Journal of Economic Behavior and Organization 152 (2018) 224–234

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

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

The reinforcement heuristic in normal form games

Carlos Alós-Ferrer*, Alexander Ritschel

Department of Economics, University of Zurich, Blümlisalpstrasse 10, Zurich 8006, Switzerland

ARTICLE INFO

Article history:
Received 16 August 2017
Revised 25 June 2018
Accepted 25 June 2018

JEL classification:
C72
C91

Keywords:
Reinforcement
Myopic best reply
Response times
Decision processes

ABSTRACT

We analyze simple reinforcement-based behavioral rules in 3×3 games through choice data and response times. We argue that there is a large overlap between reinforcement-based heuristics (win-stay, lose-shift) and the more "rational" behavioral rule of myopic best reply. However, evidence from response times shows that choices in agreement with the common prescription of those rules are comparatively fast, and choices of the form "lose-shift" occur more frequently for larger differences with bygone payoffs. Both observations speak in favor of reinforcement processes as a cognitive shortcut for apparent myopic best reply, and advise caution when interpreting behavioral results in favor of optimizing behavior.

1. Introduction

Reinforcement is one of the most basic processes underlying human learning. Accordingly, it has received widespread attention in psychology, going back to Thorndike (1911)'s "law of effect," neuroscience (e.g. Holroyd and Coles, 2002; Schönberg et al., 2007), and computer science (Sutton and Barto, 1998). Within microeconomics and game theory, it has been frequently studied as a boundedly-rational behavioral rule (see, e.g. Börgers and Sarin, 1997; Erev and Roth, 1998), as have been other rules, e.g. imitation or myopic best reply (Fudenberg and Levine, 1998; Weibull, 1995). The simplest version of reinforcement learning can be viewed as a heuristic which takes past experiences into account for the choice of upcoming actions and prescribes a shift from actions linked to negative experiences to actions associated with positive rewards: that is, "win-stay, lose-shift". This heuristic induces a bias towards past-high-reward actions which can conflict with rational behavior (outcome bias; Baron and Hershey, 1988; Dillon and Tinsley, 2008).

Evidence from neuroscience shows that reinforcement-based decisions occur extremely fast in the human brain (Holroyd and Coles, 2002; Schultz, 1998). Indeed, reinforcement is a textbook example of an *automatic process*, as conceived in dual-process theories from psychology (see, e.g., Alós-Ferrer and Strack, 2014; Kahneman, 2003; Strack and Deutsch, 2004). Those theories define automatic (or intuitive) processes as immediate, fast, unconscious, and efficient in the sense of requiring few cognitive resources. For instance, these processes capture impulsive reactions and behavior along the lines of stimulus-response schemes. The dual-process approach postulates that human decisions are mainly influenced by automatic processes and so-called *controlled* (or deliberative) processes. The latter are seen as slow, consuming cognitive resources, not instigated immediately, and reflected upon consciously. Explicit maximization of expected rewards, if conceptualized as a process, would exhibit many if not all of those characteristics.

* Corresponding author.
E-mail addresses: carlos.alos-ferrer@econ.uzh.ch (C. Alós-Ferrer), alexander.ritschel@econ.uzh.ch (A. Ritschel).

https://doi.org/10.1016/j.jebo.2018.06.014
0167-2681/

The relevance of reinforcement for economic decision making was illustrated by Charness and Levin (2005) in a binary-choice, belief-updating task where mistakes (deviations from optimization under correct Bayesian updating) could be traced to a reinforcement heuristic. In essentially the same paradigm, Achtziger and Alós-Ferrer (2014) found evidence of the conflict between reinforcement and rational optimization in the form of response time asymmetries as predicted by an explicit dual-process model. Recent psychophysiological work (Achtziger et al., 2015) found direct evidence of neural correlates of reinforcement in this paradigm and studied their relation to economic incentives. Further studies relying on this paradigm have examined the interaction of reinforcement and decision inertia (Alós-Ferrer et al., 2016), the influence of framing on reinforcement decisions (Alós-Ferrer et al., 2017), and the regulation of reinforcement processes through motivational interventions (Hügelschäfer and Achtziger, 2017).

