EJN fens European Journal of Neuroscience

European Journal of Neuroscience, Vol. 35, pp. 1036–1051, 2012 doi:10.1111/j.1460-9568.2012.08050.x

Habits, action sequences and reinforcement learning

Amir Dezfouli and Bernard W. Balleine
Brain & Mind Research Institute, University of Sydney, Camperdown, NSW 2050, Australia

Keywords: action sequence, goal-directed action, habitual action, reinforcement learning

Abstract

It is now widely accepted that instrumental actions can be either goal-directed or habitual; whereas the former are rapidly acquired and regulated by their outcome, the latter are reflexive, elicited by antecedent stimuli rather than their consequences. Model-based reinforcement learning (RL) provides an elegant description of goal-directed action. Through exposure to states, actions and rewards, the agent rapidly constructs a model of the world and can choose an appropriate action based on quite abstract changes in environmental and evaluative demands. This model is powerful but has a problem explaining the development of habitual actions. To account for habits, theorists have argued that another action controller is required, called model-free RL, that does not form a model of the world but rather caches action values within states allowing a state to select an action based on its reward history rather than its consequences. Nevertheless, there are persistent problems with important predictions from the model; most notably the failure of model-free RL correctly to predict the insensitivity of habitual actions to changes in the action–reward contingency. Here, we suggest that introducing model-free RL in instrumental conditioning is unnecessary, and demonstrate that reconceptualizing habits as action sequences allows model-based RL to be applied to both goal-directed and habitual actions in a manner consistent with what real animals do. This approach has significant implications for the way habits are currently investigated and generates new experimental predictions.

The problem with habits

There is now considerable evidence from studies of instrumental conditioning in rats and humans that the performance of reward-related actions reflects the interaction of two learning processes, one controlling the acquisition of goal-directed actions, and the other of habits (Dickinson, 1994; Balleine & O'Doherty, 2010). This evidence suggests that, in the case of goal-directed actions, action selection is governed by encoding the association between the action and its consequences or outcome. In contrast, the performance of habitual actions depends on forming an association between an action and antecedent stimuli, rather than its consequences. As a consequence, being stimulus bound, habitual actions are relatively inflexible or reflexive, and are not immediately influenced by manipulations involving the outcome. They differ from goal-directed actions, therefore, in two ways: (i) in their sensitivity to changes in the value of the outcome (Adams, 1982); and (ii) in their sensitivity to changes in the causal relationship between the action and outcome delivery (Dickinson *et al.*, 1998).

Although many replications of these effects exist in the literature, we describe, for illustration, data from a recent study in which we were able to observe both effects in the same animals in the same experiment comparing moderately trained and overtrained actions for their sensitivity to devaluation, induced by outcome-specific satiety, and contingency degradation, induced by the imposition of an omission schedule in rats. These data are presented in Fig. 1. Rats were trained to press a lever for sucrose and, after the satiety treatment, given a devaluation test (Fig. 1A). After retraining, they were given the contingency degradation test (Fig. 1B). In the first test, moderately trained rats showed a reinforcer devaluation effect; those sated on the sucrose outcome reduced performance on the lever relative to those sated on another food. In contrast, groups of overtrained rats did not differ in the test. In the second test, rats exposed to the omission contingency were able to suppress the previously trained lever press response to get sucrose, but only if moderately trained. The performance of overtrained rats did not differ from the yoked, non-contingent controls. These data confirm, therefore, the general insensitivity of habitual actions to changes in both outcome value and the action–outcome contingency.

The neural bases of goal-directed and habitual actions have also received considerable attention. Based on the results of a number of studies it appears that, whereas the acquisition of goal-directed actions is controlled by a circuit involving medial prefrontal cortex and dorsomedial striatum (DMS, or caudate nucleus), habitual actions involve connections between sensory-motor cortices and dorsolateral striatum (DLS, or putamen; Balleine *et al.*, 2007; Balleine & O'Doherty, 2010). Importantly, these two forms of action control have been argued to be at least partly competitive; at the level of the striatum, lesions or pharmacological manipulations aimed at blocking activity in the DMS render goal-directed actions habitual, whereas blocking activity in the DLS renders otherwise habitual actions goal-

Correspondence: Dr B.W. Balleine, as above.
E-mail: bernard.balleine@sydney.edu.au

The research described here and the preparation of this paper was supported by grants from the Australian Research Council #FL0992409, the National Health & Medical Research Council #633267, and the National Institute of Mental Health #MH56446.

Received 31 October 2011, revised 18 January 2012, accepted 23 January 2012

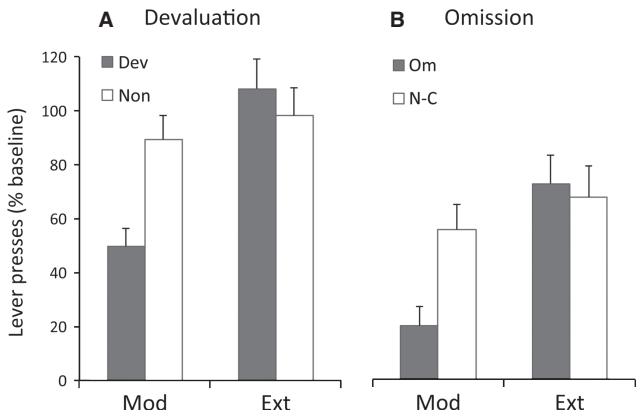

FIG. 1. Four groups of rats ($n = 8$) were trained to lever press for a 20% sucrose solution on random-interval schedules (RI1, RI15, RI30, RI60), with moderately trained rats allowed to earn 120 sucrose deliveries and overtrained rats 360 sugar deliveries (the latter involving an additional eight sessions of RI60 training with 30 sucrose deliveries per session). (A) For the devaluation assessment, half of each group was then sated either on the sucrose or on their maintenance chow before a 5-min extinction test was conducted on the levers. Moderately trained rats showed a reinforcer devaluation effect; those sated on the sucrose outcome reduced performance on the lever relative to those sated on the chow. In contrast, groups of overtrained rats did not differ in the test. Statistically we found a training × devaluation interaction ($F_{1,28} = 7.13$, $P < 0.05$), and a significant devaluation effect in the moderately trained ($F_{1,28} = 9.1$, $P < 0.05$) but not in the overtrained condition ($F < 1$). (B) For the contingency assessment, after devaluation all rats received a single session of retraining for 30 sucrose deliveries before the moderately trained and overtrained rats were randomly assigned to either an omission group or a yoked, non-contingent control group. During the contingency test the sucrose outcome was no longer delivered contingent on lever pressing, and was instead delivered on a fixed time 10-s schedule. For rats in the omission groups, responses on the lever delayed the sucrose delivery by 10 s. Rats in the yoked groups received the sucrose at the same time as the omission group, except there was no response contingency in place. As is clear, rats exposed to the omission contingency in the moderately trained group suppressed lever press performance relative to the non-contingent control, whereas those in the overtrained groups did not. Statistically, there was a training × degradation interaction ($F_{1,28} = 5.1$, $P < 0.05$), and a significant degradation effect in the moderately trained ($F_{1,28} = 7.8$, $P < 0.05$) but not in the overtrained condition ($F < 1$).

directed by the measures described above (Yin *et al.*, 2005, 2006). In addition, there is clear evidence for the involvement of a number of neuromodulators associated with these networks, particularly the input from dopamine neurons in the substantia nigra pars compacta to the DLS, which has been found to play an essential role in learning-related plasticity in the dorsolateral region and, indeed, in habits (Reynolds *et al.*, 2001; Faure *et al.*, 2005). Generally, therefore, although there is tremendous complexity in the circuitry and in the necessary conditions for plasticity associated with the goal-directed and habit-learning processes, they appear to depend on distinct, parallel cortico-basal ganglia networks that likely constitute functional loops connecting cortex, striatum and midbrain regions with feedback to the cortex via midline thalamus (Alexander & Crutcher, 1990; Kelly & Strick, 2004).

The functional specificity of these anatomical loops has provided fertile ground for computational models of action control (Frank & O'Reilly, 2006; O'Reilly *et al.*, 2010). Given the functional complexity of the circuitry involving the dorsal striatum, a key issue has been how well these computational models capture the difference between goal-directed and habitual actions. Perhaps the most influential approach has come from recent developments in a sub-field of computational theories collectively known as reinforcement learning (RL) (Sutton & Barto, 1998). The core feature of such RL models is that, in order to choose optimally between different actions, an agent needs to maintain internal representations of the expected reward available on each action and then subsequently choose the action with the highest expected value. Importantly, two forms of RL have been argued to be necessary accurately to model goal-directed and habitual actions (Daw *et al.*, 2005), known generally as 'model-based' and 'model-free' RL, respectively (Daw *et al.*, 2005; Rangel *et al.*, 2008; Redish *et al.*, 2008). The essential difference between these forms of RL lies in how they compute the values of actions. As its name implies, model-based RL depends on developing a model of the world from which values for different actions are worked out on-line by taking into account knowledge about the rewards available in each state of the world in which an animal finds itself, and the transition probabilities between those states. It then works iteratively across states to establish the value of each available action option, much as a chess player might work out which chess move to make by thinking through the consequences of various possible moves. On this view, the performance of an action depends both on its relationship to the final (what might be called the 'goal') state and the value of that state. As such, model-based RL can be applied to goal-directed actions in a relatively straightforward manner and, indeed, this model has been applied to prefrontal cortical control of decision-making and executive functions generally (Daw *et al.*, 2005).

In contrast, model-free RL does not form a model of the world but rather caches action values within states, allowing an agent to select an action based on its reward history rather than its consequences. According to the model-free RL approach, action selection involves two interacting components: one that learns to predict reward value; and another that forms a policy as to which action to select in a particular state. The goal of the model is to modify the policy stored in the actor such that over time, the action selected in a state is associated with the highest predicted reward. This is accomplished by means of a prediction error signal that computes the changes in predicted reward as the agent moves from state to state. That signal is then used: (i) to update the value predictions for each state; and (ii) to update the policy in that state. If an action moves the agent to a state predicting greater reward (a positive prediction error), then the probability of choosing that action in the future is increased. Conversely, if an action moves the agent to a state predicting less reward (a negative prediction error) then the probability of choosing that action again is decreased. As such, the policy and value signal function to bring together a state–action associative process with an error correction feedback signal to acquire and shape actions, something that strongly approximates the classic stimulus–response/reinforcement theory long postulated to support habit learning in the behavioralist era (Hull, 1943; Spence, 1956). Furthermore, analogies have been drawn in several influential papers between the anatomy and connections of the basal ganglia and possible neural architectures for implementing RL models (Montague *et al.*, 1996; O'Doherty *et al.*, 2004). A key aspect of these suggestions is that the ventral striatum learns values, whereas the dorsal striatum acquires the policies and, indeed, both of these regions receive strong inputs from the midbrain dopamine neurons thought to mediate the critical error prediction signal (Schultz *et al.*, 1997; Schultz & Dickinson, 2000).

Although Daw *et al.* (2005) proposed that model-free RL, such as that encapsulated in the actor–critic framework, can be applied directly and successfully to habits (and so were motivated to introduce a dual-process model of action control), on closer inspection this turns out not to be true. It is clear that the model-free approach anticipates the insensitivity of habitual actions to the effects of reinforcer devaluation. This is because the values that determine action selection are cached within the state, meaning that a policy can only be changed once feedback regarding a change in value is given. As tests of habit

learning are conducted in extinction, no such feedback is provided; so far, so good. The problem for traditional RL approaches to habit, however, does not lie in their application to devaluation, but in their application to contingency degradation. As shown in Fig. 1B, after a period of overtraining, during which a particular policy should have been very strongly established, the change in contingency – actually, in the case of omission, the reversal of the contingency – fails to induce any immediate change in lever press performance in the rats. From the perspective of the model-free RL, however, the error signal used to alter the policy should function perfectly well to modify performance; indeed, when an omission schedule is imposed at the end of a long period of overtraining, at which point any error should be at a minimum, the impact of both the negative error signal generated as a consequence of performing the habitual action (which now delays the reward) and the large positive error generated when the animal stops performing the action (which allows reward to be delivered) should be anticipated to have a correspondingly large effect on performance. As shown in Fig. 1B, however, the performance of real animals fails to adjust to the change in contingency, opposing this prediction of the model.

Variants of model-free RL can be designed to address the failure of this approach to deal with habits. For example, it can be argued that long periods of exposure to the stationary situation of training make the update rate of values and policies slow (Dayan *et al.*, 2000; Dayan & Kakade, 2001), rendering them rigid and reluctant to change in new environmental conditions. According to this account, contingency manipulations do not affect behavior because, after overtraining, animals cannot learn new values and policies. However, because the test is conducted in extinction, in which there is no feedback provided about the change in value, this argument cannot also explain reinforcer devaluation experiments, and habits of this kind cannot be attributed to the effect of low learning rates.

It is possible that other variants of model-free RL could be designed that account for insensitivity to both contingency degradation and reinforcer devaluation phenomena. Alternatively, perhaps the reason RL has difficulty developing a common account for both of these effects is rooted in the general equation of the representation of goal-directed and habitual actions within the model-based and model-free approach. In this paper, we are concerned with exploring this second possibility. Habits, we contend, are more complex than goal-directed actions and reflect the association of a number of, initially variably produced, actions into rapidly executed action sequences. For modeling sequences of this kind we develop a model in which we essentially preserve model-based RL and propose a new theoretical approach that accommodates both goal-directed and habitual action control within it. This model, we will show, is not only more parsimonious, it also provides a means of modeling sequences. Importantly, it also provides a model that uses the prediction error signal to construct habits in a manner that accords with the insensitivity of real animals to both reinforcer devaluation and contingency degradation. Finally, it makes several important predictions about the structure of habit learning and the performance of habitual actions generally that we believe will provide a guide to future experimentation in this area, and we discuss these issues in the final section.

Sequence learning: from closed-loop to open-loop action control

The flexibility of goal-directed actions reflects, therefore, the need for immediate, or at least rapidly acquired, solutions to new problems and, indeed, evidence suggests that in novel environments there is a strong tendency for animals to generate behavioral variation and to resist immediately repeating prior actions or sequences of actions (Neuringer, 2004; Neuringer & Jensen, 2010). Of course, the need to explore new environments requires behavioral variation; once those solutions are found, however, exploiting the environment is best achieved through behavioral stability, i.e. by persisting with a particular behavioral response. It is important to recognize that, with persistence, actions can change their form, often quite rapidly. Errors in execution and the inter-response time are both reduced and, as a consequence, actions previously separated by extraneous movements or by long temporal intervals are more often performed together and with greater invariance (Willingham *et al.*, 1989; Buitrago *et al.*, 2004a,b). With continuing repetition these action elements can become linked together and run off together as a sequence, i.e. they can become chunked (Terrace, 1991; Graybiel, 1998). Practice appears, therefore, to render variable, flexible, goal-directed actions into rapidly deployed, relatively invariant, components of action sequences, suggesting that an important way in which the form of a goal-directed action can change as it becomes habitual is via the links that it forms with other actions to generate sequences.

Generally, sequence learning and habit formation are assessed using different behavioral tasks, and are considered to be distinct aspects of automatic behavior. However, neural evidence suggests that habit learning and action sequence learning involve similar neural circuits: thus, in the first stages of learning both sequences and goal-directed actions appear to involve the prefrontal cortex and associative striatum (Miyachi *et al.*, 1997; Lehéricy *et al.*, 2005; Poldrack *et al.*, 2005; Bailey & Mair, 2006, 2007). However, as they become more routine in their expression, performance of both habits and action sequences involves the sensorimotor striatum (Miyachi *et al.*, 1997, 2002; Yin *et al.*, 2004, 2006; Lehéricy *et al.*, 2005; Poldrack *et al.*, 2005). Evidence suggests that a cortico-striatal network parallel to that implicated in goal-directed action involving sensorimotor cortices together with the DLS in rodents mediates the transition to habitual decision processes associated with stimulus–response learning (Jog *et al.*, 1999; Barnes *et al.*, 2005). Changes in the DLS appear to be training related (Costa *et al.*, 2004; Hernandez *et al.*, 2006; Tang *et al.*, 2007), and to be coupled to changes in plasticity as behavioral processes become less flexible (Costa *et al.*, 2006; Tang *et al.*, 2007). Correspondingly, whereas overtraining causes performance to become insensitive to reinforcer devaluation, lesions of DLS reverse this effect rendering performance once again sensitive to devaluation treatments (Yin *et al.*, 2004). Likewise, muscimol inactivation of DLS has been found to render otherwise habitual performance sensitive to changes in the action–outcome contingency (Yin *et al.*, 2006) and to abolish the stimulus control of action selection during discrimination learning (Featherstone & McDonald, 2004, 2005; Balleine *et al.*, 2009). Importantly, there is recent evidence that the DLS is involved in sequence learning. For example, Jin & Costa (2010) trained rats to make a fixed number of lever presses to earn food, and not only found evidence of response sequencing and chunking on this task, they also recorded neuronal activity in the DLS where they found phasic changes in the activity of neurons associated with either the first or last response in a sequence. They also observed a reduction in neuronal ensemble variability in the DLS at sequence boundaries. This start/stop activity was not a function of action value or of timing.

The most important feature of sequence learning is 'the interdependency of actions' (Shah, 2008). Through the process of sequence learning, action control becomes increasingly dependent on the history of previous actions and independent of environmental stimuli, to the point that, given some triggering event, the whole sequence of actions is expressed as an integrated unit. Take, for example, a typical

sequence-learning experiment such as the serial reaction time task (SRTT; Nissen & Bullemer, 1987). In this task a subject is required to elicit a specific action in response to a specific cue. For example, cues can be asterisks that are presented on a computer screen, and the corresponding responses require the subject to press the keys that spatially match the cues' positions that can appear in either a random or sequential order. In the 'sequential trials' condition, the position of the cues is repeated in a pattern such that the position of the next stimulus can be predicted given the previous one. In the 'random trials' condition, stimuli are presented in random order and, thus, given the previous stimuli, the next one cannot be predicted.

On a simple conception of the learning process in the SRTT, the subject takes an action and, if it is successful, the association between that action and the cue strengthens. Then, the subject waits for the next cue, and takes the action that has the strongest association with the cue. From a control theory point of view, this learning process can be characterized as a 'closed-loop control system' in which, after each response, the controller (here the subject), receives a 'feedback' signal from the environment to guide future actions. In the SRTT, after taking an action the subject receives two types of feedback: reward feedback and state feedback. Reward feedback is the reward received after each response, for example, a specific amount of juice, and is used for learning stimulus–response associations. State feedback is given by the presentation of the next cue after the response that signals the new state of the environment, and is used by the subject to select its next action. The term 'closed-loop' commonly refers to the loop created by the feedback path (Fig. 2A). In the absence of this feedback, the control system is called 'open-loop' (Fig. 2B; Astrom & Murray, 2008), according to which the actions that the subject takes are not dependent on the presented cues or the received rewards.

Clearly, closed-loop control is crucial for learning. Without feedback, the agent cannot learn which response is correct. It also needs state feedback from the environment because the correct response differs in each state and, as such, without knowing the current state, the agent cannot make the correct response. However, in the 'sequential trials' condition of the SRTT, the subject could potentially maintain a high level of performance without using the state feedback; indeed, evidence suggests that is exactly what they do in accord with open-loop control. Reaction time, defined as the time between the stimulus onset and the onset of the behavioral response, is the primary measure in the SRTT. Evidence from rodents (Schwarting, 2009), non-human primates (Hikosaka *et al.*, 1995; Matsuzaka *et al.*, 2007; Desmurget & Turner, 2010) and humans (Nissen & Bullemer, 1987; Keele *et al.*, 2003) suggests that, after training, reaction times are shorter in the 'sequential trials' than the 'random trials' condition. In fact, if the subject is permitted to respond during the inter-trial delay, then the reaction time can even be negative, i.e. the next action is often executed before the presentation of the next stimulus (Matsuzaka *et al.*, 2007; Desmurget & Turner, 2010). Matsuzaka *et al.* (2007) reported that, with increasing training, the number of these 'predictive responses' increases up to the point that in almost all trials, the subject responds in a predictive manner without the visual cue (see also Miyashita *et al.*, 1996), indicating that the number of predictive responses increases as a sequence becomes well learned.

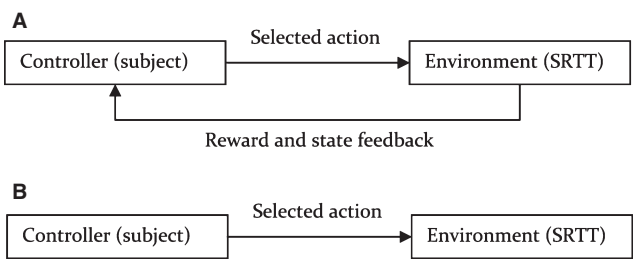

FIG. 2. (A) A closed-loop control system. After the controller executes an action it receives cues regarding the new state of the environment and a reward. (B) An open-loop control system in which the controller does not receive feedback from the environment. SRTT, serial reaction time task.

The occurrence of very short or negative reaction times in 'sequential trials' implies that, after sufficient learning, selection of an action is mostly dependent on the history of previous actions and less dependent on the external stimuli (visual cues). In fact, even if the subject does not respond predictively before stimulus presentation, it is clear that the decision as to which action to take on the next trial is made 'before' stimulus presentation. In a sequential button-push task, Matsumoto *et al.* (1999) trained a monkey to execute a series of three button pushes in response to illumination of the buttons in a fixed cued sequence. After training, the monkey was tested in a condition in which the third button in the sequence was located in a position different from its position during training. They found that, during the first and sometimes the second trial, the monkeys would continue to push the third button of the learned sequence even if one of the other targets was illuminated. Similarly, Desmurget & Turner (2010) reported when the first stimuli of a random trial followed, by coincidence, the pattern of stimuli from a learned sequence, the animal responded as if the next stimuli will be drawn from the learned sequence.

It appears, therefore, that, during the first stages of training, the subject learns the association between cues and responses. At this stage, action selection is under closed-loop control and relies on the observation of the cues. In the case of random trials, action selection remains closed-loop through the course of learning. In 'sequential trials', however, with further learning, action selection switches to open-loop control in which the execution of successive actions is not related to the current state of the environment, something that leads to the expression of action 'chunks' (Lashley, 1951; Sakai *et al.*, 2003). When actions are expressed in chunks, both state identification, based on visual cues, and action evaluation appear to be bypassed. Endress & Wood (2011), for example, note that successful sequence learning requires view-invariant movement information, i.e. rather than depending on the relation between visual cues and movement information in allocentric space, as goal-directed actions do (Willingham, 1998), sequential movements appear to depend on position-based encoding in egocentric space. Hence, chunked in this way, the performance of sequential movements is largely independent of environmental feedback, allowing for very short reaction times in the open-loop mode.

Another line of evidence consistent with the cue-independency notion of habitual behavior comes from place/response learning tasks in animals (Tolman *et al.*, 1946; Restle, 1957). In this type of task, rats begin each trial at the base of a T-maze surrounded by environmental cues (e.g. windows, doors), and are trained to find food at the end of one arm (e.g. the right, or east, arm). Following this training period, they are given probe tests in which the maze is rotated 180° (with respect to the cues), and thus the start point will be at the opposite side of the maze. Results show that after moderate training, at the choice point the animal turns in the opposite direction to that previously learned (i.e. towards the west arm; place strategy), suggesting that action control is state-dependent and based on the environmental cues (closed-loop action control). However, after overtraining, rats switch and at the test they take the same action that they learned in the initial training (i.e. they turn right; a response strategy), indicating that overtraining renders action selection at the choice point independent of

the environmental cues and the state identification process (open-loop action control; Ritchie *et al.*, 1950; Packard & McGaugh, 1996). Similarly, in more complex mazes in which a sequence of actions is required to reach the goal, removal of environmental cues does not affect performance of a learned sequence of egocentric movements (body turns), but disrupts the use of a place strategy (Rondi-Reig *et al.*, 2006). Learning the maze in a cue-deficient environment, but not in a cue-available environment, in which decision-making should minimally rely on state-guided action control is impaired by inactivation of DLS (Chang & Gold, 2004). Few studies have addressed functional differences between the DLS and DMS in place/response learning; however, in general it seems that the DMS is involved in goal-directed decision-making (the place strategy), and the DLS is involved in habitual responses (the response strategy; Devan & White, 1999; Devan *et al.*, 1999; Yin & Knowlton, 2004; Moussa *et al.*, 2011), consistent with the role of these striatal sub-regions in instrumental conditioning and SRTT.

Based on the mentioned similarities in neural and behavior aspects of increasing automaticity in SRTT ('sequential trials'), maze learning and instrumental conditioning, we assume that action sequence formation is the underlying process of these modalities of habitual behavior. In order to formalize this assumption, in the next section we use RL to provide a normative approach to modeling changes in performance during action sequence learning. Next, in the Results, we show how this model applies to the different forms of habitual behavior.

An alternative perspective: RL and action sequence formation

Overview of the decision-making process

RL is a computational approach to learning different types of instrumental associations for the purpose of maximizing the accrual of appetitive outcomes and minimizing aversive ones (Sutton & Barto, 1998). In the RL framework, stimuli are referred to as 'states', and responses as 'actions'. Hereafter we use these terms interchangeably.

A typical RL agent utilizes the following components for the purpose of learning and action selection: (i) state identification – the agent identifies its current state based on the sensory information received from its environment (e.g. visual cues); (ii) action selection – given its current state, and its knowledge about the environment, the agent evaluates possible actions then selects and executes one of them; (iii) learning – after executing an action, the agent enters a new state (e.g. receives new visual cues) and also receives a reward from the environment. Using this feedback, the agent improves its knowledge about the environment. This architecture is a closed-loop decision-making process because the action-selection process is guided by the current state of the agent, which is identified based on sensory inputs received from the environment. As we discussed in the previous section, action selection in sequence learning is not guided by environmental stimuli, and so does not require a state identification process. To explain sequence learning, therefore, we need to modify this framework. In the following sections we will introduce a mixed open-loop/closed-loop architecture for this purpose. Before turning to that issue, we shall first take a closer look at the learning process in RL.

Average reward RL

We denote the state of the agent at time t as s_t. In each state, there is a set of possible actions, and the agent selects one of them for execution, that we denote with a_t. The agent spends d_t time steps in state s_t (commonly referred to as the 'state dwell time') and, after that, by taking a_t, it enters a new state, s_{t+1}, and receives reward r_t. The next state of the agent, the amount of reward received after taking an action, and the state dwell times, depend on the dynamics of the environment, which are determined, respectively, by the transition function, the reward function and transition time function. The transition function, denoted by $P_{sa}(s')$ indicates the probability of reaching state s' upon taking action a in state s. $R(s)$ denotes the reward function, which is the amount of reward the agent receives in state s. Finally, $D(s)$, the transition time function, is the time spent in state s (dwell time). The time that the agent spends in a state is the sum of the time it takes the agent to make a decision, and the time it takes for new stimuli to appear after executing an action.

The goal of the agent is to select actions that lead to a higher average reward per time step, and this is why this formulation of the RL is called 'average reward RL' (Mahadevan, 1996; Tsitsiklis & Roy, 1999; Daw & Touretzky, 2000, 2002). This average reward, denoted by $\bar{R}$, can be defined as the total rewards obtained, divided by the total time spent for acquiring those rewards:

$$\bar{R} = \frac{r_0 + r_1 + r_2 + \ldots}{d_0 + d_1 + d_2 + \ldots}. \quad (1)$$

To choose an action amongst several alternatives, the agent assigns a subjective value to each state–action pair. This subjective value is denoted by $Q(s, a)$, and represents the value of taking action a in state s (Watkins, 1989). These Q-values are learned such that an action with a higher Q-value leads to more reward in a shorter time compared with an action with a lower Q-value.

The first determinant of $Q(s, a)$ is the immediate reward that the agent receives in s, which is $R(s)$. Besides the immediate reward the agent receives, the value of the next state the agent enters is also important: actions through which the agent reaches a more valuable state are more favorable. Thus, assuming that the agent reaches state s' by taking action a, the value of the next state, $V(s')$, is the second determinant of $Q(s, a)$ and is assumed to be proportional to the reward the agent gains in the future by taking its 'best action' in the state s'. In general, for any state s, $V(s)$ is defined as follows:

$$V(s) = \max_a Q(s, a). \quad (2)$$

The final determinant of $Q(s, a)$ is the amount of time the agent spends in the state. If the agent spends a long time in a state, then it will lose the opportunity of gaining future rewards. In fact, losing $D(s)$ time steps in state s is equal to losing $D(s)\bar{R}$ reward in the future. Given these three determinants, the value of taking an action in a state can be computed as follows:

$$Q(s, a) = R(s) - D(s)\bar{R} + E[V(s')] \quad (3)$$

where the expectation in the last term is over s':

$$Q(s, a) = R(s) - D(s)\bar{R} + \sum_{s'} P_{sa}(s')V(s') \quad (4)$$

As the above equation implies, computing Q-values requires knowledge of the transition probabilities, the reward functions and the state dwell times, which together constitute a 'model of the environment'. However, without a prior model of the environment, the agent can estimate these quantities through its experience with the environment. For example, $R(s)$ can be estimated by averaging immediate reward received in state s. In the same manner, $D(s)$ can be

computed as the average of waiting times in state s. An estimation of $P_{sa}(s')$ can be made by counting the number of times taking action a in state s leads to state s'. Given the model of the environment, Q-values can be derived from Eqn 4 using dynamic programming algorithms, such as 'value-iteration' (Puterman, 1994; Mahadevan, 1996). Because these methods of value computation rely on the model of the environment, they are called 'model-based' value computation methods. Using these state–action pairs, the agent can guide its actions toward ones that lead to a higher average reward rate.

Returning to the overview of the decision-making process in RL, in (i) the agent identifies its current state, s_t and then feeds s_t into Eqn 4, allowing the value of different actions, Q-values, to be computed. These Q-values guide the action-selection process, and the agent takes the appropriate action (ii). By taking an action, the agent enters a new state, receives a reward, and measures the time from entering the previous state, s_t, to entering the new state, s_{t+1}. Finally, using these quantities, the model of the environment is updated (iii).

Action sequence formation

When an agent starts learning in a new environment all the decisions are based on model-based action selection, i.e. after entering a new state, the agent computes Q-values using the process introduced in the previous section and chooses one of the actions that tends to have the higher Q-value. Under certain conditions, however, it may be more beneficial for the agent to execute actions in a sequence without going through the action-selection process. First, we discuss the process of sequence formation and, in the next section, how action sequences interact with the model-based action selection.

We start by reviewing the environmental conditions in which action sequences form. Figure 3 shows three environments in which, by taking action A_1 in state S, the agent enters state S' or S'' with equal probability. In states S' and S'' two actions are available, A_2 and A'_2. Figure 3A provides an example of the kind of environment in which an action sequence forms, i.e. in states S' and S'', action A_2 is the best action. An example of this environment is the situation in which pressing a lever (action A_1) leads to the illumination of, say, a light (state S') or a tone (state S'') with equal probability, both of which signal that by entering the magazine (action A_2), the rat can gain a desirable outcome, and by not entering the magazine (action A'_2) it gains nothing. As a consequence, after taking action A_1 in S the agent does not need to check the upcoming state but can execute A_2 irrespective of the next state, either S' or S'' (light or tone). In this situation, actions A_1 and A_2 form an action sequence consisting of A_1 and A_2. Hereafter, we call these action sequences 'macro actions', and denote them, for example, in this situation with $\{A_1A_2\}$. Actions that are not macro, for example A_1 or A_2, are called 'primitive actions'.

Figure 3B shows a situation in which an action sequence does not form. In state S', action A_2 is the best action, but in state S'', action A'_2 is the best. In the context of the previous example, illumination of the light indicates that by entering the magazine, the animal will gain a desirable outcome; but presentation of the tone indicates that entering the magazine is not followed by a desirable outcome. Here, after taking A_1 in S, the animal cannot select an action without knowing the upcoming state, and thus a macro action does not form.

Figure 3C shows a more challenging example. In state S', A_2 is the best action. In state S'', A_2 is not the best action, but it is slightly worse than the best action, A'_2 (e.g. two drops of a liquid reward, vs. one drop of liquid reward). Does a sequence form in this case? To answer this question, we need a cost-benefit analysis, i.e. what the agent gains by executing actions A_1 and A_2 in sequence and what it loses. Assume it decides to always execute A_2 after A_1. If the next state is S', then it loses nothing, because action A_2 is the best action in state S'. But, if the next state is S'', by taking A_2 instead of the best action, A'_2, the agent loses some of the future rewards. The amount of these reward losses is equal to the difference between the value of action A_2, $Q(S'', A_2)$, and the value of the best action, $V(S'')$, which will be $Q(S'', A_2) - V(S'')$, that we denote by $A(S'', A_2)$. $A(S'', A_2)$ can be interpreted as the advantage of taking action A_2 in state S'' instead of the best action (Baird, 1993; Dayan & Balleine, 2002). In this example, because the agent enters state S'' after state S

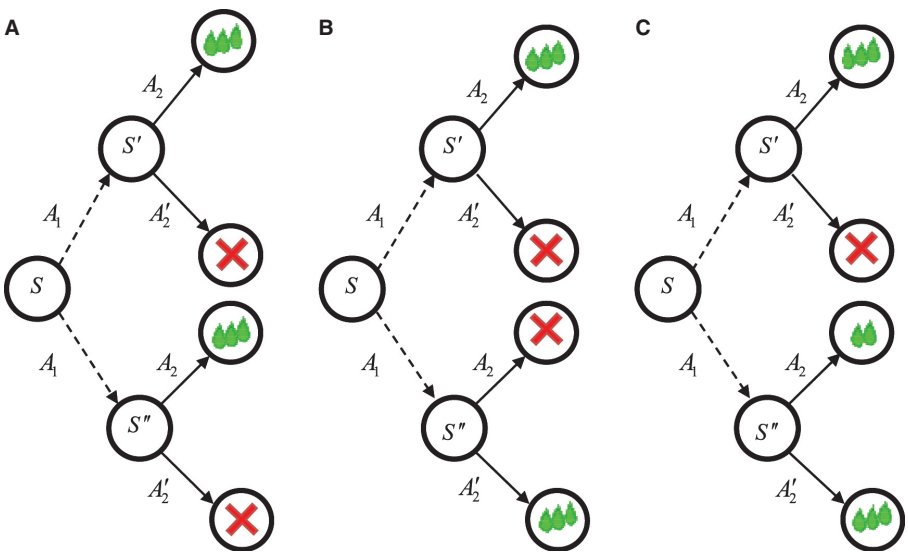

FIG. 3. Environmental constraints on sequence formation. (A) An example of an environment in which action sequences will form. Action A_1 leads to two different states with equal probability, in both of which action A_2 is the best action, and thus action sequence $\{A_1A_2\}$ forms. (B) An example of an environment in which action sequences do not form. Action A_1 leads to two different states with equal probability, in one of which action A_2 is the best action and, in another, action A'_2 is the best action. As a consequence, an action sequence $\{A_1A_2\}$ does not form. (C) An example of an environment in which the process of sequence formation is non-trivial. Action A_1 leads to two different states with equal probability, in one of which action A_2 is the best action, but in the other action A'_2 is the best action (although it is a little bit worse than the rival best action).

only half the time, the total cost of executing A_1 and A_2 in sequence will be 0.5 $A(S'', A_2)$.

Generalizing from the previous example, the cost of executing the macro action $\{aa'\}$ in state s is equal to:

$$C(s, a, a') = E[Q(s', a') - V(s')] = E[A(s', a')] \quad (5)$$

where expectation over the next state, s', given the previous action and the previous state, is:

$$C(s, a, a') = \sum_{s'} P_{sa}(s')[A(s', a')]. \quad (6)$$

Using the above equation, the term $C(s, a, a')$ can be computed based on the model-based approaches described. The agent computes $Q(s', a')$ and $V(s')$ using Eqn 4, and then $C(s, a, a')$ is calculated by Eqn 6. However, this means that at each decision point, deciding whether to execute an action sequence, Eqn 6 should be evaluated for all currently possible actions, and all possible subsequent actions. This will likely impose a heavy processing load on the decision-making process, and could considerably increase the latency of action selection. It turns out, however, that $C(s, a, a')$ can be estimated efficiently using samples of the temporal difference error signal (TD error signal).

The TD error signal experienced after taking action a_t in state s_t is defined as follows:

$$\delta_t = [r_t - d_t\bar{R} + V(s_{t+1})] - V(s_t) \quad (7)$$

Based on Eqn 3, the term $r_t - d_t\bar{R} + V(s_{t+1})$ is a sample of $Q(s_t, a_t)$. Thus, δ_t will be a sample of A (s_t, a_t), and hence $C(s, a, a')$ can be estimated using samples of the TD error signal. By taking action a_{t-1} in state s_{t-1}, and action a_t in state s_t, $C(s_{t-1}, a_{t-1}, a_t)$ will be updated as follows:

$$C_t(s_{t-1}, a_{t-1}, a_t) = (1 - \eta_C)C_{t-1}(s_{t-1}, a_{t-1}, a_t) + \eta_C\alpha_t\delta_t \quad (8)$$

where η_C is the learning rate, and α_t is a factor, which equals 1 when the environment is deterministic (see Appendix for more details). As mentioned above, extensive evidence from animal and human studies suggests that the TD error signal is coded by the phasic activities of mid-brain dopamine neurons (Schultz *et al.*, 1997; Schultz & Dickinson, 2000). Thus, besides being more efficient, utilizing the error signal for the purpose of sequence learning provides a neurally plausible way for computing the cost of sequence-based action selection, $C(s, a, a')$.

Up to now, we have only considered one side of the trade-off, which is the cost of sequence-based action selection. What are the benefits of sequence-based action selection? As discussed in the previous section, expression of a sequence of actions is faster than selecting actions one by one, based on the action evaluation process. This can be for several reasons; for example, identification of the current state by processing environmental stimuli can be time consuming, and the evaluation of actions using a model-based process is slower than having solely to select the next action from the sequence. Besides being faster, executing actions without going through the decision-making process makes it possible to perform a simultaneous task that requires decision-making resources. Here, we focus on the first advantage of sequence learning.

Assume that selecting the next action of the current sequence is τ time steps faster than selecting an action based on the action evaluation process. Saving τ time steps is equivalent to gaining $\bar{R}\tau$ more reward in the future (Niv *et al.*, 2007). This provides the other side of the trade-off: if the benefit of sequence-based action selection, $\bar{R}\tau$, exceeds its costs, $-C(s, a, a')$, then the macro action $\{aa'\}$ replaces action a in state s. Otherwise, if the macro action is already formed, it decomposes to its constitutent actions, and action a replaces the macro action $\{aa'\}$:

$$\text{If } -C(s, a, a') < \bar{R}\tau \text{ then:} \begin{cases} \text{replace action } a \text{ with the macro action} \\ \{aa'\} \text{ in state } s \\ \text{else} \\ \text{replace the action macro action } \{aa'\} \\ \text{with action a in state s} \end{cases} \quad (9)$$

After a macro action is added, it can be concatenated with other actions to form a longer macro action. For example, macro action $\{aa'\}$ can be concatenated with another action, say a'', and form the macro action $\{aa'a''\}$. It is important to recognize that, during execution of a macro action, primitive actions are not evaluated and thus the TD error signal is not computed, which means the cost of the action sequence, $C(s, a, a')$, is not updated after it is formed. This implies that a sequence should only form after the agent is certain about the estimated costs and benefits of the sequence; otherwise, the agent could stick to a sub-optimal sequence for a long period of time. This implies that action sequences should not form during early stages of instrumental learning because a high degree of certainty requires sufficient experience of the environment and hence more training time. In the current example, we did not model the agent's certainty about its estimations. Instead we assumed a large initial value for the cost of sequence formation, and chose a slow learning rate for that cost (η_C), something that ensures sequences form only after the environment is well learned.

A mixed model-based and sequence-based architecture

Assume that the agent is in state S in which several choices are available (Fig. 4), two of which are primitive actions (A_1 and A_2), and one of which is a macro action (A_3). In state S the agent uses model-based action evaluation, and selects one of the available actions for execution, which can be either a primitive action or a macro action. After completion of a primitive action the agent enters a new state in which it again uses a model-based evaluation for selecting subsequent actions. However, if the selected action is the macro action, A_3, its execution is composed of taking a sequence of primitive actions (M_1... M_4). Nevertheless, upon completion of the macro action, the agent identifies its new state (S_3), and uses model-based action evaluation again for selection of the next action.

The above scenario involves mixed model-based and sequence-based action control. At the choice points, actions are selected based

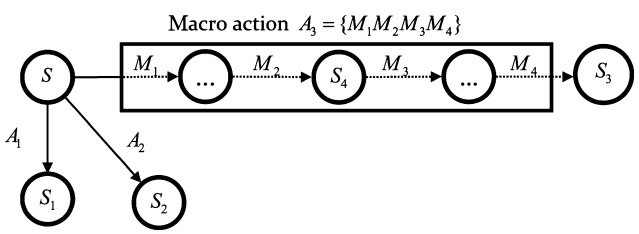

FIG. 4. Mixed model-based and sequence-based decision-making. At state S_0, three actions are available: A_1, A_2, A_3, where A_1 and A_2 are primitive actions, and A_3 is a macro action composed of primitive actions M_1… M_4. If at state S, the macro action is selected for execution, the action control transfers to the sequence-based controller, and actions M_1… M_4 become executed. After the termination of the macro action control returns back to model-based decision-making at state S_3.

on the model-based evaluations. However, during the execution of a sequence of actions, they are selected based on their sequential order, without going through the evaluation process. As discussed in the previous section, this sequence-based action selection can lead to higher average reward rates, in comparison to a pure model-based decision-making system, which is the benefit of sequence-based action selection. However, it can lead to a maladaptive behavior if the environment changes after action sequences have formed. For example, assume that after the macro action {*aa*'} has formed, the environment changes so that the execution of action a' after action a no longer satisfies the cost–benefit analysis presented in the previous section – say the change causes the value of the state to which action a' leads to decrease significantly – as a consequence, taking action a' after action a will no longer be the best choice. If action control is sequence-based, it is the previous action that determines the next action and not the consequences of the action. Hence, the agent will continue to take action a' even though it is not the most appropriate action.

Ultimately, in this situation, we expect the macro action to decompose to its components so that the agent can consider other alternative actions other than action a'. However, this does not happen instantly after the environment has changed and, thus at least for a while, the agent will continue behaving maladaptively. As mentioned in the previous section, after the macro action has formed, its cost, $C(s, a, a')$, is not updated, because the system is working on the open-loop mode and the TD error signal is not computed to update $C(s, a, a')$. As a consequence, the cost side of the sequence formation trade-off is relatively insensitive to environmental changes. The other side of the trade-off, $\bar{R}\tau$, however, is sensitive to environmental changes: if the environment changes so that executing the macro action leads to a decrease in the average reward the agent experiences, $\bar{R}$, then this circumstance motivates decomposition of the macro. In fact, this model predicts that, if the environmental changes do not alter the average reward, then the agent will continue to take action a' after a, even if the change introduces better alternatives other than taking action a'. Nevertheless, if the change causes a decrease in the average reward, then the macro action will decompose, and the responses will adapt to the new situation. However, this cannot happen instantly after the change because it takes several trials before the average reward adapts to the new condition.

The above feature of the model implies different sensitivity of sequence-based responses of the model after an environmental change compared with the situation where responses are under model-based control. As an example, in Fig. 4 assume that the action A_1 is the best action, i.e. it has the highest Q-value among actions A_1, A_2 and A_3, and so the agent takes action A_1 more frequently than the others. Now, assume that the environment changes, and the value of the state that action A_1 leads to (state S_1) dramatically decreases. The next time that the agent is making a decision in state S, it evaluates the consequences of action A_1 using Eqn 3, and finds out that A_1 is no longer the best action, and adapts its behavior instantly to the new conditions. Evidence for the effect of this type of environmental change on the behavior comes from an experiment (Ostlund *et al.*, 2009) in which rats were trained on two action sequences for two outcomes, i.e. R1 → R2 → O1 and R2 → R1 → O2. After this training, either O1 or O2 was devalued, and performance of the two sequences (macro actions {R1R2} and {R2R1}) were assessed in extinction. Results show that the performance of the sequence that leads to the devalued outcome decreases, which implies that performance of a macro action (e.g. {R1 R2}) is immediately sensitive to the value of the states to which it leads (O1). It is worth mentioning that although the values of individual actions in the sequence are not updated, the total reward gained by executing the action sequence is learned by the model-based system.

Compare the above situation with one in which an environmental change causes a decrease in the value of one of the states visited during a sequence, for example state S_4. Here, an action other than M_2 would now be more optimal but, because action selection is under sequence-based control, the choice of M_2 is derived by its previous action, M_1, rather than the value of the state that M_2 leads to, i.e. S_4. After several trials, because taking action M_2 does not lead to reward, the average reward that the agent experiences decreases, and the sequence should then decompose into its elements. At this point, the control of actions will return to the model-based system and choices adapt to the new environmental conditions. In the next section we show that insensitivity to reinforcer devaluation and contingency degradation is due to this type of environmental change.

Simulation of the model: sequence learning and habit formation

Having described the model, we are now in a position to establish whether it can provide an accurate account of: (i) sequence learning, such as that observed in SRTT; and (ii) instrumental conditioning, particularly the shift in sensitivity of instrumental actions to reinforcer devaluation and contingency degradation during the course of overtraining (see Appendix for implementation details).

Sequential and random trials of sequence learning

As already noted, when a sequence of stimuli is predictable, such as in the 'sequential trials' of the SRTT, along with the progress of learning as a result of sequence-based action selection, reaction times decline. In contrast, when the sequence of stimuli is random, as it is in the 'random trials' condition of SRTT, reaction times do not decrease substantially during the course of learning. Here, we simulated the model described previously in a task similar to SRTT. After each correct button press the model receives one unit of reward. In the 'sequential trials' condition, after each correct button press the next stimulus in a fixed sequence is presented, otherwise the sequence restarts and the first stimulus is presented. Here, we assume that the order of stimuli is S_0 to S_3, where the correct button press in S_0 is B_0, B_1 in S_1, etc. In the 'random trials' condition, the next stimulus is selected randomly. It is assumed that it takes 400 ms to make a decision using the model-based method, and 100 ms to elicit a response under sequence-based action control.

The temporal dynamics of sequence formation is depicted in Fig. 5. After several learning trials, the agent learns the model of the environment (the rewards in states, delays in states and the consequences of each action) and, as a consequence, the probability of taking the correct actions increases, which implies that the agent gains more rewards and, thus, the average reward that the agent receives, $\bar{R}$, increases. This increase in the average reward implies that a significant number of the rewards that could have been gained in the future are being lost due to time taken for model-based action selection, and this favors the transition of action control to the sequence-based method, which is faster. At the same time, the cost of sequence-based action selection, $C(S_0, B_0, B_1)$ decreases (Fig. 5A), which means that the agent has learned B_1 is always the action that should be taken after B_0. Eventually, the benefit of sequence-based action selection becomes larger than its cost and, at that stage, the macro action $\{B_0B_1\}$ replaces the B_0 action (Fig. 5B). Later, actions B_2 and B_3 form the macro action $\{B_2B_3\}$ and, finally, the two previously formed macro actions concatenate, and the macro action $\{B_0B_1B_2B_3\}$ is formed. In addition, as shown in Fig. 5A, after a macro

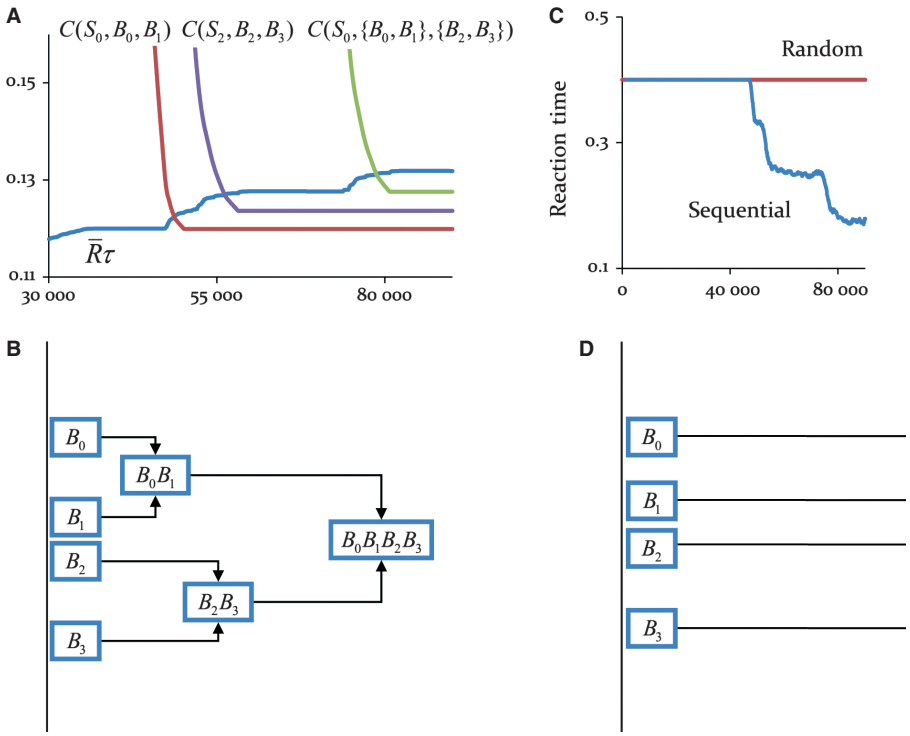

FIG. 5. The dynamics of sequence learning in sequential and random trials of SRTT. (A, B) As the learning progresses, the average reward that the agent gains increases, indicative of a high cost of waiting for model-based action selection. At the same time, the cost of sequence-based action selection decreases (top panel), which means that the agent has discovered the correct action sequences. Whenever the cost becomes less than the benefit, a new action sequence forms (bottom panel). The abscissa axis shows the number of action selections. (C) Reaction times decrease in sequential trials as a result of sequence formation but they remain constant in the random trials of SRTT because, (D) no action sequence forms. Data reported are means over 10 runs.

action is formed the average reward that the agent gains increases due to faster decision-making.

Figure 5C shows the reaction times. As the figure shows, by forming new action sequences, reaction times decrease up to the point that only selection of the action sequence is based on model-based action control, and all subsequent button presses during the sequence are based on sequence-based action control. Figure 5C also shows the reaction times in the case of 'random trials', which remain constant largely because the sequence of stimuli is not predictable and the cost of sequence-based action selection remains high so that no action sequence forms (Fig. 5D).

Instrumental conditioning

In this section, we aimed to validate the model in instrumental conditioning paradigms. Figure 6A depicts a formal representation of a simple instrumental conditioning task. The task starts in state S_0 and the agent has two options: the press lever (PL) action; and enter magazine (EM) action. By taking action PL, and then action EM, the agent enters state S_0 in which it receives one unit of reward ($r = 1$). All other actions, for example taking action EM before action PL, leads to no reward. Entering state S_1 is cued for example by a 'click' produced by the pellet dispenser, or 'buzz' of the pump if sucrose solution is the reward.

After several learning trials, the agent learns the model of the environment; the value of the PL action exceeds the value of action EM, and the probability of taking action PL increases. As the probability of taking PL increases, the agent gains more rewards and, hence, the average reward $\bar{R}$ increases. Simultaneously, the cost of sequence-based action selection decreases, which means the agent has learned that PL is always the action that should be taken after the EM action

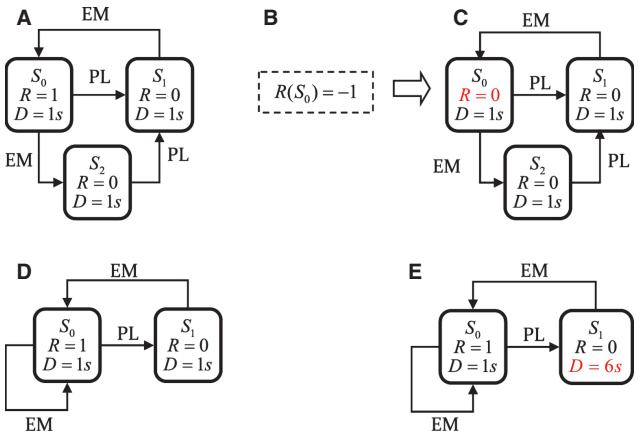

FIG. 6. Formal representation of instrumental conditioning tasks. (A) Instrumental learning: by taking the press lever (PL) action, and then enter magazine (EM) action, the agent earns a reward of magnitude one. By taking EM action in state S_2 and PL action in state S_1, the agent remains in the same state (not shown in the figure). (B, C) Reinforcer devaluation: the agent learns that the reward at state S_0 is devalued, and then is tested in extinction in which no reward is delivered. (D) Non-contingent training: unlike as in (A), reward is not contingent on the PL action, and the agent can gain the reward only by entering the magazine. (E) Omission training: taking the PL action causes a delay in the reward delivery, and the agent should wait 6 s before it can gain the reward by entering the magazine.

and, as a consequence, the macro action {EM, PL} replaces the EM action (Fig. 7A). From that point, the action PL is always taken after EM.

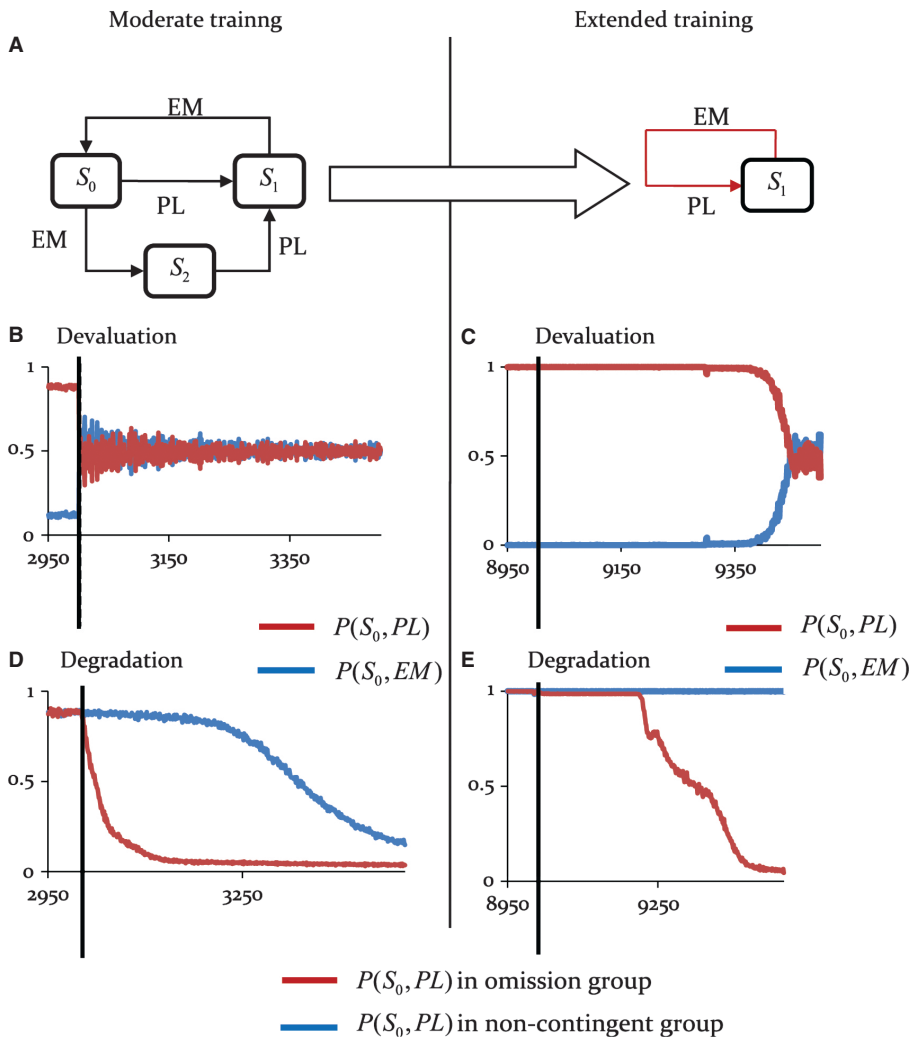

FIG. 7. Sensitivity of the model to reinforcer devaluation and contingency manipulations before and after sequence formation. (A) In the moderate training condition, actions are selected based on the model-based evaluation (left panel) but, after extended training, the selection of the press lever (PL) action is potentiated by its previous action [here enter magazine (EM); right panel]. (B) After the devaluation phase (shown by the solid-line), the probability of pressing the lever decreases instantly if the model is moderately trained. The abscissa axis shows the number of action selections. (C) After the devaluation phase behavior does not adapt until the action sequence decomposes and control returns to the model-based method. (D) In a moderately trained model the probability of selecting action PL starts to decrease in the contingency degradation condition, although the rate of decrease is greater in the case of omission training. (E) When training is extensive, behavior does not adjust and the non-contingent and omission groups perform at the same rate until the sequence decomposes. Data reported are means over 3000 runs.

The schematic representation in Fig. 6A corresponds to a continuous reinforcement schedule, in which each lever press is followed by a reinforcer delivery (e.g. Adams, 1982). However, in most experimental settings, animals are required to execute a number of lever presses (in the case of ratio schedules), or press the lever after an amount of time has passed since the previous reward delivery (in the case of interval schedules) in order to obtain reward. One approach to analysing these kinds of experiments using the paradigm illustrated in Fig. 6A is through application of the 'response unit hypothesis' (Skinner, 1938; Mowrer & Jones, 1945), according to which the total set of lever presses required for reward delivery is considered as a single unit of response. For example, in the case of fixed ratio schedules, if 10 lever presses are required to produce reinforcement, the whole 10 lever presses are considered as a single unit, corresponding to the action PL in Fig. 6A. In ratio schedules this hypothesis is supported by the observation that, early in learning, the animal frequently takes the EM action, which, with the progress of learning, tends to occur only after the last lever press (Denny *et al.*, 1957; Overmann & Denny, 1974; Platt & Day, 1979). Following this training, animals are given an extinction test, in which rewards are not delivered (but in which state S_1 is cued). During the course of extinction, the response unit mainly preserves its form and only the last response is likely to be followed by the EM action (Denny *et al.*, 1957). Further, the average number of executed response units in extinction is independent of the number of lever presses required for reinforcer delivery, indicating that the whole response unit is being extinguished, and not individual lever presses (Denny *et al.*, 1957; Overmann & Denny, 1974; Platt & Day, 1979). In the case of interval schedules, because reinforcement of a lever press depends on the time that has passed since the previous reward delivery, the 'response unit hypothesis' has to be generalized to temporal response unit structures in which the animal continues to lever press for a certain amount of time (instead of for a certain number of times), in the form of 'bouts' of lever pressing (Shull & Grimes, 2003), or nose poking (Shull *et al.*, 2002).

In fact, in the previous section, in the course of analysing SRTT, we applied the 'response unit hypothesis' by treating the action of pressing the button as a single response unit, which of course can be broken into

smaller units. Similarly, in the case of maze learning, the action of reaching the choice point from starting point can be thought as a sequence of steps. Is the formation of such response units (e.g. PL action composed of homogenous set of responses) through the action sequence formation method proposed in the previous section, or do they form in a level different from that in which macro action {*EM*, *PL*} forms? We leave the answer to this question for future works.

In the next two sections, we investigated the behavior of the model, based on Fig. 6A, when an environmental change occurs both before and after sequence formation.

Reinforcer devaluation before vs. after action sequence formation

As described above, in reinforcer devaluation studies, the value of the outcome of an action is reduced offline through some treatment (such as specific satiety or taste aversion learning) and the performance of the action subsequently assessed in extinction. There is a considerable literature demonstrating that, with moderate training, instrumental performance is sensitive to this change on value, whereas after more extended training it is not (cf. Fig. 1A).

To assess the accuracy of the model it was simulated using the procedure depicted in Fig. 6. The procedure has three steps. The first step (Fig. 6A) is the instrumental learning phase, described in the previous section. The next step (Fig. 6B) models the devaluation phase in which the model learns that the reward obtained in state S_0 is devalued ($r = -1$). The third phase is the test conducted under extinction conditions, i.e. reward is not delivered in state S_0 ($r = 0$). The critical question here is whether the model chooses action PL in state S_0. As noted previously, experimental evidence shows that after moderate training, the agent chooses action PL, whereas after extended training it does not. Figure 7B shows the probability of taking action PL after moderate training (in this case after 3000 action selections). As the figure shows, because action selection is under model-based control, when the reward in state S_0 is devalued, the value of taking PL action in state S_0 is immediately affected and, as such, the probability of taking action PL decreases.

The same is not true of overtrained actions. Figure 7C shows the sensitivity of responses to devaluation after extended training (9000 action selections). At this point the action sequence {*EM*, *PL*} has been formed and, as the figure shows, unlike moderate training the agent continues taking action PL after reinforcer devaluation. This is because action selection is under sequence-based action control, and selecting action PL is driven by the previous action (EM) rather than the value of the upcoming state. After several learning trials, because the experiment is conducted in extinction and no reward is received, the average reward decreases, which means deciding faster is not beneficial, and causes the macro action {*EM*, *PL*} to decompose to action EM and action PL. At this point, behavioral control should return to the model-based system, and the probability of taking action PL should adjust to the new conditions induced by devaluation. It is worth noting that the prediction that extensive devaluation testing in extinction will cause habitual behavior to revert to goal-directed control has not been assessed experimentally largely because extinction usually produces such a profound suppression of performance that any change in behavioral control would be difficult to detect.

Contingency degradation before vs. after action sequence formation

In the first section we began by pointing out that habits are not just insensitive to reinforcer devaluation but also to the effects of degrading the instrumental contingency. A good example of this is the failure of habits to adjust to the imposition of an omission schedule, as shown in Fig. 1B. Having learned that lever pressing delivers food, the omission schedule reverses that relationship such that food becomes freely available without needing to lever press to receive it. However, in the experimental group, lever pressing delays free food delivery; hence, the rats now have to learn to stop lever pressing to get the reward. The ability to stop responding in the omission group is compared with rats given exposure to a zero contingency between lever pressing and reward delivery (the non-contingent control). As shown previously (e.g. Dickinson *et al.*, 1998; Fig. 1B) in this situation, rats who are given moderate instrumental training are able to withhold their lever press action to get food; the omission group responds less than the control group when the omission schedule is introduced. When lever pressing has been overtrained, however, the rats are insensitive to omission and cannot withhold their lever press responses compared with the control group.

For the simulation of non-contingent reward delivery in the control condition, the model underwent the instrumental conditioning procedure described in the previous section (Fig. 6A), and then the model was simulated in the task depicted in Fig. 6D. The agent can obtain reward by taking action EM without performing action PL, and hence reward delivery is no longer contingent upon action PL. For the simulation of the omission schedule, after instrumental training (Fig. 6A), the agent was exposed to the schedule depicted in Fig. 6E. The difference between this schedule and the non-contingent schedule is that, after pressing the lever, the model must wait 6 s before obtaining the reward by taking action EM, which models the fact that rewards are delayed if the animal chooses action PL under the omission schedule. The behavior of the model after moderate training is depicted in Fig. 7D. As the figure shows, after the introduction of the degradation procedure, the probability of taking action PL starts to decrease. The rate of decrease is faster in the case of the omission schedule, in comparison to the non-contingent schedule. This is because the final value of action PL in the omission condition is lower than the final value of the non-contingent condition and, therefore, the values adjust faster in the omission case.

Figure 7E shows the effect of omission and degradation after more extended training when action selection is under sequence-based control. As the figure shows, in both the control and omission conditions, the probability of selecting the action fails to adjust to the new conditions and the agent continues selecting action PL. However, in the omission condition, because rewards are delayed as a result of pressing the lever, the average reward starts to decrease and, after sufficient omission training, the action sequence decomposes, and behavior starts to adapt to the new conditions. In the case of non-contingent reward delivery, because the average reward remains constant, the model predicts that the agent will continue pressing the lever, even after extended exposure to the non-contingent schedule. Again, this prediction has not been assessed in the literature largely because exposure to non-contingent reward tends to have a generally suppressive effect on performance, as competition from EM responding increases and action PL begins to extinguish. As is the case after reinforcer devaluation testing, it can be challenging to assess changes in performance when responding has reached a behavioral floor.

Although somewhat idealized relative to the effects observed in real animals, it should be clear that, in contrast to simple RL, sequence learning and habitual actions are both accurately modeled by this mixed model-based and sequence-based architecture. The implications of this model for the behavior of real animals and for theories of goal-directed and habitual action control are described below.

The implications of the model

The development of the mixed model-based and sequence-based architecture we describe and simulate here was motivated by the fact that simple RL fails accurately to model the performance of habitual instrumental actions. One reason for this failing, we believe, lies in the way model-free RL treats the TD error signal during the acquisition of habitual actions. When applied to a specific action, the TD error actually ensures that overtrained, habitual actions will be more sensitive to degradation of the instrumental contingency than goal-directed actions. This is the opposite pattern of results to that observed in real animals (see Fig. 1). We also argued that the problem lies in the way that habitual actions are described, i.e. they are typically regarded as essentially identical to goal-directed actions in their representational structure; a response (whether a button push, handle turn, lever press, etc.) can be nominally either goal-directed or habitual in the usual case. We propose, to the contrary, that as goal-directed actions become habitual they grow more complex in the sense that they become more thoroughly integrated, or chunked, with other motor movements such that, when selected, these movements tend to run off in a sequence. Under this scheme, rather than establishing the value of individual actions, the TD error provides feedback regarding the cost of this chunk or sequence of actions, an approach that we show successfully models the insensitivity of overtrained actions to both reinforcer devaluation and degradation of the instrumental contingency.

Implications and predictions: habitual responses

The solution we propose to the issue that habits pose to RL is, therefore, to view them as a form of action sequence under the control of a model-based RL process. There are two important implications, and sets of predictions, generated by this kind of approach. The first is for the nature of habits themselves. Viewed as sequences of movements, as opposed to action primitives, habitual responses should be anticipated to have qualities that differentiate them from goal-directed actions, many of which will be similar to those of skilled movements (Newell *et al.*, 2001; Buitrago *et al.*, 2004a). First, we should anticipate observing a reduction in the variation in the movements around the action, for example if, say, lever pressing is chunked with other movements, the frequency of those movements should increase whereas the incidence of other extraneous movements should decline. Second, the inter-response interval between any candidate set of sequential movements should decline with practice. More critically: (i) the incidence of these factors should be predicted to correlate with the acquisition of habits by other measures, such as reduced sensitivity to reinforcer devaluation and contingency degradation; and (ii) manipulations known to affect habits, such as lesions or inactivation of DLS (Yin *et al.*, 2008) or of the dopaminergic input to this region from the substantia nigra pars compacta (Faure *et al.*, 2005), should be expected to affect these measures and alter actions to make their performance more similar to that observed when demonstrably goal-directed.

In fact, there is already considerable evidence for some of these effects in the literature examining the acquisition of skilled movements (Wolpert *et al.*, 2001; Wolpert & Flanagan, 2010), although it is well beyond the scope of the current paper to review that literature here. Of greater direct relevance is evidence from studies examining the performance of animals in various mazes, particularly those that have been heavily utilized to study the acquisition of habitual performance, such as the T-maze. For example, Jog *et al.* (1999) overtrained rats in a T-maze and found that, as component responses performed between start and end points in the maze declined in latency, neural activity in the DLS specific to those points also gradually declined to the extent that task-related activity was limited to the first and last responses in the maze, i.e. the points at which any response sequence or chunk should have been initiated and terminated. Similar effects have been observed by Barnes *et al.* (2005) and, indeed, using response reversals, they were also able to observe a collapse of the sequence, i.e. both inter-maze responses and neural activity initially associated with those responses reemerged along with task-irrelevant movements. The inter-maze responses again declined with continued training post-reversal. Hence, changes in both striatal unit activity and incidental behavioral responses tracked the development of the sequence, as has been observed recently using a homogeneous sequence of lever press responses, described above (Jin & Costa, 2010), and in a head movement habit (Tang *et al.*, 2007). Finally, Kubota *et al.* (2009) reported observing electrophysiological effects associated with both the overall response sequence and the component response primitives as these were rapidly remapped onto new stimuli presented during the course of T-maze performance, suggesting that the mice (in this case) maintained separate representations of the sequence and component movements. Interestingly, Redish and colleagues (Schmitzer-Torbert & Redish, 2004; Johnson *et al.*, 2007) reported neural signals in the striatum associated with sequence learning in a novel navigation task composed of a series of T-mazes reordered across days. Striatal activity differed across different sequences of turns in the maze, but was also highly correlated across repeated sequences suggesting, again, the co-occurrence of movement-specific and sequence-specific activity in the striatum.

The role of sensorimotor striatum in action sequences is not without controversy (Bailey & Mair, 2006; Turner & Desmurget, 2010). For example, inactivation of the globus pallidus internus, the principal output of the sensorimotor striatum, does not increase the reaction time in the 'sequential trials' of SRTT (Desmurget & Turner, 2010), which suggests performance of action sequences is not dependent on the sensorimotor striatum. However, although execution, as opposed to the encoding, of sequence-based actions may not be dependent on sensorimotor striatum, it can be argued that transfer of control to sequence-based action control can be independent of its output in the pallidum. Based on this assumption, disrupting the output should not necessarily be accompanied by an increase in the reaction time, because it does not cause transfer of control to model-based action control, but should result in an increase in the number of errors because it cannot exert control over the motor control circuitry, consistent with the experimental observation (Desmurget & Turner, 2010).

Implications and predictions: action sequences

The second implication of the current analysis is for the way in which the processes controlling goal-directed and habitual actions should be thought to interact. A number of investigators have noted the apparently competitive nature of these forms of action control, and some have suggested that these processes may compete for access to the motor system. Using this general approach, previous computational accounts have successfully explained the effect of reinforcer devaluation on instrumental responses in different stages of learning (Daw *et al.*, 2005; Keramati *et al.*, 2011), the effect of habit formation on reaction times (Keramati *et al.*, 2011), and the effect of the complexity of state identification on the behavioral control (Shah & Barto, 2009). All these approaches shares a common 'flat' architecture in which the goal-directed and the habitual systems work in parallel at

the same level, and utilize a third mechanism, called an arbitration mechanism, to decide whether the next action will be controlled by the goal-directed or the habit process. They differ from the hierarchical architecture used here in which the goal-directed system stands at a higher level, with the role of the habit process limited to efficiently implementing decisions made by the goal-directed system in the form of macro actions. This structural difference itself raises some important behavioral predictions. For example, in the hierarchical structure, the goal-directed system treats macro actions as integrated units, and thus the action evaluation process is blind to the change in the value of states visited during execution of the macro action. Thus, in Fig. 4, goal-directed action evaluation in state S depends only on the value of states S_1, S_2, S_3 and the total reward obtained through executing the macro action. As a consequence, changing the value of state S_4 has no immediate impact on the decisions made at state S. In contrast, in a 'flat' architecture, goal-directed action selection is conducted by searching all the consequences of possible actions and, thus, a change in the value of state S_4 should immediately affect action selection in state S. Another prediction of a sequence-based conception of habits is that, if an agent starts decision-making in a state in which an action sequence has been learned (state S), it will immediately show habit-like behavior, such as insensitivity to outcome devaluation. In contrast, if it starts decision-making somewhere in the middle of a sequence (e.g. in the test phase the task starts in an intermediary state such as S_4), it will not show habit-like behavior. This is because in the current conception of the model, an action sequence can be launched only in the state in which is has been learned.

The problem of mixed closed-loop and open-loop decision-making has been previously addressed and discussed in the literature, but the solutions proposed differ from that suggested here. In the model proposed here, the inputs for the mixed architecture come from fast action control in the open-loop mode, whereas in previous work they have come from a cost associated with sensing the environment (Hansen *et al.*, 1996), or the complexity of modeling the environment (Kolter *et al.*, 2010). From a control point of view, in most situations (especially stochastic environments), open-loop action control is considered to be inappropriate. As such, in hierarchical approaches to RL, the idea of macro actions is usually generalized to closed-loop action control (Barto & Mahadevan, 2003). Behavioral and neural signatures of this generalized notion have been found in previous studies (Haruno & Kawato, 2006; Botvinick, 2008; Botvinick *et al.*, 2009; Badre & Frank, 2012; Ribas-Fernandes *et al.*, 2011). Here, instead of using this generalized hierarchical RL, we utilized a hierarchical approach with a mixed open-loop and closed-loop architecture that allows us to incorporate the role of action sequences in habit-learning. Further, to examine the potential neural substrates of this architecture, we developed a method based on the TD error for learning macro actions, though various alternative methods have also been proposed for this purpose (Korf, 1985; Iba, 1989; Randløv, 1998; Mcgovern, 2002).

With regard specifically to sequences, the effect of overtraining on reaction times has been addressed in instrumental conditioning models (Keramati *et al.*, 2011), which often interprets them as the result of transition to habitual control, which, because it is fast, results in a reduction in reaction times. However, that model predicts a decrease in reaction times in both 'random' and 'sequential trials' of SRTT. This is because, on that approach, a habit forms whenever the values of the available choices are significantly different, irrespective of whether the sequence of states is predictable. As such, because in both 'sequential' and 'random trials' of SRTT the values of correct and incorrect responses are different, the model predicts that habits will form and reaction times decrease in both cases, which is not consistent with the evidence. In the sequence-learning literature, the issue of learning sequences of responses and stimuli using TD error has been addressed (Berns & Sejnowski, 1998; Bapi & Doya, 2001; Nakahara *et al.*, 2001; Bissmarck *et al.*, 2008). However, because these models are generally developed for the purpose of visuo-motor sequence learning, it is not straightforward to apply them to instrumental conditioning tasks. Likewise, the effect of the predictability of stimuli (i.e. random vs. sequential trials in the SRTT) on reaction times is not directly addressed in these models, which makes it hard to compare them in SRTT.

One important restriction of the proposed model relates to the fact that it may seem implausible to assume that, after an action sequence has been formed, the individual actions are always executed together. To address this issue it can, for example, be assumed that, occasionally, the model-based controller interrupts the execution of actions during the performance of an action sequence in order to facilitate learning new action sequences. Along the same lines, it can be assumed that action sequences do not replace primitive actions (as proposed in this paper), but are added as new actions to the list of available actions. Finally, investigating why some kinds of reinforcement schedules lead to habits (action sequences) whilst others do not, is an interesting issue and each of these will be addressed in future work.

Abbreviations

DLS, dorsolateral striatum; DMS, dorsomedial striatum; EM, enter magazine; PL, press level; RL, reinforcement learning; SRTT, serial reaction time task; TD, temporal difference.

References

Adams, C.D. (1982) Variations in the sensitivity of instrumental responding to reinforcer devaluation. *Q. J. Exp. Psychol. B*, **34B**, 77–98.

Alexander, G.E. & Crutcher, M.D. (1990) Functional architecture of basal ganglia circuits: neural substrates of parallel processing. *Trends Neurosci.*, **13**, 266–271.

Astrom, K.J. & Murray, R.M. (2008) *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton University Press, Princeton, NJ.

Badre, D. & Frank, M.J. (2012) Mechanisms of hierarchical reinforcement learning in cortico-striatal circuits 2: evidence from fMRI. *Cereb. Cortex*, **22**, 527–536.

Bailey, K.R. & Mair, R.G. (2006) The role of striatum in initiation and execution of learned action sequences in rats. *J. Neurosci.*, **26**, 1016–1025.

Bailey, K.R. & Mair, R.G. (2007) Effects of frontal cortex lesions on action sequence learning in the rat. *Eur. J. Neurosci.*, **25**, 2905–2915.

Baird, L.C. (1993) *Advantage Updating (No. WL–TR-93-1146)*. Wright Laboratory, Wright-Patterson Air Force Base Ohio.

Balleine, B.W. & O'Doherty, J.P. (2010) Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, **35**, 48–69.

Balleine, B.W., Delgado, M.R. & Hikosaka, O. (2007) The role of the dorsal striatum in reward and decision-making. *J. Neurosci.*, **27**, 8161–8165.

Balleine, B.W., Liljeholm, M. & Ostlund, S.B. (2009) The integrative function of the basal ganglia in instrumental conditioning. *Behav. Brain Res.*, **199**, 43–52.

Bapi, R.S. & Doya, K. (2001) Multiple forward model architecture for sequence processing. In Sun, R. & Giles, C.L. (Eds), *Sequence learning: paradigms, algorithms, and applications*. Springer Verlag, NY, pp. 309–320.

Barnes, T.D., Kubota, Y., Hu, D., Jin, D.Z. & Graybiel, A.M. (2005) Activity of striatal neurons reflects dynamic encoding and recoding of procedural memories. *Nature*, **437**, 1158–1161.

Barto, A.G. & Mahadevan, S. (2003) Recent advances in hierarchical reinforcement learning. *Discrete Event Dyn. S.*, **13**, 41–77.

Berns, G.S. & Sejnowski, T.J. (1998) A computational model of how the basal ganglia produce sequences. *J. Cogn. Neurosci.*, **10**, 108–121.

Bissmarck, F., Nakahara, H., Doya, K. & Hikosaka, O. (2008) Combining modalities with different latencies for optimal motor control. *J. Cogn. Neurosci.*, **20**, 1966–1979.
Botvinick, M.M. (2008) Hierarchical models of behavior and prefrontal function. *Trends Cogn. Sci.*, **12**, 201–208.
Botvinick, M.M., Niv, Y. & Barto, A.G. (2009) Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition*, **113**, 262–280.
Buitrago, M.M., Ringer, T., Schulz, J.B., Dichgans, J. & Luft, A.R. (2004a) Characterization of motor skill and instrumental learning time scales in a skilled reaching task in rat. *Behav. Brain Res.*, **155**, 249–256.
Buitrago, M.M., Schulz, J.B., Dichgans, J. & Luft, A.R. (2004b) Short and long-term motor skill learning in an accelerated rotarod training paradigm. *Neurobiol. Learn. Mem.*, **81**, 211–216.
Chang, Q. & Gold, P.E. (2004) Inactivation of dorsolateral striatum impairs acquisition of response learning in cue-deficient, but not cue-available, conditions. *Behav. Neurosci.*, **118**, 383–388.
Costa, R.M., Cohen, D. & Nicolelis, M.A.L. (2004) Differential corticostriatal plasticity during fast and slow motor skill learning in mice. *Curr. Biol.*, **14**, 1124–1134.
Costa, R.M., Lin, S.-C., Sotnikova, T.D., Cyr, M., Gainetdinov, R.R., Caron, M.G. & Nicolelis, M.A.L. (2006) Rapid alterations in corticostriatal ensemble coordination during acute dopamine-dependent motor dysfunction. *Neuron*, **52**, 359–369.
Daw, N.D. & Touretzky, D.S. (2000) Behavioral considerations suggest an average reward TD model of the dopamine system. *Neurocomputing*, **32–33**, 679–684.
Daw, N.D. & Touretzky, D.S. (2002) Long-term reward prediction in TD models of the dopamine system. *Neural Comput.*, **14**, 2567–2583.
Daw, N.D., Niv, Y. & Dayan, P. (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.*, **8**, 1704–1711.
Dayan, P. & Balleine, B.W. (2002) Reward, motivation, and reinforcement learning. *Neuron*, **36**, 285–298.
Dayan, P. & Kakade, S. (2001) Explaining away in weight space. In Leen, T.K., Dietterich, T.G. & Tresp, V. (Eds), *Advances in Neural Information Processing Systems 13*. MIT Press, Denver, CO, USA, pp. 451–457.
Dayan, P., Kakade, S. & Montague, P.R. (2000) Learning and selective attention. *Nat. Neurosci.*, **3**, 1218–1223.
Denny, M.R., Wells, R.H. & Maatsch, J.L. (1957) Resistance to extinction as a function of the discrimination habit established during fixed-ratio reinforcement. *J. Exp. Psychol.*, **54**, 451–456.
Desmurget, M. & Turner, R.S. (2010) Motor sequences and the basal ganglia: kinematics, not habits. *J. Neurosci.*, **30**, 7685–7690.
Devan, B.D. & White, N.M. (1999) Parallel information processing in the dorsal striatum: relation to hippocampal function. *J. Neurosci.*, **19**, 2789–2798.
Devan, B.D., McDonald, R.J. & White, N.M. (1999) Effects of medial and lateral caudate-putamen lesions on place- and cue-guided behaviors in the water maze: relation to thigmotaxis. *Behav. Brain Res.*, **100**, 5–14.
Dickinson, A. (1994) Instrumental conditioning. In Mackintosh, N.J. (Ed.) *Animal Cognition and Learning*. Academic Press, London, pp. 4–79.
Dickinson, A., Squire, S., Varga, Z. & Smith, J.W. (1998) Omission learning after instrumental pretraining. *Q. J. Exp. Psychol. B*, **51**, 271–286.
Endress, A.D. & Wood, J. (2011) From movements to actions: two mechanisms for learning action sequences. *Cogn. Psychol.*, **63**, 141–171.
Faure, A., Haberland, U., Condé, F. & El Massioui, N. (2005) Lesion to the nigrostriatal dopamine system disrupts stimulus-response habit formation. *J. Neurosci.*, **25**, 2771–2780.
Featherstone, R.E. & McDonald, R.J. (2004) Dorsal striatum and stimulus-response learning: lesions of the dorsolateral, but not dorsomedial, striatum impair acquisition of a stimulus-response-based instrumental discrimination task, while sparing conditioned place preference learning. *Neuroscience*, **124**, 23–31.
Featherstone, R.E. & McDonald, R.J. (2005) Lesions of the dorsolateral striatum impair the acquisition of a simplified stimulus-response dependent conditional discrimination task. *Neuroscience*, **136**, 387–395.
Frank, M.J. & O'Reilly, R.C. (2006) A mechanistic account of striatal dopamine function in human cognition: psychopharmacological studies with cabergoline and haloperidol. *Behav. Neurosci.*, **120**, 497–517.
Graybiel, A.M. (1998) The basal ganglia and chunking of action repertoires. *Neurobiol. Learn. Mem.*, **70**, 119–136.
Hansen, E.A., Barto, A.G. & Zilberstein, S. (1996) Reinforcement learning for mixed open-loop and closed loop control. In Mozer, M., Jordan, M.I. & Petsche, T. (Eds). *Advances in Neural Information Processing Systems NIPS*, Vol. 9. The MIT Press, Denver, CO, USA, pp. 1026–1032.
Haruno, M. & Kawato, M. (2006) Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops: fMRI examination in stimulus-action-reward association learning. *Neural. Netw.*, **19**, 1242–1254.
Hernandez, P.J., Schiltz, C.A. & Kelley, A.E. (2006) Dynamic shifts in corticostriatal expression patterns of the immediate early genes Homer 1a and Zif268 during early and late phases of instrumental training. *Learn. Mem.*, **13**, 599–608.
Hikosaka, O., Rand, M.K., Miyachi, S. & Miyashita, K. (1995) Learning of sequential movements in the monkey: process of learning and retention of memory. *J. Neurophysiol.*, **74**, 1652–1661.
Hull, C.L. (1943) *Principles of Behavior*. Appleton, New York, NY.
Iba, G.A. (1989) A heuristic approach to the discovery of macro-operators. *Mach. Learn.*, **3**, 285–317.
Jin, X. & Costa, R.M. (2010) Start/stop signals emerge in nigrostriatal circuits during sequence learning. *Nature*, **466**, 457–462.
Jog, M.S., Kubota, Y., Connolly, C.I., Hillegaart, V. & Graybiel, A.M. (1999) Building neural representations of habits. *Science*, **286**, 1745–1749.
Johnson, A., van der Meer, M.A.A. & Redish, A.D. (2007) Integrating hippocampus and striatum in decision-making. *Curr. Opin. Neurobiol.*, **17**, 692–697.
Keele, S.W., Ivry, R., Mayr, U., Hazeltine, E. & Heuer, H. (2003) The cognitive and neural architecture of sequence representation. *Psychol. Rev.*, **110**, 316–339.
Kelly, R.M. & Strick, P.L. (2004) Macro-architecture of basal ganglia loops with the cerebral cortex: use of rabies virus to reveal multisynaptic circuits. *Prog. Brain Res.*, **143**, 449–459.
Keramati, M., Dezfouli, A. & Piray, P. (2011) Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Comput. Biol.*, **7**, e1002055.
Kolter, Z., Plagemann, C., Jackson, D.T., Ng, A. & Thrun, S. (2010) A probabilistic approach to mixed open-loop and closed-loop control, with application to extreme autonomous driving. *Proc. of the IEEE Int.Conf.on Robotics & Automation (ICRA)*. Anchorage, Alaska, USA.
Korf, R.E. (1985) Macro-operators: a weak method for learning. *Artif. Intell.*, **26**, 35–77.
Kubota, Y., Liu, J., Hu, D., DeCoteau, W.E., Eden, U.T., Smith, A.C. & Graybiel, A.M. (2009) Stable encoding of task structure coexists with flexible coding of task events in sensorimotor striatum. *J. Neurophysiol.*, **102**, 2142–2160.
Lashley, K.S. (1951) The problem of serial order in behavior. In Jeffress, L.A. (Ed.) *Cerebral Mechanisms in Behavior*. Wiley, NY, pp. 112–136.
Lehéricy, S., Benali, H., Van de Moortele, P.-F., Pélégrini-Issac, M., Waechter, T., Ugurbil, K. & Doyon, J. (2005) Distinct basal ganglia territories are engaged in early and advanced motor sequence learning. *Proc. Nat. Acad. Sci. USA*, **102**, 12566–12571.
Matsumoto, N., Hanakawa, T., Maki, S., Graybiel, A.M. & Kimura, M. (1999) Nigrostriatal dopamine system in learning to perform sequential motor tasks in a predictive manner. *J. Neurophysiol.*, **82**, 978–998.
Mahadevan, S. (1996) Average reward reinforcement learning: foundations, algorithms, and empirical results. *Mach. Learn.*, **22**, 159–195.
Matsuzaka, Y., Picard, N. & Strick, P.L. (2007) Skill representation in the primary motor cortex after long-term practice. *J. Neurophysiol.*, **97**, 1819–1832.
Mcgovern, E.A. (2002) *Autonomous Discovery of Temporal Abstractions From Interaction With An Environment*. University of Massachusetts, Amherst.
Miyachi, S., Hikosaka, O., Miyashita, K., Kárádi, Z. & Rand, M.K. (1997) Differential roles of monkey striatum in learning of sequential hand movement. *Exp. Brain Res.*, **115**, 1–5.
Miyachi, S., Hikosaka, O. & Lu, X. (2002) Differential activation of monkey striatal neurons in the early and late stages of procedural learning. *Exp. Brain Res.*, **146**, 122–126.
Miyashita, K., Rand, M.K., Miyachi, S. & Hikosaka, O. (1996) Anticipatory saccades in sequential procedural learning in monkeys. *J. Neurophysiol.*, **76**, 1361–1366.
Montague, P.R., Dayan, P. & Sejnowski, T.J. (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.*, **16**, 1936–1947.
Moussa, R., Poucet, B., Amalric, M. & Sargolini, F. (2011) Behavior contributions of dorsal striatal subregions to spatial alternation behavior. *Learn. Mem.*, **18**, 444–451.
Mowrer, O.H. & Jones, H. (1945) Habit strength as a function of the pattern of reinforcement. *J. Exp. Psychol.*, **35**, 293–311.
Nakahara, H., Doya, K. & Hikosaka, O. (2001) Parallel cortico-basal ganglia mechanisms for acquisition and execution of visuomotor sequences - a computational approach. *J. Cogn. Neurosci.*, **13**, 626–647.

Neuringer, A. (2004) Reinforced variability in animals and people: implications for adaptive action. *Am. Psychol.*, **59**, 891–906.
Neuringer, A. & Jensen, G. (2010) Operant variability and voluntary action. *Psychol. Rev.*, **117**, 972–993.
Newell, K.M., Liu, Y.T. & Mayer-Kress, G. (2001) Time scales in motor learning and development. *Psychol. Rev.*, **108**, 57–82.
Nissen, M.J. & Bullemer, P. (1987) Attentional requirements of learning: performance measures evidence from. *Cogn. Psychol.*, **19**, 1–32.
Niv, Y., Daw, N.D., Joel, D. & Dayan, P. (2007) Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology*, **191**, 507–520.
O'Doherty, J.P., Dayan, P., Schultz, J., Deichmann, R., Friston, K. & Dolan, R.J. (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, **304**, 452–454.
O'Reilly, R.C., Herd, S.A. & Pauli, W.M. (2010) Computational models of cognitive control. *Curr. Opin. Neurobiol.*, **20**, 257–261.
Ostlund, S.B., Winterbauer, N.E. & Balleine, B.W. (2009) Evidence of action sequence chunking in goal-directed instrumental conditioning and its dependence on the dorsomedial prefrontal cortex. *J. Neurosci.*, **29**, 8280–8287.
Overmann, S.R. & Denny, M.R. (1974) The free-operant partial reinforcement effect: a discrimination analysis. *Learn. Motiv.*, **5**, 248–257.
Packard, M.G. & McGaugh, J.L. (1996) Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiol. Learn. Mem.*, **65**, 65–72.
Platt, J.R. & Day, R.B. (1979) A hierarchical response-unit analysis of resistance to extinction following fixed-number and fixed-consecutive-number reinforcement. *J. Exp. Psychol. Anim. Behav. Process.*, **5**, 307–320.
Poldrack, R.A., Sabb, F.W., Foerde, K., Tom, S.M., Asarnow, R.F., Bookheimer, S.Y. & Knowlton, B.J. (2005) The neural correlates of motor skill automaticity. *J. Neurosci.*, **25**, 5356–5364.
Puterman, M.L. (1994) *Markov Decision Processes*. Wiley Interscience, New York.
Randløv, J. (1998) Learning macro-actions in reinforcement learning. In Kearns, M.J., Solla, S.A. & Cohn, D.A. (Eds). *Advances in Neural Information Processing Systems 11*. MIT Press, Denver, CO, USA, pp. 1045–1051.
Rangel, A., Camerer, C. & Montague, P.R. (2008) A framework for studying the neurobiology of value-based decision making. *Nat. Rev. Neurosci.*, **9**, 545–556.
Redish, A.D., Jensen, S. & Johnson, A. (2008) A unified framework for addiction: vulnerabilities in the decision process. *Behav. Brain Sci.*, **31**, 415–437; discussion 437–487.
Restle, F. (1957) Discrimination of cues in mazes: a resolution of the place-vs.-response question. *Psychol. Rev.*, **64**, 217–228.
Reynolds, J.N., Hyland, B.I. & Wickens, J.R. (2001) A cellular mechanism of reward-related learning. *Nature*, **413**, 67–70.
Ribas-Fernandes, J.J.F., Solway, A., Diuk, C., McGuire, J.T., Barto, A.G., Niv, Y. & Botvinick, M.M. (2011) A neural signature of hierarchical reinforcement learning. *Neuron*, **71**, 370–379.
Ritchie, B.F., Aeschliman, B. & Pierce, P. (1950) Studies in spatial learning: VIII. Place performance and the acquisition of place dispositions. *J. Comp. Physiol. Psychol.*, **43**, 73–85.
Rondi-Reig, L., Petit, G.H., Tobin, C., Tonegawa, S., Mariani, J. & Berthoz, A. (2006) Impaired sequential egocentric and allocentric memories in forebrain-specific-NMDA receptor knock-out mice during a new task dissociating strategies of navigation. *J. Neurosci.*, **26**, 4071–4081.
Sakai, K., Kitaguchi, K. & Hikosaka, O. (2003) Chunking during human visuomotor sequence learning. *Exp. Brain Res.*, **152**, 229–242.
Schmitzer-Torbert, N. & Redish, A.D. (2004) Neuronal activity in the rodent dorsal striatum in sequential navigation: separation of spatial and reward responses on the multiple T task. *J. Neurophysiol.*, **91**, 2259–2272.
Schultz, W. & Dickinson, A. (2000) Neuronal coding of prediction errors. *Annu. Rev. Neurosci.*, **23**, 473–500.
Schultz, W., Dayan, P. & Montague, P.R. (1997) A neural substrate of prediction and reward. *Science*, **275**, 1593–1599.
Schwarting, R.K.W. (2009) Rodent models of serial reaction time tasks and their implementation in neurobiological research. *Behav. Brain Res.*, **199**, 76–88.
Shah, A. (2008) *Biologically-Based Functional Mechanisms of Motor Skill Acquisition*. University of Massachusetts, Amherst.
Shah, A. & Barto, A.G. (2009). