goal, is independent of the kicker's choice. Table 1 contains a hypothetical distribution of

choices and reinforcements for a player who was engaged in 50 penalty kicks.

choice of kicker	left (k_1)		right (k_2)	
	40		10	
reinforcement	success (s_1)	failure	success (s_2)	failure
	24	16	6	4

Table 1: A sample distribution of choices in penalty kick situations

According to equation (1), the strict matching law holds in this example because

$$\frac{k_1}{k_2} = \frac{40}{10} = 4 = \frac{24}{6} = \frac{s_1}{s_2}.$$

In spite of many empirical confirmations, the strict rule of equation (1) has occasionally failed to describe behaviour in psychological experiments. Two of the main systematic deviations are known as bias and under-/overmatching. Both kinds of deviations have been accounted for by extending the strict matching law to the generalised matching law (e.g. Baum, 1974; McDowell, 2013a):

$$\frac{k_1}{k_2} = \beta \cdot \left(\frac{s_1}{s_2}\right)^{\alpha}. \tag{2}$$

Equation (2) contains two free parameters $\alpha, \beta \in (0, \infty)$. In the psychological literature, the matching law is said to hold if there exist $\alpha, \beta \in (0, \infty)$ such that condition (2) is true. Consequently, the matching law has been seen as a set of infinitely many equations.

Figure 1 illustrates the range of relations between the behaviour ratio $\frac{k_1}{k_2}$ and the reinforcement ratio $\frac{s_1}{s_2}$ that is captured by the generalised matching law. The diagonals indicate strict matching ($\beta = \alpha = 1$). The left-sided graph contains the generalised matching law with $\alpha = 1$ and different values of β. In the right-sided

graph, β is set to 1, and α is varied between 0.1 and 10.

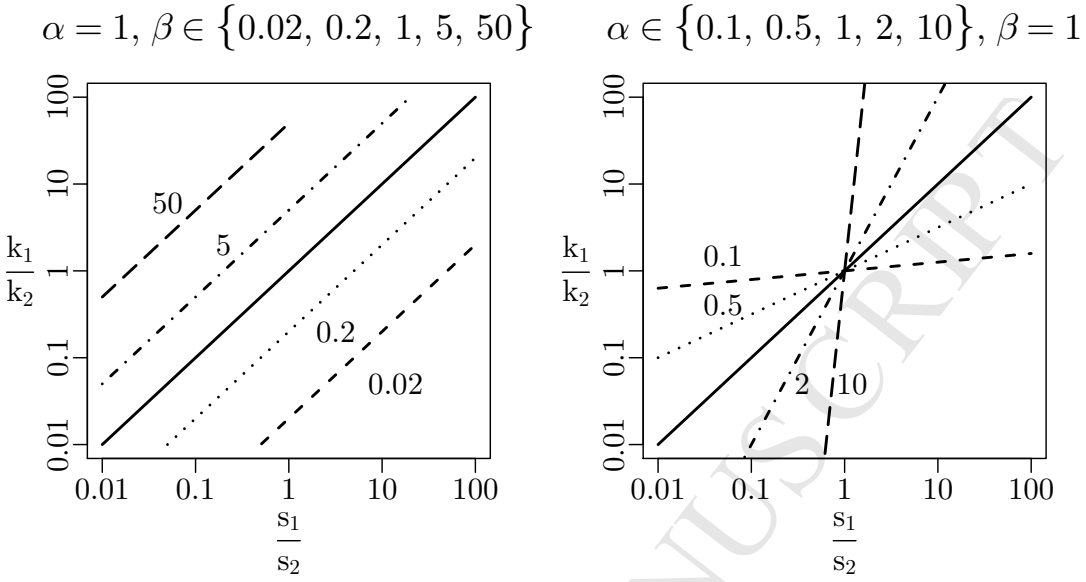

Figure 1: The generalised matching law as given by equation (2) with different parameter settings

By setting $\beta \neq 1$, the generalised matching law captures bias. This is a systematic preference for one of the alternatives (Baum, 1974). As pictured by the left-sided graph of figure 1, the first alternative is chosen more often than predicted by the strict matching law if $\beta > 1$ $\left(\frac{k_1}{k_2} > \frac{s_1}{s_2}\right)$. If $\beta < 1$ $\left(\frac{k_1}{k_2} < \frac{s_1}{s_2}\right)$, the second alternative appears more frequently. As summarised by Baum (1974), bias mostly originates from the presence of asymmetries in choice alternatives or reinforcements. For example, if the choice of one alternative requires additional effort or if the reinforcements differ in their amount or quality, an individual prefers one alternative over the other.

The second parameter of equation (2) accounts for two systematic deviations that have been called undermatching and overmatching (Baum, 1979). In case

ACCEPTED MANUSCRIPT

of overmatching, alternatives with high relative frequencies of reinforcement are chosen more often than predicted by the strict matching law. In the right-sided graph of figure 1, this is modelled by $\alpha > 1$. In contrast, undermatching ($\alpha < 1$) stems from a systematic preference for the less reinforced alternative.

In experimental studies, undermatching is observed more often than overmatching. For example, undermatching occurs if a frequent switching between the alternatives is possible (Baum, 1979). Another reason of undermatching is the presence of interrelated deprivation rates (Baum and Nevin, 1981; Green and Freed, 1993). If the consumption of one resource increases the demand for another resource (e.g. the consumption of food may increase the demand for water), an increase in reinforcement of one alternative raises the frequency of choosing the other one. Moreover, since undermatching implies that low relative frequencies of choice are adjusted to somewhat higher levels, this behaviour might be interpreted as experimenting (see e.g. Vollmer and Bourret, 2000; Kangas et al., 2009).

3 The matching law as micro-foundation

When attempting to establish the matching law as micro-foundation of social phenomena, the transition between micro- and macro-level must be solved. Figure 2 illustrates the setting in reference to Coleman's (1990, p. 646) micro-macro scheme. As explained in the following, the matching law is regarded as the outcome of repeatedly applying a rule of decision-making.

A first problem appears when using the generalised matching law of equation (2) instead of the strict matching law of equation (1). The parameters $\alpha, \beta \in (0, \infty)$ must be specified. While, in the psychological literature, the parameters

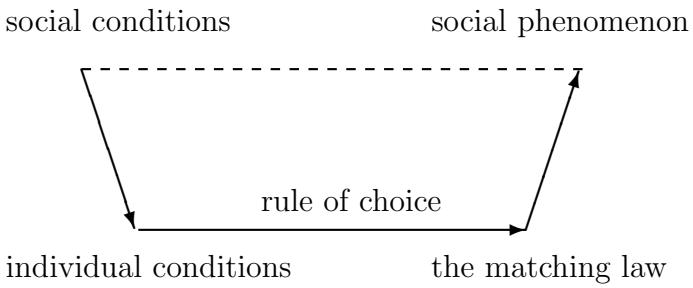

Figure 2: The matching law in Coleman's micro-macro scheme

are usually chosen empirically as best fit to the data, this approach runs the risk of defining a tautology (Rachlin, 1971). A solution is the theoretical derivation of the parameters from social and individual conditions of choice. This issue is further discussed in section 7.

More severely, even if the parameters $\alpha, \beta \in (0, \infty)$ are given and the total number of choices $k_1 + k_2$ is fixed, multiple combinations of (k_1, k_2, s_1, s_2) are possible for equation (2) to hold. Consequently, it is not possible to predict an outcome from the matching law itself. Instead, a rule of decision-making that leads to a particular combination of (k_1, k_2, s_1, s_2) is needed.

Another problem is that the matching law cannot be directly mapped to a social phenomenon. Since the relevant variables (k_1, k_2, s_1, s_2) are aggregated over a period of time, the matching law prevents the prediction of decisions at a certain point in time and, thus, the derivation of a social phenomenon.

For example, the previously mentioned penalty kicks are actually social situations of strategic interdependence (e.g. Palacios-Huerta, 2003). Both the kicker and the goal keeper must decide concurrently between the left and right side of the goal. The former decides where to kick the ball. The latter decides where to jump in order to block the ball. For a given situation of repeated penalty kicks,

the matching law predicts a matching of choice and reinforcement ratios for both players separately, i.e.

$$\frac{k_{K1}}{k_{K2}} = \frac{s_{K1}}{s_{K2}} \text{ and } \frac{k_{G1}}{k_{G2}} = \frac{s_{G1}}{s_{G2}},$$

where K denotes the kicker and G the goal keeper. Since the matching law does not predict the decisions of the players at a particular penalty situation, nothing can be said about the outcome of this situation (e.g. whether both players choose the left side of the goal) or how this social outcome changes over time.

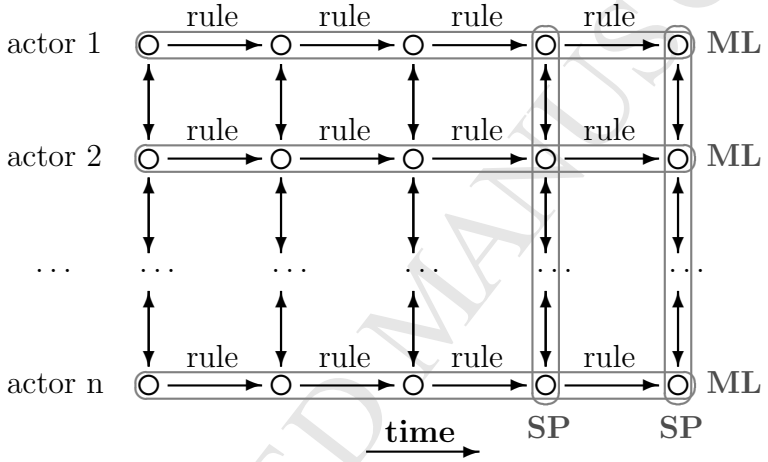

Figure 3: Diagram of repeated interactions among multiple actors; the matching law (ML) denotes a temporal average over a sequence of choices; a social phenomenon (SP) is commonly measured as cross-sectional outcome

For the general case of more than two actors, this point is clarified in the diagram of figure 3. It shows the interaction between multiple actors over time. Each actor is assumed to choose repeatedly one of several alternatives by following some rule. The circles stand for points in time. A horizontal arrow from one circle to another displays a choice. As indicated by the vertical arrows, the actors may influence each other in their decisions. The horizontally stretched ovals (ML) mark the fact that the matching law refers to aggregated measures of a sequence of deci-

sions. For instance, the matching law may state that actor 1 chooses a particular action in 20% of the time. However, a social phenomenon (SP) commonly refers to a cross-sectional or longitudinal measurement, e.g. how many actors emitted a certain action at one or several points in time. Therefore, it corresponds to one or several ovals that vertically stretch over the circles of the diagram. Apparently, a vertically recorded social phenomenon is not compatible with the horizontally spreading matching law. A macro-level derivation requires a set of rules that directly specify individual choices.

4 The model of melioration learning

As argued in the previous section, the matching law by itself cannot be used to derive a social phenomenon. Since it denotes an individual's temporal average, no prediction is made about the cross-sectional average of a group of actors. A solution is provided by a mechanism of decision-making that identifies the particular actions. At the same time, this rule should result in the matching law at the individual level. In other words, it should lead to values (k_1, k_2, s_1, s_2) that correspond to equation (2) in the long run. Multiple rules exist that meet these requirements. With melioration learning, a very simple one is analysed in this paper.

4.1 Melioration learning

Melioration learning was introduced as explanation of the matching law by Vaughan and Herrnstein (1987). In regard to the model of section 2, it states that the frequency of choosing alternative 1 increases if it comes with the currently highest average value. The average value may be as simple as its success rate $\frac{s_1}{k_1}$. But also

the subjective utility of reinforcements can be captured by melioration learning and the matching law, respectively (see section 7).

