Animal Behaviour 104 (2015) 87–114

Contents lists available at ScienceDirect

Animal Behaviour

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

A model for the evolution of reinforcement learning in fluctuating games

Slimane Dridi*, Laurent Lehmann

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

ARTICLE INFO

Article history:
Received 17 June 2014
Initial acceptance 18 June 2014
Final acceptance 16 January 2015
Available online 11 April 2015
MS. number: 14-00499R

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

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

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

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

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

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

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

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

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

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

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

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