In this work, we take a further step in the study of reinforcement heuristics in economic settings by moving beyond binary-choice tasks and studying the explicit relation between reinforcement and myopic payoff maximization in strategic decisions. Hence, we study reinforcement processes in a more complex setting which results in longer decisions times than, e.g., standard neuropsychological experiments. We concentrate on two-player, 3×3 asymmetric normal form games. In this setting, the microeconomics literature has devoted a great deal of attention to myopic best reply, a behavioral rule which maximizes the own payoff assuming the other player will repeat her action, and which can be assumed to have a more deliberative/controlled nature than reinforcement.

Previous work has analyzed paradigms where, by design, reinforcement and more deliberative behavior could either conflict or be aligned (e.g. Achtziger and Alós-Ferrer, 2014). In other settings, however, there might be a great degree of overlap between the prescriptions of reinforcement and those of myopic best reply. In the present work we specifically explore to what extent reinforcement can act as a shortcut for (apparent) optimization in strategic situations with explicit feedback. Suppose a player's last action delivered the best possible payoff. Reinforcement will then prescribe to repeat the choice (win-stay). By definition, however, that choice is the best reply if the opponent stays put. Likewise, suppose the last action did not deliver the best possible payoff. Reinforcement will prescribe to choose a different action (lose-shift). But, again by definition, the current choice cannot be the myopic optimum, and hence myopic best reply also prescribes to shift. In principle, the "shift" prescribed by reinforcement is arbitrary, but if payoffs are observable (as, e.g., if the payoff table is known), the observable maximum becomes salient and the shift will often be in its direction, leading to an apparent myopic best reply. In view of these observations, we postulate that reinforcement processes might often act as cognitive shortcuts resulting in choices indistinguishable from myopic best reply.[1]

If choices coming from reinforcement and those prescribed by myopic best reply cannot be distinguished, how can this hypothesis be substantiated? There are two possible avenues. The first relies on response times. As explained above, reinforcement processes are automatic and can be expected to lead to shorter response times than alternative processes. In contrast, myopic best reply involves explicit maximization and can be assumed to be deliberative, hence relatively slow. Hence, if the choices favored by both processes are actually due to the involvement of reinforcement processes, one should expect shorter response times (compared to other choices), while if they are due to explicit maximization, response times should be longer.

The second avenue is bygone payoffs. Suppose a player obtains a payoff which is not the maximum possible one given the opponent's strategy. If that maximum payoff is observable, the deviation with respect to it is a cardinal measure of experienced disappointment. Define experienced regret as the difference between the maximum possible payoff (that of the best reply) and the actually obtained payoff. By definition, regret is zero if and only if the player has chosen a best reply, and strictly positive if not (this will be directly observable in our experiment). Myopic best reply, considered as a behavioral rule, prescribes to change strategy whenever a best reply has not been chosen. In contrast, reinforcement processes are stimulus-response mappings which take the win-loss information as an input. The loss information also carries a measurement of stimulus strength which in turn yields a variation in the responses. Hence, standard formalizations of reinforcement take the probability of a shift to be increasing in the degree of the loss, which is just the experienced regret as defined above. That is, reinforcement should be triggered more often for larger experienced regret. Hence, if observed choices follow from reinforcement processes, one should observe a dependence on experienced regret.

To study these questions, we conducted an experiment ($N = 144$) where participants played 3×3 games against other players repeatedly. In order to isolate the decision processes of interest, in our experiment players had full knowledge of their own payoff tables, but were not aware of the payoffs of the opponents. In this way, we aimed to eliminate a number of potential confounds, as e.g. imitation or social preferences. Also, in this simple design the maximum (bygone) payoff associated with a choice is directly observable, and regret is simply the difference between the highest payoff associated with the opponent's choice and the actually received payoff. To make this even simpler, each of the different payoff tables used in the experiment contain only three different numerical payoffs, hence maximum payoffs and regret levels are easily observable.

The experiment was divided in short parts of 13 rounds each, and after each part both the opponent and the payoff table were changed. This was an explicit design decision in order to prevent convergence or long-run effects, and rather concentrate on the decision processes.

[1] Indeed, the obvious evolutionary reason for the existence of automatic processes is that in certain situations they are adapted and support (near-)optimal decisions while saving cognitive costs.