More specifically, melioration predicts that the relative frequency $\frac{k_1}{k_1+k_2}$ of having chosen alternative 1 increases if $\frac{s_1}{k_1} > \frac{s_2}{k_2}$, and decreases if $\frac{s_1}{k_1} < \frac{s_2}{k_2}$. Herrnstein (1990a, p. 219) states that

> "[t]he melioration process continues until the stronger response displaces all others, or, because the reinforcement returns from an alternative may depend on its level of occurrence, equilibrium is attained with several alternatives left in the response set, each yielding the same returns per unit at a given allocation among them [..]."

While several empirical confirmations of melioration exist (Vaughan, 1981; Mazur, 1981; Vaughan and Herrnstein, 1987; Herrnstein et al., 1993; Antonides and Maital, 2002; Tunney and Shanks, 2002; Yechiam et al., 2003; Neth et al., 2005), it is generally regarded as too simple to account for real human behaviour (e.g. Shteingart and Loewenstein, 2014). More appropriate learning models have been suggested in the past (see section 6). But melioration may be sufficiently accurate when predicting social phenomena instead of individual behaviour. Especially in complex social interactions, the application of simple rules of choice enables the understanding of the underlying mechanisms, leads to clearer predictions and, hence, supports the empirical test of the hypotheses.

4.2 The model

In the past, mathematically rigorous representations of melioration learning have been suggested (e.g. Brenner and Witt, 2003; Loewenstein, 2010). In line with

Neth et al. (2006) and Gureckis and Love (2009), this paper employs an algorithm that is perfectly consistent with the ideas of Vaughan and Herrnstein (1987) and builds on a well-established algorithm of reinforcement learning. More precisely, melioration is implemented as instance of Q-learning with ε-greedy selection.

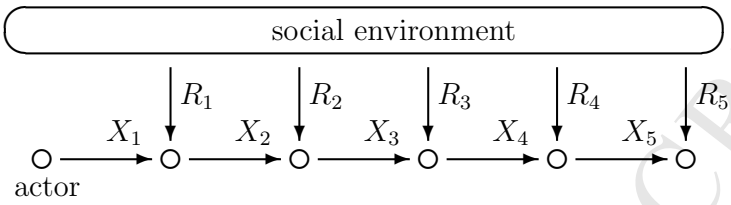

Figure 4: The situation of sequential decision-making

Q-learning was introduced by Watkins (1989) and is a popular algorithm of artificial intelligence (Sutton and Barto, 1998; van Otterlo and Wiering, 2012). Given a finite set of choice alternatives E and a reward sequence $(R_t)_{t=1}^{\infty}$ with values in $[0, \infty)$, it defines a choice sequence $(X_t)_{t=1}^{\infty}$ with values in E. Figure 4 illustrates this process. An actor is seen as repeatedly emitting an action $X_t \in E$ and awaiting a response R_t. The decisions take place along discrete time steps $t \in \mathbb{N}$. In the model of section 2, the response R_t is either 0 or 1 indicating whether a reinforcement occurs. But in general, R_t can return any positive real value that stands for the subjective utility of a reinforcement (see section 7). In the following, the more general case is assumed.

In combination with ε-greedy selection, Q-learning describes behaviour that myopically optimises upcoming rewards and explores alternatives from time to time. More specifically, the actor is assumed to maintain a set of Q-values $\{Q(j)\}_{j \in E}$ and a set of frequencies $\{K(j)\}_{j \in E}$ (for details, see the algorithm in the appendix). Initially, all values are set to zero. At every round $t \in \mathbb{N}$, an alternative $X_t = e \in E$

is chosen randomly with probability $\varepsilon \in (0,1)$ or greedily otherwise. Greedy choice means that the alternative with the currently highest Q-value is selected. If multiple alternatives have the highest value, one of them is chosen randomly. Afterwards, the corresponding frequency $K(e)$ is increased by one and, given the reward $R_t = y$, the Q-value $Q(e)$ is updated by the following rule:

$$Q(e) \leftarrow Q(e) + \frac{1}{K(e)} \cdot (y - Q(e)) \,. \tag{3}$$

If $y_1, y_2, \ldots, y_{K(e)} \in [0, \infty)$ denote the values of all rewards that were received for an action $e \in E$ until a certain point in time, $\frac{1}{K(e)} \sum_{l=1}^{K(e)} y_l$ is the average value of this action. Given equation (3), it holds that

$$Q(e) = \begin{cases} \frac{1}{K(e)} \sum_{l=1}^{K(e)} y_l & \text{, if } K(e) > 0, \\ 0 & \text{, if } K(e) = 0. \end{cases}$$

Consequently, a Q-value $Q(e)$ gives the average reward of alternative $e \in E$. If $y_1, y_2, \ldots, y_{K(e)} \in \{0, 1\}$, this value corresponds to the success rate $\frac{s_e}{k_e}$.

As long as an actor chooses an action with the currently highest Q-value, the relative frequency of this action increases as required by Vaughan and Herrnstein (1987). Therefore, updating the Q-values by equation (3) and always choosing the action with the highest Q-value resembles melioration learning. The main difference between the model and the ideas of Vaughan and Herrnstein (1987) is the exploration rate ε. The maximally exploiting strategy of greedily selecting an action with the highest Q-value has the disadvantage of exclusively choosing a single alternative as soon as it has been reinforced by a strictly positive reward. A trade-off between the exploitation of the currently best actions and the exploration

ACCEPTED MANUSCRIPT

of other actions is made by introducing some level of erratic behaviour (Sutton and Barto, 1998, p. 28).

Furthermore, the exploration rate is a realistic assumption about human behaviour, which naturally includes mistakes, misinterpretations, and the occasional trial of alternative behaviour (e.g. Selten, 1975; Bendor, 1987). Multiple studies have shown that random perturbations affect social outcomes (Macy and Tsvetkova, 2015). In particular, it was argued that success depends on fine-tuning the trade-off between exploration and exploitation (March, 1991). While no exploration undermines the adaptation to changes in the environment and compromises long-term prosperity, high exploration misses out on the gains of emitting the currently best actions. It is, therefore, conceivable that humans maintain a small but effective level of exploration that balances environmental changes and helps to learn profitable behavioural regularities.

Finally, it was claimed in the introduction that the proposed model of melioration learning leads to the matching law. In proposition 1 (see appendix), this is shown to hold under certain conditions of stationarity. In general, it must be assumed that, during the process of learning, each Q-value $Q(e)$ converges to some fixed values $Q^*(e)$. For example, this condition holds if the expected values of the rewards R_t depend only on action X_t and no other previous actions. In many social situations, those kinds of constraints cannot be ensured. This impedes the analytical derivation of social phenomena. The model can still be applied to strategic situations by means of computer simulations.

ACCEPTED MANUSCRIPT

5 Application to the volunteer's dilemma

This section illustrates the proposed approach of how to explain a social phenomenon by the matching law. The purpose of this endeavour is to derive empirical predictions for a particular situation of strategic decision-making. In the following, the volunteer's dilemma is considered. In the case of repeated interactions, adequate empirical studies of this situation have not been conducted yet (see section 5.4). Nevertheless, in some situations, the melioration learning model more accurately conforms to results from earlier experiments than predictions that follow from the mixed Nash equilibrium, which is prominent in the respective literature.

Since the matching law cannot be employed directly, melioration learning is applied instead. An advantage of the underlying Q-learning algorithm is its applicability to a wide range of situations. While it has already been analysed in two-actor settings (Wunder et al., 2010; Kianercy and Galstyan, 2012), studies of situations with more than two actors are largely absent.

The volunteer's dilemma (Diekmann, 1985) captures a range of sociologically relevant situations. In particular, it includes the problem of the "diffusion of responsibility" (Darley and Latané, 1968), in which a single person loses his readiness to volunteer if other candidates are present. The dilemma arises, for example, in cases of emergencies or accidents if one of several persons could provide first aid or call for help (Diekmann, 1985).

More specifically, the volunteer's dilemma represents a situation with $n \in \mathbb{N}$ actors ($n > 1$), each of whom must decide between volunteering or being idle. A collective good is provided as soon as one member of the group volunteers. While this results in a utility $u \in (0, \infty)$ for everyone, the act of volunteering entails a

		number of other volunteers				
		0	1	2	...	$n-1$
single actor	volunteer	$u-c$	$u-c$	$u-c$	$u-c$	$u-c$
	be idle	0	u	u	u	u

Table 2: The representation of the volunteer's dilemma from the perspective of a single actor; all actors must choose between volunteering or being idle

cost $c \in (0, u)$. As shown in table 2, an actor obtains utility $u - c$ if she volunteers and u if someone else does. In case of no volunteer, everyone receives zero utility.

When using game theory for analysis, the volunteer's dilemma has n pure Nash equilibria in which exactly one actor volunteers. Additionally, there is a mixed Nash equilibrium with the probability of volunteering given by $p = 1 - \left(\frac{c}{u}\right)^{\frac{1}{n-1}}$, for each actor (Diekmann, 1985, p. 607).

5.1 Learning to volunteer

In order to derive predictions about the actors' behaviour, particular examples of the volunteer's dilemma were analysed by computer simulations. A total of 10 000 actors were divided into groups of size $n \in \{2, 3, \ldots, 10\}$. All actors employed the melioration learning model with $\varepsilon = 0.1$. The utility of the collective good was set to $u = 10$. The costs of volunteering c was varied between 1 and 7.

Each simulation lasted for 1 000 time steps, which means that every actor encountered 1 000 instances of the volunteers dilemma. While, in real life, situations that resemble this dilemma occur rarely, the predictions can still be applied to predict human behaviour. First, most of the simulations have already converged after

a couple of hundred interactions. Second, humans can be assumed to learn faster than modelled by melioration. The simplicity of the model is, thus, compensated by a longer period of learning.

In a first set of simulations, the groups of actors were regarded as fixed. This means that the same actors interacted repeatedly with each other in situations of the volunteer's dilemma. This setting corresponds to any association, department, or group of friends that regularly needs a volunteer to, for example, complete a task or organise an annual party.

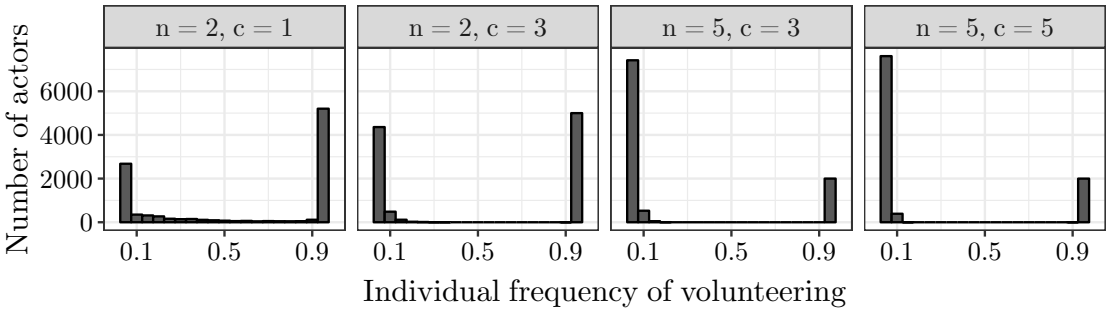

Figure 5: Histograms over individual long-term frequencies of volunteering; for each actor, the frequency of volunteering is calculated over the whole period of 1 000 interactions; results are shown for four simulations with fixed groups

The analysis revealed that the actors learn to coordinate their choices to a state in which exactly one actor per group volunteers. This corresponds to the pure Nash equilibrium and is illustrated by four sample histograms in figure 5. The x-axis depicts the individual frequency of volunteering aggregated over all 1 000 choices during one simulation run. Each panel depicts an entire simulation, and the bars add up to 10 000 actors. For instance, in the first two simulations with $n = 2$, approximately 5 000 actors almost always volunteer with a relative frequency of volunteering above 0.9. The remaining actors are mostly idle. Due to the group

size of $n = 2$, this means that one actor per group volunteers, and the other one is idle. According to the melioration model, neither actor has incentive to switch to the other alternative. The actor who volunteers would receive a reward of 0 instead of $u - c$. The utility of the idle actor would fall from u to $u - c$. The same is observed in groups of size $n = 5$. Exactly one actor per group volunteers, which makes up a total of 2 000 actors with a high relative frequency of volunteering.

Since there is exactly one permanent volunteer in each group, melioration learning produces optimal behaviour in simulations with fixed groups. Accordingly, the population-wide rate of volunteering is $\frac{1}{n}$ and, thus, decreases hyperbolically with n. This is seen in the left-sided plot of figure 6. The relative frequency of volunteers is largely independent of the costs c. Even though the benefit of volunteering $(u - c)$ decreases with increasing costs, the single volunteer benefits from her decision because other actors volunteer very seldom. Furthermore, since there is one volunteer per group, the relative frequency of groups without volunteers is very low in most simulations. This is seen in the right-sided plot of figure 6.

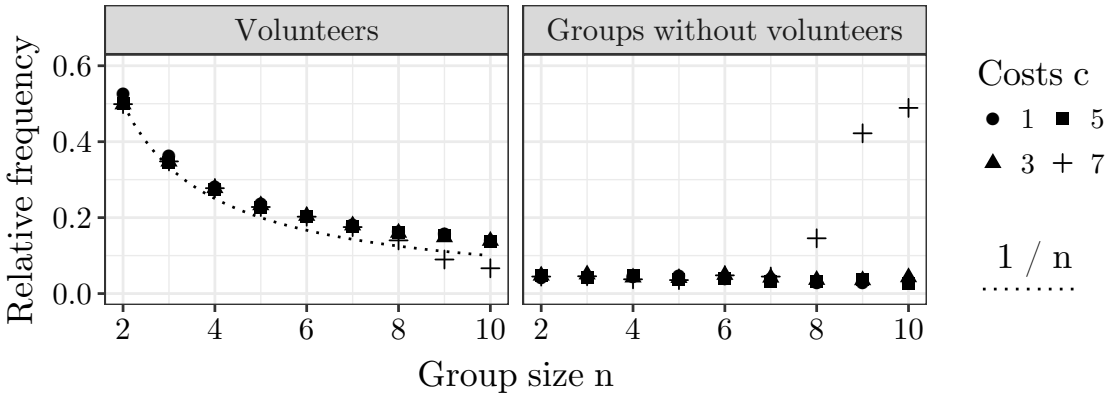

Figure 6: The effect of group size and costs on the population-wide rate of volunteers; groups are fixed; also the rate of groups without volunteers is depicted

A deviation from the inverse of the group size $\frac{1}{n}$ appears if $c = 7$ and $n \geq 8$. Volunteering drops more rapidly, which means that no permanent volunteer exists and many groups end up without volunteer. This effect is due to the exploration rate ε. When following the melioration learning model, the actors try different actions with probability $\varepsilon = 0.1$. The probability of at least one actor volunteering by exploration is given by $1 - \left(1 - \frac{\varepsilon}{2}\right)^{n-1}$. It follows that the expected value of being idle is at least $u \cdot \left(1 - \left(1 - \frac{\varepsilon}{2}\right)^{n-1}\right) = 10 \cdot (1 - 0.95^{n-1})$. With $c = 7$ and $n \geq 8$, this value is greater than the reward of volunteering on purpose ($u - c = 3$). Consequently, every actor obtains a higher reward from being idle than from volunteering, and no permanent volunteer is established. In other words, melioration leads to a situation with only erratic volunteering if $u - c$ is low and n is high.

5.2 Anonymous groups

Next to fixed groups, the simulations were repeated with anonymous groups. In this setting, the actors interacted with different partners at each round. Because the group composition changes continuously, the actors cannot coordinate their decisions to an equilibrium with a single permanent volunteer. Each actor has to choose an alternative without any experience on the decisions of the other actors.

The simulations show that, while some actors still choose one alternative exclusively, the results are considerably different to the ones of fixed groups. Related to figure 5, figure 7 exhibits histograms over the individual relative frequencies of volunteering. In contrast to the previous figure, each bar consists of one histogram with three bins of different width. For example, the first bar of the first chart of

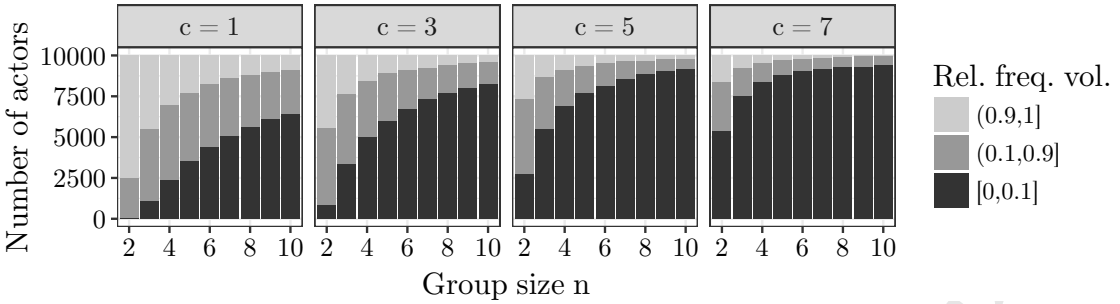

Figure 7: Histograms over individual long-term frequencies of volunteering in the simulations with anonymous groups; each bar depicts one histogram with three bins of different width: $[0, 0.1], (0.1, 0.9], (0.9, 1]$

figure 7 corresponds to the first chart of figure 5 ($n = 2$ and $c = 1$). The bin-width is chosen in order to roughly distinguish between three types of actors: actors who mostly volunteer (the interval $(0.9, 1]$), actors who almost never volunteer (the interval $[0, 0.1]$), and the ones who occasionally volunteer (the interval $(0.1, 0.9]$). The plots show the size of each proportion.

Similar to the previous setting with fixed groups, the majority of actors is either mostly volunteering or mostly idle. In opposition to the former results, the fraction of actors who mostly volunteer decreases with costs c. Especially, if c is large, the fraction of permanent volunteers is below $\frac{1}{n}$.

Averaged over the whole population, the individual level of volunteering decreases with group size in correspondence with the numeric predictions of the mixed Nash equilibria. In figure 8, the simulation results are plotted as points, and the predictions of the Nash equilibrium are drawn as lines. It is also seen that the number of actors who volunteer decreases with the costs of volunteering.

In contrast to fixed groups, a cost effect appears because the chance of encountering another volunteer in the next dilemma situation depends on the overall number of volunteers in the population. More specifically, the difference between

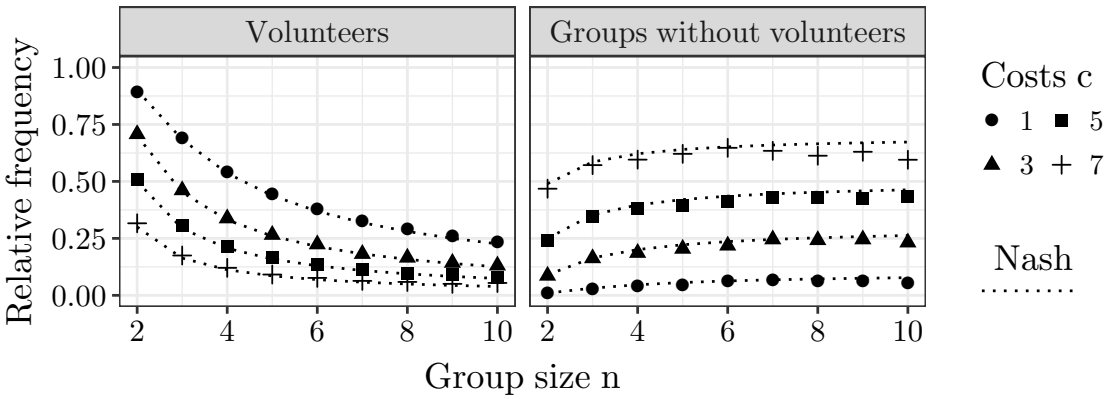

Figure 8: The effect of group size and costs on the population-wide rate of volunteers; anonymous groups; also the rate of groups without volunteers is depicted

the expected rewards of volunteering and being idle is

$$(u - c) - p_v \cdot u, \tag{4}$$

with p_v being the probability of another actor volunteering. In fixed groups, the probability p_v is basically stable at $1 - \left(1 - \frac{\varepsilon}{2}\right)^{n-1}$ from the perspective of the single permanent volunteer. For the other actors, $p_v \approx 1$. Unless the costs c and the group size n are too high (see example above), the difference (4) is always positive for the single volunteer and negative for the other actors.

In the anonymous setting, on the other hand, p_v depends on the number of volunteers in the population. The reward of being idle ($p_v \cdot u$) increases if many volunteers exists and decrease if only few actors volunteer. An actor changes to volunteering if p_v is low and to being idle if p_v is high. At the equilibrium point of

$$(u - c) = p_v \cdot u, \tag{5}$$

ACCEPTED MANUSCRIPT

the actors mainly stick to their choices. Since this point depends on the costs c, the overall rate of volunteers changes with c. Furthermore, condition (5) corresponds to the mixed Nash equilibrium, which explains the match between melioration learning and that solution in the present setting.

As seen in figure 7, the actors are either always idle or occasional volunteers if costs are high. Therefore, the number of volunteers is not sufficient to obtain at least one volunteer per group. This is pictured in the right-sided plot of figure 8. More than half of the groups are without a volunteer if c is high. But also in the case of low costs, this rate is greater than in fixed groups.

5.3 The asymmetric volunteer's dilemma

The volunteer's dilemma is a highly simplified representation of real situations. Generally, it cannot be assumed that the costs of volunteering is the same for all actors. The basic version of the dilemma is easily extended to an asymmetric version in which the actors differ in their strength and, hence, in costs of volunteering. In the following simulations, a strong and a weak actor type is assumed. At first, the costs of volunteering are cut in half for the stronger ones. According to the mixed Nash equilibrium, the probability of a strong actor being idle is twice as high as the corresponding probability of weak actors (Diekmann, 1993, p. 77, eq. 4). This rather counter-intuitive hypotheses does not match empirical findings (Diekmann, 1993). A pure Nash equilibrium would be a more plausible solution, especially if a strong actor is the single volunteer.

The simulations were run with anonymous groups and half of the actors being strong (results of simulations with fixed groups are similar, as shown in figure 12

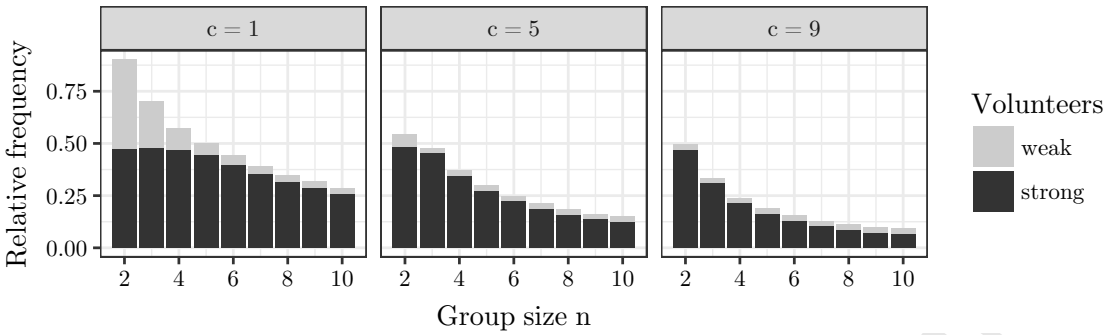

Figure 9: The effect of group size in an asymmetric volunteer's dilemma; the costs of volunteering for strong actors are $0.5c$; 50% of the actors are strong

in the appendix). In contrast to the mixed Nash equilibrium, melioration learning predicts that strong actors are more likely to volunteer than weak actors (see figure 9). In groups of size two, all the strong actors and some weak ones volunteer. With increasing group size, the rate of volunteers decreases at a similar rate as in the symmetric version. But the weak actors cease to volunteer first.

The observation that strong actors are more likely to volunteer is in line with the previous finding of volunteering decreasing with its costs. If an actor is strong and costs are low, the marginal benefit of volunteering $(u - c) - p_v \cdot u$ is positive for a greater range of probabilities p_v. Consequently, the threshold of p_v for a weak actor to be idle is generally lower than this threshold for strong actors. As n increases, the probability p_v of encountering another volunteer increases. As soon as the threshold of weak actors is passed, they change to being idle. This takes place before the threshold of the strong ones is reached.

Furthermore, in a stable state, it is conceivable that $(u - c) > p_v \cdot u$ for some, mainly strong, actors and $(u - c) < p_v \cdot u$ for other actors. This conflicts with the Nash equilibrium, which requires an equality $(u - c) = p_v \cdot u$. In order to achieve

this equality, p_v must be greater for strong than for weak actors. In other words, strong actors must be less likely to volunteer than weak ones for the mixed Nash equilibrium to hold. In contrast, melioration learning predicts the more intuitive result of strong actors bearing the costs of volunteering.

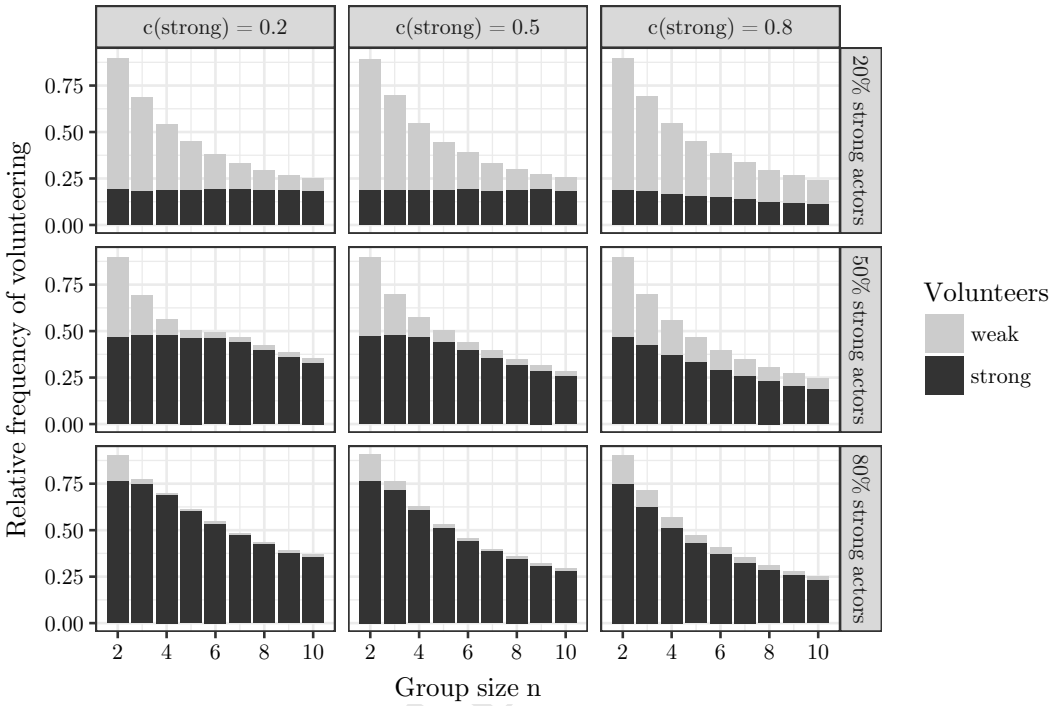

Figure 10: The effect of group size in the asymmetric volunteer's dilemma; $c = 1$; the strength and the proportion of strong actors vary between 0.2 and 0.8

In the simulations shown in figure 10, the proportion and strength of the strong actors are varied. The latter ranges between $c(strong) = 0.2$ (strongest actors with lowest costs) and $c(strong) = 0.8$ (less strong but still lower costs than weak actors). The costs of weak actors are fixed at $c = 1$. If only 20% of the population is strong, all of them volunteer independent of their strength and group size. In the case of 80% strong actors, almost no weak actor has to volunteer and the burden of volunteering is borne by the stronger ones.

ACCEPTED MANUSCRIPT

5.4 Comparison with empirical findings

In empirical studies of the basic, symmetric, version of the volunteer's dilemma, the individual likelihood of volunteering was observed to decrease with group size (e.g. Darley and Latané, 1968). This finding is in line with the previous simulations because they reproduce the mixed or the pure Nash equilibrium. Both concepts explain a negative relationship between group size and volunteering.

Nevertheless, the empirically observed decline in the relative frequency of volunteers is generally not as sharp as implied by the mixed equilibrium or the inverse of the group size (see for example Franzen, 1995, or Goeree et al., 2017). This also means that the predictions of melioration are incorrect. But this conclusion may be premature. There are some caveats when comparing the simulation results of the previous sections to experimental studies. For instance, in Franzen (1995), the actors interacted only once in a volunteer's dilemma. In this case, learning is limited and takes place exclusively during the instructions of the experiment.

In the experiments of Goeree et al. (2017), subjects participated in 20 consecutive rounds of the volunteer's dilemma. Some learning can be assumed. Nevertheless, the predictions of melioration diverge from the results of the experiments. This might be due to differences between the experimental setting and the theoretical assumptions. More specifically, while the actors of the simulations were not aware of any other group member or the table of payoffs, the participants of the experiment were more thoroughly informed about the situation. It can be suspected that the latter used this information in order to coordinate their choices. If the subjects of the experiment were not aware of the structure of the situation, the findings might have corresponded to the predictions of the simulations.

ACCEPTED MANUSCRIPT

6 Discussion

The model of melioration learning is shortly discussed in this section. In particular, its empirical validity is assessed, and alternative models are mentioned. Melioration has been categorised as model of reinforcement learning (Brenner, 2006). Reinforcement learning refers to the simple idea of behaviour being more likely to reoccur if followed by a positive experience and diminishing over time if provoking negative reactions. Two other models of reinforcement learning that have been used in the social sciences are the Bush-Mosteller and the Roth-Erev model (Roth and Erev, 1995; Skyrms and Pemantle, 2000; Flache and Macy, 2002).

Moreover, an active research field of the computer sciences engages in the analysis of reinforcement learning (called RL; Sutton and Barto, 1998). The melioration algorithm of section 4.2 constitutes a relatively trivial instance of the Q-learning method, which is, in turn, just one of several RL methods.

As pointed out above, melioration learning is able to account for some of the observed behaviour in situations of repeated choice (see also Sakai et al., 2006, p. 1092). But generally, there is "tremendous heterogeneity in reports on human operant learning" (Shteingart and Loewenstein, 2014, p. 94). In particular, melioration has often been regarded as too simple to accurately represent the complexity of human decision-making (Barto et al., 1990, p. 593). Even experiments with animals revealed that changes in behaviour occur more rapidly than predicted by melioration learning (Gallistel et al., 2001; Sugrue et al., 2004). This indicates that, instead of long-term averages such as the Q-values, subjects maintain a temporally local representations of reinforcement rates, which include only the most recent outcomes. Other learning models such as the one of Corrado et al. (2005)

account for rapid changes. The latter "differs from melioration with respect to both the quantity that drives behavioural change and the temporal window over which that quantity is computed" (Corrado et al., 2005, p. 611).

Further models of reward-driven behaviour are listed in Sakai et al. (2006). All of them exhibit the matching law in their steady states. For example, Sakai and Fukai (2008a) argued for the usage of actor-critic learning because it exhibits the matching law in steady states and no direct representation of the average values (Q-values) is needed.

A somewhat different decision rule was proposed by McDowell (2013b). The author modelled learning as an evolutionary process within the individual. A "population" of different behavioural alternatives is assumed, and every decision is followed by a "reproduction" of successful actions. The relative number of an alternative in the population resembles its probability of choice. In a series of simulations, McDowell and colleagues showed that this model leads to matching behaviour (McDowell, 2004; McDowell and Caron, 2007; McDowell et al., 2008; McDowell and Popa, 2010). Also behaviour that deviates from the matching law was reproduced if this behaviour had been observed in laboratory experiments.

In summary, the empirical status of melioration learning is disputed, and alternative models of learning have been suggested. Because most of the reinforcement learning models entail the matching law in their steady states, there is no theoretical reason to prefer one model over the others. If the most realistic representation is wanted, only a neural network model (e.g. Loewenstein and Seung, 2006) might be appropriate because it is closest to the physical basis of human decision-making. However, the implementation of neural networks is demanding, and the mechanisms are difficult, if not impossible, to comprehend.

The advocated model of melioration has several advantages. First, in contrast to other models of melioration, it closely corresponds to the original formulation of Vaughan and Herrnstein (1987), which stated that the relative frequencies, and not the probabilities of choice, change in accordance with the differences in reinforcement. Second, the model omits probabilities of choice. Therefore, it is not necessary to assume that humans behave stochastically (apart from the exploration rate). Third, Q-learning is one of the most popular and widely studied RL techniques. Many results about its convergence properties already exist and can be appropriated for an application in social theory. Fourth, Q-learning is one of the simplest RL methods, and simplicity is a desired property of simulation models (Axelrod, 1997, p. 18). The rules of decision-making should be kept simple in order to ease the understanding of the results and to reduce the time of computation.

Finally, although experiments revealed deviations from the melioration model on the individual level, its predictions might be sufficiently accurate on a social level. Only if deviations are observed on the social level, the introduction of more advanced behavioural assumptions is justified. And even if melioration learning turns out to be too simple, it may serve as valid starting point for further investigations. Instead of using another learning model, the melioration algorithm can be adjusted in order to account for empirical results. For example, a rapid change in behaviour is facilitated if, instead of all previous encounters, a smaller time frame is used to calculate the Q-values. Alternatively, the speed of learning is increased by adding eligibility traces (Sutton and Barto, 1998, ch. 7).

A rapid change in behaviour also takes place if a new state of the environment is recognised. In its general form, Q-learning allows the discrimination between environmental states (Watkins and Dayan, 1992). However, the state of the world

ACCEPTED MANUSCRIPT

cannot be assumed to be just given to the actors. They must learn which environmental aspects are relevant for the reinforcement mechanisms and whether two stimuli indicate the same state or two different states. Thus, a theory of stimulus discrimination is needed to complement the reinforcement learning process (Shteingart and Loewenstein, 2014). Some approaches have already been studied (e.g. Sakai et al., 2006). There are also extensions of RL techniques that can deal with continuous state sets (van Hasselt, 2012) or with limited information about the current state of the environment (Spaan, 2012).

Furthermore, Q-learning, like other reinforcement learning methods, is goal-dependent. The Q-values are learned with respect to particular preferences, which are expressed by the subjective values of the results. If the preferences change, new Q-values must be learned. In contrast, actors can use their experiences to build a mental representation of the environment. Subsequently, they are able to update behaviour in the pursue of new goals. For example, associative learning denotes a process in which the actors learn associations between states of the environment.

There is some progress in the integration of associative learning techniques in models of reinforcement learning (Alonso and Mondragón, 2006; Veksler et al., 2014). Similarly, actors can be designed to learn the reward functions of a given situation (Sutton and Barto, 1998, ch. 9; Hester and Stone, 2012). However, with those extensions, the simple ideas of reinforcement learning are abandoned. The presence of a mental image of the environment and the processing of this image are properties of belief learning models. It can be speculated that, in the end, a good model of human learning should integrate elements of both reinforcement and belief-based models (see also Camerer and Ho, 1999).

ACCEPTED MANUSCRIPT

7 The matching law and optimal behaviour

In this section, the relation between the matching law and the optimal solution of choice is discussed. As suggested by Rachlin et al. (1976, 1980, 1981), this can be achieved by integrating the matching law into economic consumer theory. This approach also allows to theoretically derive model parameters from individual preferences and situational constraints.

When applying the strict matching law of equation (1) to a situation of repeated decision-making, the value of the reinforcements should be independent of the chosen alternative. For example, the value of a goal in a penalty kick situation is the same regardless of the chosen side. In experiments with different types of reinforcements, systematic deviations from the strict matching law have been observed. This led to the formulation of the generalised matching law (equation (2)). Although the generalised version was able to account for most of the observed behaviour in experimental studies (e.g. Baum, 1979; Pierce and Epling, 1983; Herrnstein, 1997; McDowell, 2005, 2013a), the arbitrary choice of the free parameters in order to fit the model to the data reduces the falsifiability of the matching law and its applicability as micro-level assumption.

On the contrary, the parameters should be derived theoretically from situational constraints and individual preferences. This is a general problem in behavioural psychology, and, early on, Howard Rachlin and colleagues (Rachlin et al., 1976, 1980, 1981) argued that the introduction of microeconomic theory in experimental research might help with this problem. Following this previous work, a situation of repeated choice, such as a behavioural experiment or a real-world setting, is modelled by the consumer problem of microeconomic theory. The free

parameters of the generalised matching law are avoided by substituting the frequencies of reinforcement (s_1 and s_2 of equation (1)) by more adequate measures of an actor's utilities. This procedure is in line with the interpretations of the matching law by Herrnstein (1997) and Gray and Tallman (1984).

The consumer problem of microeconomic theory was also used by Coleman (1990) in his linear system of actions. But instead of predicting behaviour by utility maximisation, this section employs the matching law as behavioural assumption. This increases the validity of the model because an empirical regularity is set in place of a normative description of behaviour. However, this extension of economic consumer theory requires additional assumptions about the situation. In addition to a set of outcomes and an actor's preferences, also the behaviour must be explicitly modelled. Moreover, restrictions are applied for the matching law to be introduced as solution to this situation. In line with previous research on this subject (Herrnstein and Prelec, 1991), this restricted version of the consumer problem will be called the problem of distributed choice.

7.1 The problem of distributed choice

For any $m \in \mathbb{N}$, a set $E = \{e_1, e_2, \ldots, e_m\}$ is regarded as set of choice alternatives. An actor is assumed to repeatedly choose one of the alternatives of E. To be consistent with the existing literature, it is distinguished between behaviour and outcomes. The behaviour of an actor is modelled by the set of choice distributions

$$\mathcal{P} := \left\{ \boldsymbol{p} = (p_{e_1}, p_{e_2}, \ldots, p_{e_m}) \in [0,1]^E \;\middle|\; \textstyle\sum_{j \in E} p_j = 1 \right\}.$$

Each element of $\mathcal{P}$ denotes a frequency distribution over the alternatives E.

As customary in microeconomic theory, the outcomes are given by a budget set $\mathcal{X}^E$ such that an element $\boldsymbol{x} = (x_{e_1}, x_{e_2}, \ldots, x_{e_m}) \in \mathcal{X}^E$ specifies the relative amounts of reinforcement that are obtained after choosing the corresponding alternatives. Situational constraints affect the relationship between behaviour and outcome, which is commonly specified by a budget function $g : \mathcal{P} \to \mathcal{X}^E$. In case of a single decision, a budget function maps an action to an objective outcome. Given a behaviour distribution $\boldsymbol{p} \in \mathcal{P}$, a distribution of outcomes over the alternatives E is returned.

An example that illustrates the formal specifications arises in the daily selection of lunch (Herrnstein and Prelec, 1991). In a highly simplified model, an actor chooses between only two different items, e.g. pizza (P) and salad (S). The distribution $\boldsymbol{p} = (p_P, p_S)$ gives the relative frequencies of choice over a period of several days, e.g. 80% pizza and 20% salad. A budget function maps this behaviour to the objective results (x_P, x_S). The elements x_P and x_S represent the shares of pizzas and salads, respectively, including the characteristics of the food such as the crusts and toppings of the pizzas or the composition of the salads.

In addition to the specification of situational constraints by a budget function, assumptions about the preferences of an actor have to be made. For instance, an actor may show a general preference for pizza but dislikes a thin crust. Similarly, she may prefer no pepper in her salad. Additionally, satiation or deprivation effects change the subjective value of another unit of pizza. If pizza is chosen every day, its average value is lower than if chosen every second day.

In economic theory, preferences are represented by utility functions. In regard to the matching law, one fundamental requirement is that the actor's preferences over the set $\mathcal{X}^E$ can be described by a non-negative and additive utility function.

More specifically, let $\succsim$ be a relation on $\mathcal{X}^E$ that describes the actor's preferences, which means that, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}^E$, $\boldsymbol{x} \succsim \boldsymbol{y}$ if $\boldsymbol{x}$ is preferred to $\boldsymbol{y}$. It is required that there exists a function $u : \mathcal{X}^E \to [0, \infty)$ such that, for every $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}^E$,

$$\boldsymbol{x} \succsim \boldsymbol{y} \Leftrightarrow u(\boldsymbol{x}) \geq u(\boldsymbol{y})$$

and that u can be additively decomposed into $u_{e_1}, u_{e_2} \ldots, u_{e_m} : \mathcal{X} \to [0, \infty)$:

$$u(x_{e_1}, \ldots, x_{e_m}) = \sum_{j \in E} u_j(x_j) \text{ for every } (x_j)_{j \in E} \in \mathcal{X}^E.$$

The theory of additive conjoint measurement (Fishburn, 1970; Krantz et al., 1971) specifies necessary and sufficient conditions for the existence of an additive decomposition of the utility function.

The definition of both a budget and a utility function allows to distinguish between situational constraints and individual preferences, which eases the modelling of the situation (see section 7.2). For the purpose of a compact presentation, the two functions are concatenated to a single function $v := u \circ g$, which is called value function. Similarly, component value functions $v_j : \mathcal{P} \to [0, \infty)$, $j \in E$ can be defined such that $v(\boldsymbol{p}) = \sum_{j \in E} v_j(\boldsymbol{p})$.

The following definition of a problem of distributed choice summarises the prerequisites. It extends the consumer problem by explicitly modelling the behaviour $\mathcal{P}$ and restricts the situation such that the matching law can be applied.

Definition 1. *Let $m \in \mathbb{N}$, $E = \{e_1, e_2, \ldots, e_m\}$, and $\mathcal{P}$ as defined above. A* ***problem of distributed choice*** *is given by a pair (E, v) if $v : \mathcal{P} \to [0, \infty)$ and*

there exist functions $v_{e_1}, v_{e_2} \ldots, v_{e_m} : \mathcal{P} \to [0, \infty)$ *with*

$$v(\boldsymbol{p}) = \sum_{j \in E} v_j(\boldsymbol{p}) \text{ for every } \boldsymbol{p} \in \mathcal{P}.$$

For a given $\boldsymbol{p} = (p_j)_{j \in E} \in \mathcal{P}$, $v_j(\boldsymbol{p})$ returns the share of the total value $v(\boldsymbol{p})$ that is associated with the choice of alternative $j \in E$. If $p_j > 0$, then

$$\overline{v}_j(\boldsymbol{p}) := \frac{v_j(\boldsymbol{p})}{p_j}$$

is the average value of alternative $j \in E$. It gives the mean value of an reinforcement that is received after the choice of j. In situations in which the value of every reinforcement is constant and standardised to one, the average value equals the rate of reinforcement (corresponding to $\frac{s_j}{k_j}$ in equation (1)). This motivates the following definition of the matching law.

Definition 2. *Given a problem of distributed choice* (E, v)*, the* ***matching law*** *holds for* $\boldsymbol{p} = (p_j)_{j \in E} \in \mathcal{P}$ *if, for all* $i, j \in E$ *with* $p_i, p_j > 0$*,*

$$\overline{v}_i(\boldsymbol{p}) = \overline{v}_j(\boldsymbol{p}). \tag{6}$$

Figure 11 shows a definition of the value functions in case of the lunch example from above. The average values of pizza $\overline{v}_P$ and salad $\overline{v}_S$ decrease with the relative frequencies of choosing the respective alternatives (note that $p_S = 1 - p_P$). This may, for instance, result from a utility function that accounts for satiation effects. According to definition 2, the matching law holds for the following frequency distributions: $(p_P, p_S) \in \{(0, 1), (1, 0), (0.875, 0.125)\}$. The trivial distributions $(0, 1)$

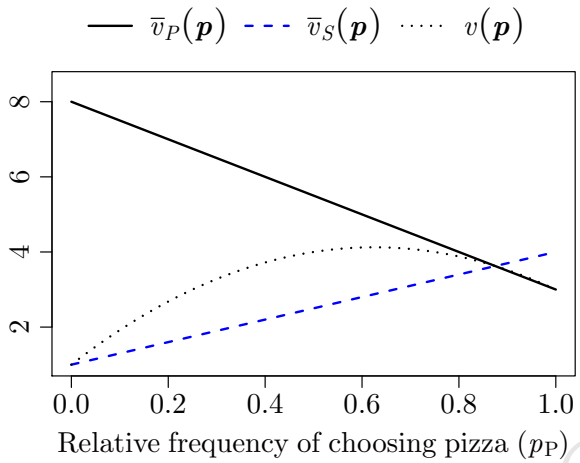

Figure 11: An example of the problem of distributed choice: The maximum of the overall value $v(\boldsymbol{p})$ diverges from the matching law, which holds if the average value of pizza equals the average value of salad: $\overline{v}_P(\boldsymbol{p}) = \overline{v}_S(\boldsymbol{p})$; $\boldsymbol{p} = (p_P, p_S)$

and $(1, 0)$ are elements of the matching law because there is no pair $i, j \in E$ with $p_i, p_j > 0$. The third distribution $(0.875, 0.125)$ corresponds to the intersection of the curves $\overline{v}_P$ and $\overline{v}_S$. Since this is the only intersection, there are no other elements of $\mathcal{P}$ for which the matching law holds.

The maximum of the overall value $v(\boldsymbol{p})$ is given at $\boldsymbol{p}^* = (0.625, 0.375)$. If the actor aligns her behaviour along this distribution, she would optimise her overall outcome. It is a general and important property of the matching law that it may diverge from optimal behaviour (Rachlin et al., 1980; Vaughan and Herrnstein, 1987; Herrnstein, 1990a,b; Herrnstein and Prelec, 1991; Herrnstein and Prelec, 1992). Consequently, the matching law can be regarded as alternative to this classic economic prediction of individual behaviour.

Nevertheless, various authors have argued that optimal behaviour results in the matching law under certain conditions (Rachlin et al., 1976; Staddon and Motheral, 1978; Rachlin et al., 1980; Herrnstein, 1982). A general specification of

these conditions was recently given by Kubanek (2017). The author showed that the matching law is observed whenever an actor optimises her behaviour and there exists a strictly monotonic function f such that

$$\frac{dv_e(\boldsymbol{p})}{dp_e} = f\left(\frac{v_e(\boldsymbol{p})}{p_e}\right), \text{ for all } e \in E. \tag{7}$$

Additionally, Kubanek (2017) discovered that the matching law is necessarily optimal if $f' > 0$ and $f(x) < x$, which implies diminishing value returns.

In the example of figure 11, the average values $\overline{v}_P$ and $\overline{v}_S$ are linear functions of p_P and p_S:

$$\overline{v}_P = b_P + a_P \cdot p_P \text{ and } \overline{v}_S = b_S + a_S \cdot p_S,$$

with $b_P, a_P, b_S, a_S \in \mathbb{R}$. The derivations of v_P and v_S are, thus, given by

$$\frac{dv_P(\boldsymbol{p})}{dp_P} = b_P + 2 \cdot a_P \cdot p_P \text{ and } \frac{dv_S(\boldsymbol{p})}{dp_S} = b_S + 2 \cdot a_S \cdot p_S.$$

A strictly monotonic function f for condition (7) to hold could be found if $b_P = b_S$. But since $b_P \neq b_C$ in the example of figure 11, optimal behaviour and the matching law predict different behaviour. A situation in which condition (7) necessarily holds is described in the following section.

7.2 Deriving the parameters of the situation

Definition 2 generalises the strict version of the matching law (equation (1)). By including average values $\overline{v}_j$, it substitutes the rates of reinforcement $\frac{s_j}{k_j}$ by more general measures of reinforcement. Similar to the parameters of the generalised matching law, these measures have to be derived theoretically if the matching law

is applied as behavioural assumption.

As an example, a behavioural experiment with concurrent schedules of reinforcement is assumed. In this kind of experiment, an individual chooses between two or more behavioural alternatives, e.g. pressing one of several buttons. The reinforcement of either choice is triggered by a schedule. For instance, a ratio schedule may specify that every 5th choice of a certain alternative is followed by a reinforcement. Interval schedules trigger reinforcements after a certain amount of time has passed since the last reinforcement.

According to Rachlin et al. (1980), the budget function of a behavioural experiment with m alternatives ($E = \{1, 2, \ldots, m\}$) is given by

$$g(\boldsymbol{p}) = (b_1 \cdot p_1^r, b_2 \cdot p_2^r, \ldots, b_m \cdot p_m^r)\,, \text{ for all } \boldsymbol{p} = (p_1, p_2, \ldots, p_m) \in \mathcal{P}. \tag{8}$$

The parameter $r \in (0, 1]$ specifies the type of the reinforcement schedules. If $r = 1$, the alternatives are reinforced by ratio schedules. Interval schedules are characterised by $0 < r < 1$ (Rachlin et al., 1980, p. 362). The second set of parameters $(b_1, \ldots, b_m) \in [0, \infty)^m$ allows to include differences in probabilities or amounts of reinforcement.

Next to situational constraints, the preferences of the actor must be specified by a utility function. In the following, the preferences are assumed to follow a constant elasticity of substitution (CES) utility function (Dixit and Stiglitz, 1977):

$$u(\boldsymbol{x}) = \left(\sum_{j \in E} a_j \cdot x_j^\rho \right)^{\frac{1}{\rho}}, \text{ for all } \boldsymbol{x} = (x_j)_{j \in E} \in \mathcal{X}, \tag{9}$$

with $\rho \leq 1$, $\rho \neq 0$, and $a_j > 0$, for all $j \in E$.

CES utilities cover a variety of situations because they are able to account for differences in the amount or quality of reinforcements as well as for interrelated deprivation rates. A simple instance is given by the linear utility function ($\rho = 1$):

$$u(\boldsymbol{x}) = \sum_{j \in E} a_j \cdot x_j, \text{ for all } \boldsymbol{x} = (x_j)_{j \in E} \in \mathcal{X}.$$

Linear functions cover reinforcements that are perfect substitutes, e.g. if they consists of the same amount of the same resource. But also experiments with differences in the quality of reinforcements correspond to linear utility functions as long as the parameters a_j are set to appropriate values.

In case of $\rho < 1$, the resources that are used as reinforcements somehow complement each other. If, for example, two different kinds of food are used as reinforcements and both resources are subject to satiation, the subjective value of an additional unit of either resource decreases with its repeated consumption. When continuously consuming one kind, receiving the other kind from time to time actually increases the average value of the first one. Therefore, a mix of both resources can result in a higher overall value than the consumption of a single resource. In the limit $\rho \to 0$, equation (9) approaches the Cobb-Douglas utility function, which was assumed by Coleman (1990) in his study of social exchange. The assumption of CES preferences is less restrictive (see also Braun, 1994).

Any situation that can be described by the budget function of equation (8) and a CES utility function conforms to a problem of distributed choice (definition 1). This is seen by transforming the utility function to u^ρ, which is clearly an additive and non-negative function on $\mathcal{X}^E$. The concatenation of the budget function g

and the utility function u^ρ leads to the following component value functions:

$$v_j(\boldsymbol{p}) = a_j \cdot \left(b_j \cdot p_j^r\right)^\rho, \text{ for each } j \in E \text{ and } \boldsymbol{p} = (p_j)_{j \in E} \in \mathcal{P}. \tag{10}$$

Since $\frac{dv_j(\boldsymbol{p})}{dp_j} = r \cdot \rho \cdot a_j \cdot b_j^\rho \cdot p_j^{r\rho-1} = r\rho\frac{v_j(\boldsymbol{p})}{p_j}$, there exists a strictly monotonic function $f(x) = r\rho x$ such that condition (7) holds. According to Kubanek (2017), this means that optimal behaviour conforms to the matching law in this particular class of situations. With regard to the second condition of Kubanek (2017), the matching law is necessarily optimal if $0 < r\rho < 1$ and all alternatives are chosen with strictly positive frequency ($p_j > 0$ for all $j \in E$).

8 Conclusion

This paper is an attempt to integrate an often observed empirical regularity of individual decision-making into sociological research. Since this so-called matching law refers to aggregated individual behaviour, it cannot be directly established as behavioural assumption. Instead, a mechanism of decision-making is required when deriving social phenomena. One possible mechanism is given by the melioration learning model. Currently, no general results about the convergence of melioration learning exist. In simple settings (see proposition 1 in the appendix), melioration is guaranteed to converge to the matching law. But most social situations do not conform to those conditions. It is still possible to analyse them by means of computer simulations.

When applied to the volunteer's dilemma, the melioration model was shown to yield new and falsifiable predictions. Under certain conditions, the results are

in line with the pure or mixed Nash equilibrium. This means that the predictions of game theory may be valid even without its rather strict assumptions about the actors' reasoning. However, in the asymmetric version of the dilemma, predictions of melioration learning and the mixed Nash equilibrium diverge. The results from the simulations are more intuitive than the game-theoretic predictions.

There are many open problems. Most importantly, hypotheses that were derived from melioration learning must be tested empirically. Laboratory experiments are a convenient method. But melioration requires a long period of decision-making, which increases the costs of experiments. Moreover, most of the existing data is not applicable because, in those experiments, information about the structure of the situation was given to the subjects. Since this information may affect the decisions but is not considered by actors who learn by melioration, new experiments must be conducted.

By its integration into economic consumer theory, the matching law was related to optimal behaviour. While previous work has pointed to conditions under which matching and maximisation align (Kubanek, 2017), an important property of the matching law is its deviation from optimal behaviour in many situations of repeated decision-making. It can, hence, be employed as an alternative to the more common assumption of rationality, which usually refers to long-term utility maximisation. Due to the existence of many experiments that confirm the matching law as empirical regularity of individual behaviour, it should be preferred to the assumption of optimal behaviour. Additionally, only low cognitive skills and very few information about the situation are needed for an actor to behave in accordance with the matching law.

Further analyses of the relationship between optimisation and the matching

ACCEPTED MANUSCRIPT

law can be found in Baum (1981), Vaughan (1981), Herrnstein (1997), Sakai and Fukai (2008b), and Loewenstein et al. (2009). For instance, the matching law is different from maximisation because it neglects the long-term effects of present decisions. Accordingly, it has been argued that the matching law may be considered as rational choice in uncertain environments. While humans are capable of improving their outcomes if sufficient information is supplied, perfectly rational behaviour can lead to suboptimal outcomes if an actor is uncertain about the social environment, e.g. about the choices of the other actors (Flache, 2002). Likewise, a study of Sims et al. (2013) indicated that rational actors choose a suboptimal equilibrium if only few information about the situation is present. Sims et al. (2013, p. 139) concluded that "melioration can be reinterpreted not as irrational choice but rather as globally optimal choice under uncertainty".

ACCEPTED MANUSCRIPT

9 Appendix

Algorithm 1 The melioration learning algorithm

Require: initial exploration rate $\varepsilon_0 \in (0, 1)$, set of alternatives E

$t \leftarrow 0$
initialise $Q_1(j) \leftarrow 0$, for all $j \in E$
initialise $K_1(j) \leftarrow 0$, for all $j \in E$
repeat
  $t \leftarrow t + 1$
  $\varepsilon \leftarrow \frac{\varepsilon_0}{1+\sum_{j \in E} K_t(j)}$
  if $\varepsilon >$ random number between 0 and 1 (uniformly distributed) **then**
    choose a random action $X_t \leftarrow e \in E$ using a uniform distribution
  else
    choose action $X_t \leftarrow e$ such that $e \in \arg\max_{j \in E} Q_t(j)$
  end if
  observe reward $R_t = y$
  $K_{t+1}(e) \leftarrow K_t(e) + 1$
  $Q_{t+1}(e) \leftarrow Q_t(e) + \frac{1}{K_{t+1}(e)} \cdot (y - Q_t(e))$
  for all $j \neq e$ **do**
    $K_{t+1}(j) \leftarrow K_t(j)$
    $Q_{t+1}(j) \leftarrow Q_t(j)$
  end for
until termination

According to the following proposition, melioration learning, as given by algorithm 1, leads to the matching law if certain conditions of stationarity hold. These conditions include Markov decision processes (Bellman, 1957; Watkins and Dayan, 1992) and, therefore, many non-social settings.

Proposition 1. *Consider a situation of sequential decision-making and an actor who follows algorithm 1. If, for every $e \in E$, the limit of $Q_t(e)$ exists, then, for all $i, j \in E$ with $\lim_{t\to\infty} \frac{1}{t} K_t(i) > 0$ and $\lim_{t\to\infty} \frac{1}{t} K_t(j) > 0$ (a.s.),*

$$\lim_{t\to\infty} Q_t(i) = \lim_{t\to\infty} Q_t(j) \text{ (a.s.).}$$

ACCEPTED MANUSCRIPT

Proof of proposition 1. Let $Q^*(i) := \lim_{t\to\infty} Q_t(i)$ and $Q^*(j) := \lim_{t\to\infty} Q_t(j)$. Assuming that $Q^*(i) > Q^*(j)$, there must exist $t_0 \in \mathbb{N}$ such that for all $t > t_0$: $Q_t(i) > Q_t(j)$. According to algorithm 1, this implies that action j is chosen with probability ε and independently of the previous choices at every time step t with $t > t_0$. To arrive at a contradiction, it is shown that $\frac{1}{t}K_t(j) \xrightarrow{\text{a.s.}} 0$. Note that $\frac{1}{t}K_t(j) = \frac{1}{t}\sum_{l=1}^{t} \mathbf{1}_{\{X_l=j\}}$, with $\mathbf{1}_{\{X_l=j\}}$ being the indicator random variable:

$$\mathbf{1}_{\{X_l=j\}} = \begin{cases} 1 & \text{if } X_l = j, \\ 0 & \text{else.} \end{cases}$$

First, since ε converges towards zero as $t \to \infty$:

$$\mathbb{E}\left(\mathbf{1}_{\{X_l=j\}} \mid \mathbf{1}_{\{X_{l-1}=j\}}, \ldots, \mathbf{1}_{\{X_1=j\}}\right) \xrightarrow{\text{a.s.}} 0 \text{ as } l \to \infty. \tag{11}$$

Second, because the variance of $\mathbf{1}_{\{X_l=j\}}$ is between zero and one, the stability theorem of Loève (1978, p. 53) yields:

$$\frac{1}{t}\sum_{l=1}^{t} \mathbf{1}_{\{X_l=j\}} - \mathbb{E}\left(\mathbf{1}_{\{X_l=j\}} \mid \mathbf{1}_{\{X_{l-1}=j\}}, \ldots, \mathbf{1}_{\{X_1=j\}}\right) \xrightarrow{\text{a.s.}} 0. \tag{12}$$

From (11) and (12) follows $\frac{1}{t}K_t(j) \xrightarrow{\text{a.s.}} 0$, and a premise is violated. In an analogous manner, a contradiction can be drawn from the assumption $Q^*(i) < Q^*(j)$. □

Figure 12 shows results of simulations of the asymmetric volunteer's dilemma. In contrast to figure 10, the groups are fixed and not randomly assembled before every round. Nevertheless, the relative frequencies of volunteering are similar to the ones above. Strong actors are more likely to volunteer than weak actors.

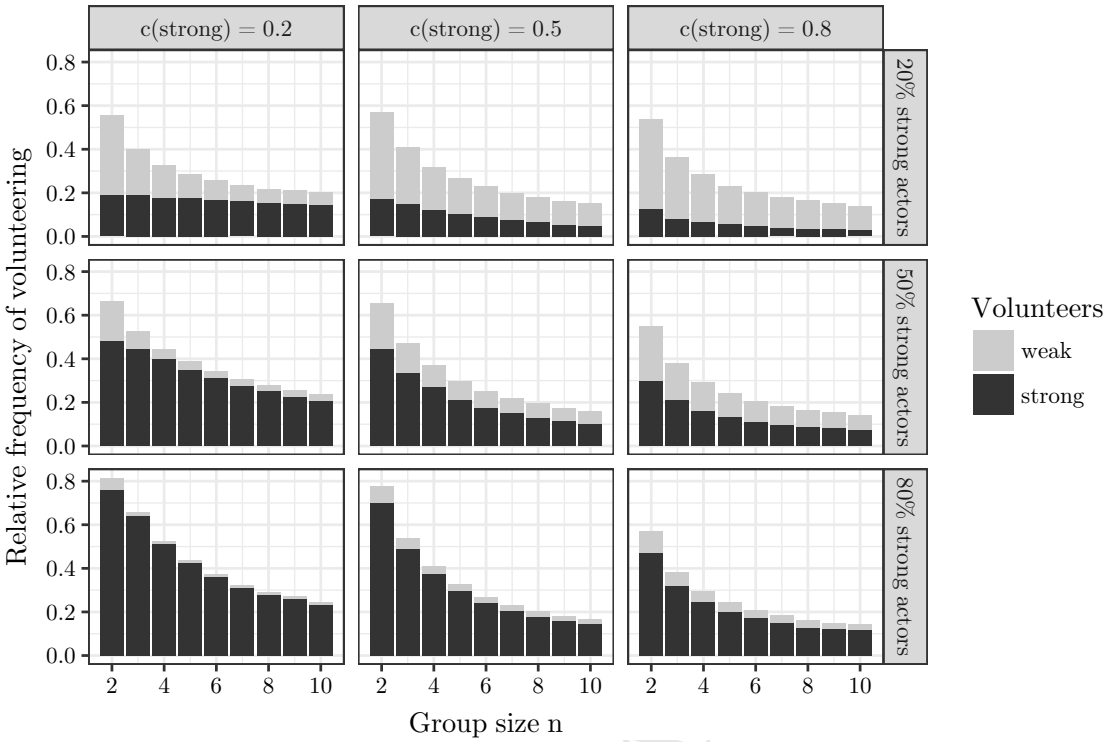

Figure 12: The effect of group size in the asymmetric volunteer's dilemma with fixed groups; $c = 1$; the strength and the proportion of strong actors vary between 0.2 and 0.8

References

Alonso, E. and E. Mondragón (2006). Associative learning for reinforcement learning: Where animal learning and machine learning meet. In E. Alonso and Z. Guessoum (Eds.), *Proceedings of the Fifth Symposium on Adaptive Agents and Multi-Agent Systems*, Paris, France, pp. 87–99.

Antonides, G. and S. Maital (2002). Effects of feedback and educational training on maximization in choice tasks: Experimental-game evidence. *The Journal of Socio-Economics 31*(2), 155–165.

ACCEPTED MA

ACCEPTED MANUSCRIPT

Axelrod, R. (1997). Advancing the art of simulation in the social sciences. *Complexity 3*(2), 16–22.

Barto, A. G., R. S. Sutton, and C. J. C. H. Watkins (1990). Learning and sequential decision making. In M. Gabriel and J. Moore (Eds.), *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, Cambridge, Mass, pp. 539–602. MIT Press.

Baum, W. M. (1974). On two types of deviation from the matching law: bias and undermatching. *Journal of the Experimental Analysis of Behavior 22*(1), 231–242.

Baum, W. M. (1979). Matching, undermatching, and overmatching in studies of choice. *Journal of the Experimental Analysis of Behavior 32*(2), 269–281.

Baum, W. M. (1981). Optimization and the matching law as accounts of instrumental behaviour. *Journal of the Experimental Analysis of Behavior 36*(3), 387–403.

Baum, W. M. and J. A. Nevin (1981). Maximization theory: Some empirical problems. *Behavioral and Brain Sciences 4*(3), 389–390.

Becker, G. S. (1981). *A Treatise on the Family*. Cambridge: Harvard University Press.

Bellman, R. E. (1957). A Markov decision process. *Journal of Mathematics and Mechanics 6*(5), 679–684.

Bendor, J. (1987). In good times and bad: Reciprocity in an uncertain world. *American Journal of Political Science 31*(3), 531–558.

ACCEPTED MANUSCRIPT

Bendor, J. (2001). Bounded rationality. In N. J. Smelser and P. B. Baltes (Eds.), *International Encyclopedia of the Social & Behavioral Sciences*, pp. 1303 – 1307. Oxford: Pergamon.

Borrero, J. C., S. S. Crisolo, Q. Tu, W. A. Rieland, N. A. Ross, M. T. Francisco, and K. Y. Yamamoto (2007). An application of the matching law to social dynamics. *Journal of Applied Behavior Analysis 40*(4), 589–601.

Braun, N. (1994). Restricted access in exchange systems. *The Journal of Mathematical Sociology 19*(2), 129–148.

Brenner, T. (2006). Agent learning representation: Advice on modelling economic learning. In L. Tesfatsion and K. L. Judd (Eds.), *Handbook of Computational Economics. Agent-based Computational Economics*, Volume 2. North-Holland.

Brenner, T. and U. Witt (2003). Melioration learning in games with constant and frequency-dependent pay-offs. *Journal of Economic Behavior & Organization 50*(4), 429–448.

Burgess, R. L. and D. Bushell (Eds.) (1969). *Behavioral Sociology. The Experimental Analysis of Social Processes*. New York and London: Columbia University Press.

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

Coleman, J. S. (1990). *Foundations of Social Theory*. Cambridge, Mass., and London, England: The Belknap Press of Harvard University Press.

ACCEPTED MANUSCRIPT

Conger, R. and P. Killeen (1974). Use of concurrent operants in small group research: A demonstration. *The Pacific Sociological Review 17*(4), 399–416.

Corrado, G. S., L. P. Sugrue, H. S. Seung, and W. T. Newsome (2005). Linear-nonlinear-Poisson models of primate choice dynamics. *Journal of the Experimental Analysis of Behavior 84*(3), 581–617.

Darley, J. M. and B. Latané (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology 8*(4), 377–383.

de Villiers, P. A. and R. J. Herrnstein (1976). Toward a law of response strength. *Psychological Bulletin 83*(6), 1131–1153.

Diekmann, A. (1985). Volunteer's dilemma. *Journal of Conflict Resolution 29*(4), 605 – 610.

Diekmann, A. (1993). Cooperation in an asymmetric volunteer's dilemma game. Theory and experimental evidence. *International Journal of Game Theory 22*(1), 75–85.

Dixit, A. K. and J. E. Stiglitz (1977). Monopolistic competition and optimum product diversity. *The American Economic Review 67*(3), 297 – 308.

Emerson, R. M. (1972). Exchange theory, part i: A psychological basis for social exchange. In J. Berger, M. Zelditch, and B. Anderson (Eds.), *Sociological Theories in Progress*, Volume 2, Chapter 3, pp. 38–57. Boston: Houghton Mifflin Company.

Fishburn, P. C. (1970). *Utility theory for decision making*. New York: Wiley.

Flache, A. (2002). The rational weakness of strong ties: Failure of group solidarity in a highly cohesive group of rational agents. *The Journal of Mathematical Sociology 26*(3), 189–216.

Flache, A. and M. W. Macy (2002). Stochastic collusion and the power law of learning: A general reinforcement learning model of cooperation. *Journal of Conflict Resolution 46*(5), 629.

Franzen, A. (1995). Group size and one-shot collective action. *Rationality and Society 7*(2), 183–200.

Gallistel, C. R., T. A. Mark, A. P. King, and P. E. Latham (2001). The rat approximates an ideal detector of changes in rates or reward: Implications for the law of effect. *Journal of Experimental Psychology: Animal Bahvior Processes 27*(4), 354–372.

Gigerenzer, G., P. M. Todd, and the ABC Research Group (1999). *Simple Heuristics that Make us Smart*. Oxford University Press.

Goeree, J. K., C. Holt, and A. M. Smith (2017). An experimental examination of the volunteer's dilemma. *Games and Economic Behavior 102*(C), 303–315.

Gray, L. N., W. I. Griffith, M. H. von Broembsen, and M. J. Sullivan (1982). Social matching over multiple reinforcement domains: An explanation of local exchange imbalance. *Social Forces 61*(1), 156–182.

Gray, L. N. and I. Tallman (1984). A satisfaction balance model of decision making and choice behavior. *Social Psychology Quarterly 47*(2), 146–159.

Gray, L. N. G. and M. H. von Broembsen (1976). On the generalizability of the law of effect: Social psychological measurement of group structure and process. *Sociometry 39*(3), 175–183.

Green, L. and D. E. Freed (1993). The substitutability of reinforcers. *Journal of the Experimental Analysis of Behavior 60*(1), 141–158.

Gureckis, T. M. and B. C. Love (2009). Short term gains, long term pains: How cues about state aid learning in dynamic environments. *Cognition 113*(3), 293–313.

Hamblin, R. L. (1977). Behavior and reinforcement: A generalization of the matching law. In R. L. Hamblin and J. H. Kunkel (Eds.), *Behavioral Theory in Sociology. Essays in Honor of George C. Homans*, pp. 469–502. New Brunswick, N.J.: transaction Books.

Hamblin, R. L. (1979). Behavioral choice and social reinforcement: Step function versus matching. *Social Forces 57*(4), 1141–1156.

Hamblin, R. L. and J. H. Kunkel (Eds.) (1977). *Behavioral Theory in Sociology. Essays in Honor of George C. Homans*. New Brunswick, N.J.: transaction Books.

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior 4*(3), 267–272.

Herrnstein, R. J. (1982). Melioration as behavioral dynamism. In M. L. Commons, R. J. Herrnstein, and H. Rachlin (Eds.), *Quantitative Analysis of Behavior,*

ACCEPTED MANUSCRIPT

Vol. II: Matching and Maximizing Accounts, pp. 433 – 458. Cambridge, Mass.: Ballinger Publishing Company.

Herrnstein, R. J. (1990a). Behavior, reinforcement and utility. *Psychological Science 1*(4), 217–224.

Herrnstein, R. J. (1990b). Rational choice theory: Necessary but not sufficient. *American Psychologist 45*(3), 356–367.

Herrnstein, R. J. (1997). *The Matching Law. Papers in Psychology and Economics*. Cambridge, Mass. & London, England: Harvard University Press.

Herrnstein, R. J., G. F. Loewenstein, D. Prelec, and W. Vaughan (1993). Utility maximization and melioration: Internalities in individual choice. *Journal of Behavioral Decision Making 6*(3), 149–185.

Herrnstein, R. J. and D. Prelec (1992). A theory of addiction. In G. Loewenstein and J. Elster (Eds.), *Choice over Time*, pp. 331 – 361. New York: Russell Sage Press.

Herrnstein, R. J. and D. Prelec (1991). Melioration: A theory of distributed choice. *Journal of Economic Perspectives 5*(3), 137–156.

Hester, T. and P. Stone (2012). Learning and using models. In M. Wiering and M. van Otterlo (Eds.), *Reinforcement Learning. State-of-the-Art*, Chapter 4, pp. 111–141. Berlin and Heidelberg: Springer.

Homans, G. C. (1961). *Social behavior. Its Elementary Forms*. London: Routledge & Kegan Paul.

ACCEPTED MANUSCRIPT

Homans, G. C. (1974). *Social behavior. Its Elementary Forms* (2. rev. ed.). New York: Harcourt Brace Jovanich, Inc.

Kangas, B. D., M. S. Berry, R. N. Cassidy, J. Dallery, M. Vaidya, and T. D. Hackenberg (2009). Concurrent performance in a three-alternative choice situation: Response allocation in a Rock/Paper/Scissors game. *Behavioural Processes 82*(2), 164–172.

Kianercy, A. and A. Galstyan (2012). Dynamics of Boltzmann Q-Learning in two-player two-action games. *Physical Review E 85*(4), 041145.

Krantz, D. H., R. D. Luce, P. Suppes, and A. Tversky (1971). *Foundations of Measurement. Volume 1: Additive and Polynomial Representations*. New York: Academic Press.

Kubanek, J. (2017). Optimal decision making and matching are tied through diminishing returns. *Proceedings of the National Academy of Sciences 114*(32), 8499–8504.

Loève, M. (1978). *Probability Theory II* (4th ed.). New York, Heidelberg, and Berlin: Springer.

Loewenstein, Y. (2010). Synaptic theory of replicator-like melioration. *Frontiers in Computational Neuroscience 4*, 17.

Loewenstein, Y., D. Prelec, and H. S. Seung (2009). Operant matching as a Nash equilibrium of an intertemporal game. *Neural Computation 21*(10), 2755–2773.

Loewenstein, Y. and H. S. Seung (2006). Operant matching is a generic outcome of

ACCEPTED MANUSCRIPT

synaptic plasticity based on the covariance betwen reward and neural activity. *Proceedings of the National Academy of Sciences 103*(41), 15224–15229.

Macy, M. and M. Tsvetkova (2015). The signal importance of noise. *Sociological Methods & Research 44*(2), 306–328.

Macy, M. W. and A. Flache (2009). Social dynamics from the bottom up. Agent-based models of social interaction. In P. Hedström and P. Bearman (Eds.), *The Oxford Handbook of Analytical Sociology*, Chapter 11, pp. 245–268. Oxford, England: Oxford University Press.

March, J. G. (1991). Exploration and exploitation in organization learning. *Organization Science 2*(1), 71–87.

Mazur, J. E. (1981). Optimization theory fails to predict performance of pigeons in a two-response situation. *Science 214*(4522), 823–825.

McDowell, J. J. (1988). Matching theory in natural human environments. *The Behavior Analyst 11*(2), 95–109.

McDowell, J. J. (2004). A computational model of selection by consequences. *Journal of the Experimental Analysis of Behavior 81*(3), 297–317.

McDowell, J. J. (2005). On the classic and modern theories of matching. *Journal of the Experimental Analysis of Behavior 84*(1), 111–127.

McDowell, J. J. (2013a). On the theoretical and empirical status of the matching law and matching theory. *Psychological Bulletin 139*(5), 1000–1028.

McDowell, J. J. (2013b). A quantitative evolutionary theory of adaptive behavior dynamics. *Psychological Review 120*(4), 731–750.

ACCEPTED MANUSCRIPT

McDowell, J. J. and M. L. Caron (2007). Undermatching is an emergent property of selection by consequences. *Behavioural Processes 75*(2), 97–106.

McDowell, J. J., M. L. Caron, S. Kulubekova, and J. P. Berg (2008). A compuational theory of selection by consequences applied to concurrent schedules. *Journal of the Experimental Analysis of Behavior 90*(3), 387–403.

McDowell, J. J. and A. Popa (2010). Toward a mechanism of adaptive behavior: Evolutionary dynamics and matching theory statics. *Journal of the Experimental Analysis of Behavior 94*(2), 241–260.

Molm, L. D. (2006). The social exchange framework. In P. J. Burke (Ed.), *Contemporary Social Psychological Theories*, pp. 24–45. Stanford, California: Stanford University Press.

Neth, H., C. R. Sims, and W. D. Gray (2005). Melioration despite more information: The role of feedback frequency in stable suboptimal performance. In *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting*, pp. 357–361.

Neth, H., C. R. Sims, and W. D. Gray (2006). Melioration dominates maximization : stable suboptimal performance despite global feedback. In R. Sun (Ed.), *CogSci/ICCS 2006; Vancouver, British Columbia, Canada*, pp. 627–632.

Olson, M. (1965). *The Logic of Collective Action. Public Goods and the Theory of Groups*. Harvard University Press.

Palacios-Huerta, I. (2003). Professionals play minimax. *The Review of Economic Studies 70*(2), 395–415.

ACCEPTED MANUSCRIPT

Pierce, W. D. and W. F. Epling (1983). Choice, matching, and human behavior. A review of the literature. *The Behavior Analyst 6*(1), 57–76.

Rachlin, H. (1971). On the tautology of the matching law. *Journal of the Experimental Analysis of Behavior 15*(2), 249–251.

Rachlin, H., R. C. Battalio, J. H. Kagel, and L. Green (1981). Maximization theory in behavioral psychology. *Behavioral and Brain Sciences 4*(3), 371–417.

Rachlin, H., L. Green, J. H. Kagel, and R. C. Battalio (1976). Economic demand theory and psychological studies of choice. In G. Bower (Ed.), *The Psychology of Learning and Motivation*, Volume 10, pp. 129–154. New York: Academic Press.

Rachlin, H., J. H. Kagel, and R. C. Battalio (1980). Substitutability in time allocation. *Psychological Review 87*(4), 355–374.

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

Sakai, Y. and T. Fukai (2008a). The actor-critic learning is behind the matching law: Matching versus optimal behaviors. *Neural Computation 20*(1), 227–251.

Sakai, Y. and T. Fukai (2008b). When does reward maximization lead to matching law? *PLoS ONE 3*(11), e3795.

Sakai, Y., H. Okamoto, and T. Fukai (2006). Computational algorithms and neuronal network models underlying decision processes. *Neural Networks 19*(8), 1091 – 1105.

ACCEPTED MANUSCRIPT

Selten, R. (1975). Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory 4*(1), 25–55.

Shteingart, H. and Y. Loewenstein (2014). Reinforcement learning and human behavior. *Current Opinion in Neurobiology 25*, 93–98.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics 69*(1), 99–118.

Sims, C. R., H. Neth, R. A. Jacobs, and W. D. Gray (2013). Melioration as rational choice: Sequential decision making in uncertain environments. *Psychological Review 120*(1), 139–154.

Skyrms, B. and R. Pemantle (2000). A dynamic model of social network formation. *Proceedings of the National Academy of Sciences 97*(16), 9340–9346.

Spaan, M. T. J. (2012). Partially observable Markov decision processes. In M. Wiering and M. van Otterlo (Eds.), *Reinforcement Learning. State-of-the-Art*, Chapter 12, pp. 387–414. Berlin and Heidelberg: Springer.

Staddon, J. E. R. and S. Motheral (1978). On matching and maximizing in operant choice experiments. *Psychological Review 85*(5), 436–444.

Sugrue, L. P., G. S. Corrado, and W. T. Newsome (2004). Matching behavior and the representation of value in the parietal cortex. *Science 304*(5678), 1782–1787.

Sunahara, D. F. and W. D. Pierce (1982). The matching law and bias in a social exchange involving choice between alternatives. *The Canadian Journal of Sociology 7*(2), 145–166.

ACCEPTED MANUSCRIPT

Sutton, R. S. and A. G. Barto (1998). *Reinforcement learning. An Introduction*. Cambridge, Massachusetts, and London, England: The MIT Press.

Tunney, R. J. and D. R. Shanks (2002). A re-examination of melioration and rational-choice. *Journal of Behavioral Decision Making 15*(4), 291–311.

van Hasselt, H. (2012). Reinforcement learning in continuous state and action spaces. In M. Wiering and M. van Otterlo (Eds.), *Reinforcement Learning. State-of-the-Art*, Chapter 7, pp. 207–251. Berlin and Heidelberg: Springer.

van Otterlo, M. and M. Wiering (2012). Reinforcement learning and markov decision processes. In M. Wiering and M. van Otterlo (Eds.), *Reinforcement Learning. State-of-the-Art*, Chapter 1, pp. 3–42. Berlin and Heidelberg: Springer.

Vaughan, W. (1981). Melioration, matching, and maximization. *Journal of the Experimental Analysis of Behavior 36*(2), 141–149.

Vaughan, W. and R. J. Herrnstein (1987). Stability, melioration, and natural selection. In L. Green and J. H. Kagel (Eds.), *Advances in Behavioral Economics*, Volume 1, pp. 185–215. Norwood, N.J.: Ablex.

Veksler, V. D., C. W. Myers, and K. A. Gluck (2014). SAwSu: An integrated model of associative and reinforcement learning. *Cognitive Science 38*(3), 580–598.

Vollmer, T. R. and J. Bourret (2000). An application of the matching law to evaluate the allocation of two- and three-point shots by college basketball players. *Journal of Applied Behavior Analysis 33*(2), 137–150.

Watkins, C. J. C. H. (1989). Learning from delayed rewards. Ph. D. thesis, University of Cambridge, England.

Watkins, C. J. C. H. and P. Dayan (1992). Q-learning. *Machine Learning 8*(3-4), 279–292.

Wiering, M. and M. van Otterlo (Eds.) (2012). *Reinforcement Learning. State-of-the-Art*. Berlin and Heidelberg: Springer.

Wilensky, U. (1999). Netlogo. `http://ccl.northwestern.edu/netlogo/`. Center for Connected Learning and Computer-Based Modeling, Northwestern University. Evanston, IL.

Wunder, M., M. Littman, and M. Babes (2010). Classes of multiagent Q-learning dynamics with ϵ-greedy exploration. In *Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel*, pp. 1167–1174.

Yechiam, E., I. Erev, V. Yehene, and D. Gopher (2003). Melioration and the transition from touch-typing training to everyday use. *Human Factors: The Journal of the Human Factors and Ergonomics Society 45*(4), 671–684.

ACCEPTED MANUSCRIPT