Neuron
Article

States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning

Jan Gläscher,[1,3,*] Nathaniel Daw,[4] Peter Dayan,[5] and John P. O'Doherty[1,2,6]
[1]Division of Humanities and Social Sciences
[2]Computation and Neural Systems Program
California Institute of Technology, Pasadena, CA 91101, USA
[3]Neuroimage Nord, Department of Systems Neuroscience, University Medical Center Hamburg-Eppendorf, 20246 Hamburg, Germany
[4]Center for Neural Science and Department of Psychology, New York University, NY 10003, USA
[5]Gatsby Computational Neuroscience Unit, University College London, London WC1N 3AR, UK
[6]Trinity College Institute of Neuroscience and School of Psychology, Trinity College Dublin 2, Ireland
*Correspondence: glascher@hss.caltech.edu
DOI 10.1016/j.neuron.2010.04.016

SUMMARY

Reinforcement learning (RL) uses sequential experience with situations ("states") and outcomes to assess actions. Whereas model-free RL uses this experience directly, in the form of a reward prediction error (RPE), model-based RL uses it indirectly, building a model of the state transition and outcome structure of the environment, and evaluating actions by searching this model. A state prediction error (SPE) plays a central role, reporting discrepancies between the current model and the observed state transitions. Using functional magnetic resonance imaging in humans solving a probabilistic Markov decision task, we found the neural signature of an SPE in the intraparietal sulcus and lateral prefrontal cortex, in addition to the previously well-characterized RPE in the ventral striatum. This finding supports the existence of two unique forms of learning signal in humans, which may form the basis of distinct computational strategies for guiding behavior.

INTRODUCTION

One of the most critical divisions in early-20th century animal learning psychology was that between behaviorist notions such as Thorndike's (Thorndike, 1933), that responses are triggered by stimuli through associations strengthened by reinforcement, and Tolman's proposal (Tolman, 1948), that they are instead planned using an internal representation of environmental contingencies in the form of a "cognitive map." Although the original debate has relaxed into a compromise position, with evidence at least in rats that both mechanisms exist and adapt simultaneously (Dickinson and Balleine, 2002), a full characterization of their different learning properties and the way that their outputs are integrated to achieve better control is as yet missing. Here, we adopt specific computational definitions that have been proposed to capture the two different structures of learning. We use them to seek evidence of the two strategies in signals measured by functional magnetic resonance imaging (fMRI) in humans learning to solve a probabilistic Markov decision task.

Theoretical work has considered the two strategies to be model-free and model-based, and has suggested how their outputs might be combined depending on their respective certainties (Daw et al., 2005; Doya et al., 2002). In a model-based system, a cognitive map or model of the environment is acquired, which describes how different "states" (or situations) of the world are connected to each other. Action values for different paths through this environment can then be computed by a sort of mental simulation analogous to planning chess moves: searching forward along future states to evaluate the rewards available there. This is termed a "forward" or "tree-search" strategy. In contrast, a model-free learning system learns action values directly, by trial and error, without building an explicit model of the environment, and thus retains no explicit estimate of the probabilities that govern state transitions (Daw et al., 2005). Because these approaches evaluate actions using different underlying representations, they produce different behaviors in experiments aimed at investigating their psychological counterparts. Most such experiments (Dickinson and Balleine, 2002) study whether animals adapt immediately to changes in the environment. For instance, in classic "latent learning" studies (Tolman and Honzik, 1930), animals are pretrained on a maze, then rewards are introduced at a particular location to probe whether subjects can plan new routes there taking into account previously learned knowledge of the maze layout. The experiment discussed here, though nonspatial, follows this scheme.

Learning in both model-based and model-free strategies is typically driven by prediction errors, albeit with different meaning and properties in each case. A prediction error is a difference between an actual and an expected outcome and this signal is commonly thought of as the engine of learning, as it is used to update expectations in order to make predictions more accurate.

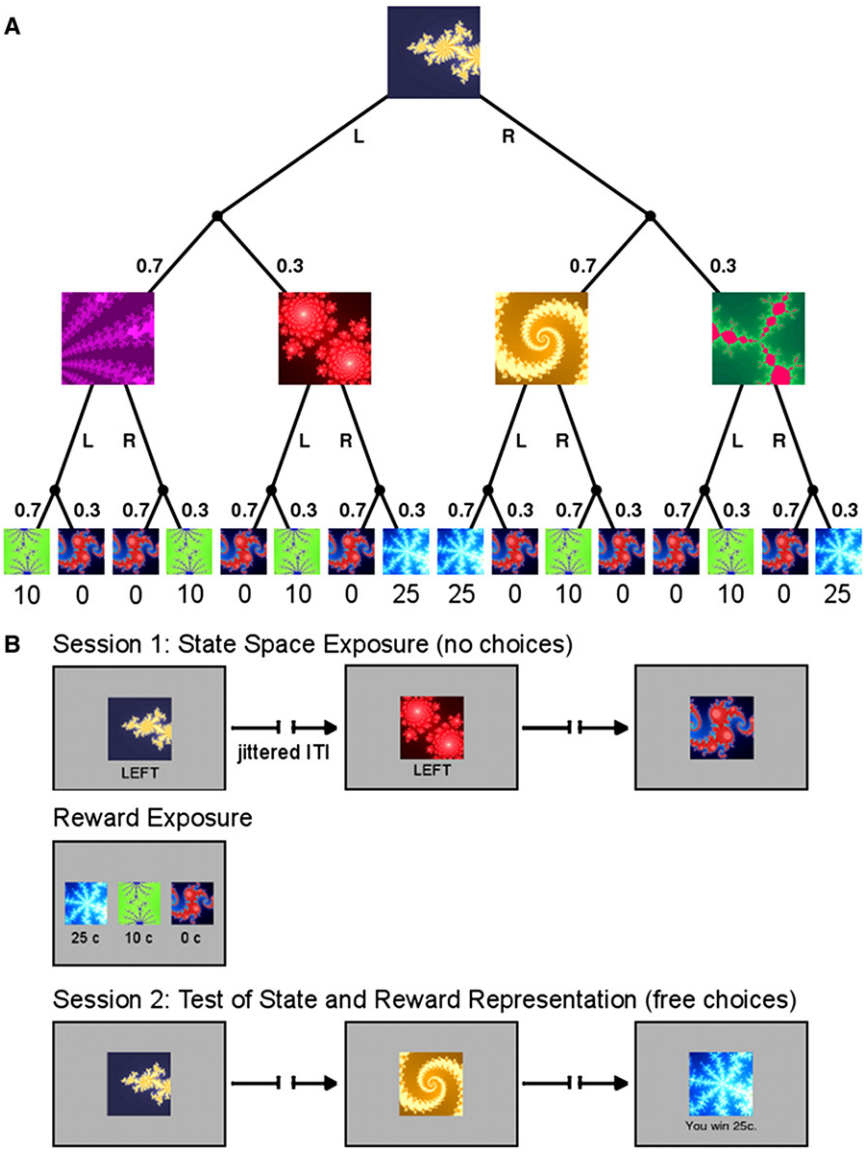

Figure 1. Task Design and Experimental Procedure

(A) The experimental task was a sequential two-choice Markov decision task in which all decision states are represented by fractal images. The task design follows that of a binary decision tree. Each trial begins in the same state. Subjects can choose between a left (L) or right (R) button press. With a certain probability (0.7/0.3) they reach one of two subsequent states in which they can choose again between a left or right action. Finally, they reach one of three outcome states associated with different monetary rewards (0¢, 10¢, and 25¢). (B) The experiment proceeded in two fMRI scanning sessions of 80 trials each. In the first session, subject choices were fixed and presented to them below the fractal image. However, subjects could still learn the transition probabilities. Between scanning sessions subjects were presented with the reward schedule that maps the outcome states to the monetary payoffs. This mapping was rehearsed in a short choice task. Finally, in the second scanning session, subjects were free to choose left or right actions in each state. In addition, they also received the payoffs in the outcome states.

In the case of model-free learning, this error signal (called the reward prediction error, RPE) amounts to the difference between the actual and expected reward at a particular state. In the context of reinforcement learning (RL), this error signal is used to learn values for action choices that maximize expected future reward (Sutton and Barto, 1998). An abundance of evidence from both single-unit recordings in monkeys (Bayer and Glimcher, 2005; Schultz, 1998; Schultz et al., 1997) and human fMRI (D'Ardenne et al., 2008) suggests that dopaminergic neurons in the ventral tegmental area and substantia nigra pars compacta exhibit a response pattern consistent with a model-free appetitive RPE. Furthermore, BOLD signals in the ventral striatum (vStr) show response properties consistent with dopaminergic input (Delgado et al., 2000, 2008; Knutson et al., 2001, 2005), most notably correlating with RPEs (Haruno and Kawato, 2006; McClure et al., 2003; O'Doherty et al., 2003).

Model-based action valuation requires predicting which state is currently expected, given previous states and/or choices. These expectations can be learned using a different prediction error, called the state prediction error (SPE), which measures the surprise in the new state given the current estimate of the state-action-state transition probabilities. The central questions for the current study are whether the human brain computes the SPE as well as the RPE, and, if so, what the different neural signatures of these two signals are. One indication that the brain may compute SPEs is that neural signals marking gross violations of expectations have long been reported, particularly using EEG (Courchesne et al., 1975; Fabiani and Friedman, 1995) and EEG in combination with fMRI (Opitz et al., 1999; Strobel et al., 2008). Unlike the prediction error signals associated with dopamine activity, which are largely reward-focused and associated with model-free RL (Holroyd and Coles, 2002), these respond to incorrect predictions of affectively neutral stimuli. Here, we study quantitatively how state predictions are learned, and seek trial-by-trial neural signals that reflect the dynamics of this learning.

We designed a probabilistic sequential Markov decision task involving choices in two successive internal states, followed by a rewarded outcome state (see Experimental Procedures). The task has the structure of a decision tree, in which each abstract decision state is represented by a fractal image (Figure 1A). In each trial, the participants begin at the same starting state and