Journal of Physiology - Paris 105 (2011) 36–44

Contents lists available at SciVerse ScienceDirect

Journal of Physiology - Paris

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

# A reinforcement learning approach to instrumental contingency degradation in rats

Alain Dutech [a], Etienne Coutureau [b,c], Alain R. Marchand [b,c,*]

[a] LORIA/INRIA, Campus Scientifique, BP 239, 54506 Vandoeuvre les Nancy, France
[b] Univ. Bordeaux, INCIA, CNRS, UMR 5287, F-33400 Talence, France
[c] CNRS, INCIA, UMR 5287, F-33076 Bordeaux Cedex, France

ARTICLE INFO

*Keywords:*
Rats
Instrumental
Prefrontal cortex
Contingency degradation
Simulation
Model-free learning
SARSA

ABSTRACT

Goal-directed action involves a representation of action consequences. Adapting to changes in action-outcome contingency requires the prefrontal region. Indeed, rats with lesions of the medial prefrontal cortex do not adapt their free operant response when food delivery becomes unrelated to lever-pressing. The present study explores the bases of this deficit through a combined behavioural and computational approach. We show that lesioned rats retain some behavioural flexibility and stop pressing if this action prevents food delivery. We attempt to model this phenomenon in a reinforcement learning framework. The model assumes that distinct action values are learned in an incremental manner in distinct states. The model represents states as *n*-uplets of events, emphasizing sequences rather than the continuous passage of time. Probabilities of lever-pressing and visits to the food magazine observed in the behavioural experiments are first analyzed as a function of these states, to identify sequences of events that influence action choice. Observed action probabilities appear to be essentially function of the last event that occurred, with reward delivery and waiting significantly facilitating magazine visits and lever-pressing respectively. Behavioural sequences of normal and lesioned rats are then fed into the model, action values are updated at each event transition according to the SARSA algorithm, and predicted action probabilities are derived through a softmax policy. The model captures the time course of learning, as well as the differential adaptation of normal and prefrontal lesioned rats to contingency degradation with the same parameters for both groups. The results suggest that simple temporal difference algorithms with low learning rates can largely account for instrumental learning and performance. Prefrontal lesioned rats appear to mainly differ from control rats in their low rates of visits to the magazine after a lever press, and their inability to initially detect weak contingency changes.



## 1. Introduction

Goal-directed behaviour requires a representation of action outcome and an ability to adapt the action when this outcome changes. In rodents as in humans, the prefrontal cortex contributes to these functions (Balleine and O'Doherty, 2010). Rats with lesions of the medial prefrontal cortex (mPFC) actually learn an instrumental task (lever pressing for a food reward) at a normal rate, but the response acquired appear insensitive to tests of goal-directed behaviour (Killcross and Coutureau, 2003; Coutureau et al., 2009). The mPFC is required to adapt to contingency degradation, i.e. a weakening of the correlation between food delivery and lever pressing (Hammond, 1980; Balleine and Dickinson, 1998), as has been shown in a design in which the outcome is equally probable in the presence or absence of an instrumental action (Corbit and Balleine, 2003). The neural mechanisms of such a deficit in mPFC-lesioned rats are still poorly understood. Adaptation to contingency requires a learning process that integrates novel observations of unpredicted reward deliveries with a previously acquired action-reward association. As such, it may be described within the reinforcement learning framework (Sutton and Barto, 1998). Although reinforcement learning processes occurring in the striatum have been proposed to underlie instrumental learning (Joel et al., 2002; Kim et al., 2009), the role of the prefrontal cortex in this learning remains elusive (Daw et al., 2005; Frank and Claus, 2006; Maia, 2009).

We have recently demonstrated the involvement of dopaminergic mechanisms within the prelimbic area of the mPFC in the adaptation to contingency changes (Naneix et al., 2009). Dopamine signals from ventral midbrain dopaminergic neurons are known to be modulated by uncertainty and delays in rewards delivery (Fiorillo et al., 2003; Kobayashi and Schultz, 2008). Thus, new learning could be driven by the delivery of non-contingent rewards that occur in the absence of lever pressing and elicit a dopaminergic prediction error signal. Such a signal is indeed at the root of

* Corresponding author. Address: CNRS, UMR 5287, INCIA, Bât. B2 av. Facultés, F-33405 Talence Cedex, France. Tel.: +33 5 40 00 24 58; fax: +33 5 40 00 87 43.
*E-mail addresses:* alain.dutech@loria.fr (A. Dutech), etienne.coutureau@u-bordeaux1.fr (E. Coutureau), alain.marchand@u-bordeaux1.fr (A.R. Marchand).

0928-4257/$ - see front matter © 2011 Published by Elsevier Ltd.
doi:10.1016/j.jphysparis.2011.07.017

most reinforcement-learning accounts of instrumental behaviour (Maia, 2009). The fact that prefrontal-lesioned rats do not adapt to contingency changes may result from several factors which are difficult to disentangle using purely behavioural criteria. These factors include for instance a loss of flexibility, a different sensitivity to internal or external events, an impaired memory for their own actions or an imperfect representation of the various states of the environment. Within a reinforcement learning framework, it is possible to integrate several of these factors as parameters in a model and to fit the model to the rats' behaviour, in an attempt to characterize normal and lesioned rats using different parameter sets. Moreover, this approach suggests novel ways to analyse the rat's behaviour, to determine whether behaviour indeed depends on supposedly distinct states that are required for theoretical modelling.

In the present study, we used temporal difference (TD) learning to test the hypothesis that prefrontal-lesioned rats have difficulties in parsing the flow of events so as to detect changing relationship between the rat's own actions and rewards. We examined this issue using a combined behavioural and simulation approach, with the following steps: Behavioural data (Coutureau et al., submitted for publication) were first collected by training normal rats and rats with lesions of the mPFC in a standard operant task, followed by a contingency degradation phase. Then, a detailed analysis of behavioural sequences was conducted to identify differences in behaviour that might underlie deficits in adaptation to contingency changes. Finally, a reinforcement-learning model was developed and trained using real event sequences, in order to determine whether different sets of parameters underlie the behavioural performance of normal and lesioned rats. Identifying such differences in model parameters would provide valuable clues as to the operations performed in the mPFC.

Free operant learning raises special difficulties for reinforcement learning because time is not divided into a series of discrete trials which would provide a natural support for the Markov processes on which the TD algorithm is based (Daw et al., 2006). We attempted to capture working memory with a model that focuses on the span of working memory for successive events, based on event sequences and essentially disregarding time. The model makes use of the SARSA algorithm which may be biologically plausible (Niv et al., 2006) and incorporates real sequences of actions and events to train the model and to adjust model parameters.

## 2. Material and methods

### 2.1. Behavioural data

The behavioural experiments that served as basis for simulation were conducted in a set of eight operant cages (Imetronic, Pessac, France) which allowed online control of reward delivery and time-stamped recording of events such as lever-presses and head entries into the food magazine. They involved a series of instrumental training sessions during which 12 rats bearing neurotoxic lesions of the mPFC (Prelimbic + Infralimbic areas, Fig. 1) and 14 control rats were trained when hungry to lever press for food pellets in a free operant task (Fig. 2). The training phase consisted of two sessions of magazine training and seven sessions of rewarded lever presses with rewards delivered at progressively increasing intervals, from about 3 rewards/min (fixed interval, FI20 schedule) during the first two sessions to an average of 1 reward/min (variable interval, VI60 schedule) during the last four sessions (Fig. 2B). Although most of the lever presses were not rewarded, each pellet obtained resulted from a press on the lever, without delay. The rats were then switched in a test phase to one of two possible new action-outcome contingencies (Yin et al., 2006). In the negative contingency condition, the animals could obtain a pellet by abstaining from pressing the lever for a fixed time (20 s) and a new pellet was delivered every 20 s in the absence of lever presses. In the zero-contingency condition, reward delivery was yoked to that of an animal in the negative contingency condition, and thus independent of lever pressing (Fig. 2C). The results showed that mPFC-lesioned rats persisted in pressing the lever at a high rate in the zero-contingency condition, but not in the negative contingency condition, thereby demonstrating that the behaviour of mPFC-lesioned rats is not inflexible. Rather, they appear unable to detect weak contingency changes.

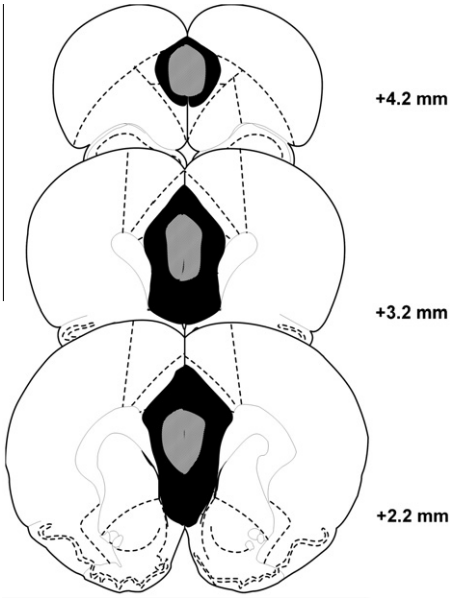


**Fig. 1.** Schematic representation of the medial prefrontal cortex lesions. Black area: maximal lesions; grey area: minimal lesions. Antero-posterior coordinates of frontal sections refer to Bregma. Cell loss occurred in both the prelimbic and infralimbic parts of the medial prefrontal cortex and spared the anterior cingulate cortex.

### 2.2. Modelling the task with TD learning

The continuous nature of the task in free operant behaviour is a challenge for reinforcement learning models, unless using the less tractable framework of semi-Markov Decision Processes (Bradtke and Duff, 1995). At any instant, the rat may choose between various actions such as visiting the food magazine or pressing the lever. However, magazine visits will not lead to the same outcome depending on whether or not the rat has previously pressed the lever. This emphasizes the need to define distinct states in this free operant situation. To circumvent the absence of trial structure, we chose to define states using an event-sequence model that captures the limited capacity of working memory for sequences of consecutive events, without explicit reference to their time of occurrence.

In the conditioning boxes, only lever presses and magazine entries are automatically detected. Other common actions in rats, such as moving around, sniffing, rearing or grooming may occur during "waiting". They can only be inferred from the absence of recorded actions for some time. Three distinct actions were therefore considered: pressing the lever (p), visiting the magazine (v), and other (unrecorded) actions (u) that were assumed to occur without repetition after a specified time interval (parameter $\tau$-other) had elapsed in the absence of other actions. Obviously, a given action is less likely to influence subsequent behaviour after the animal has engaged in other, perhaps unrecorded, activities. Parameter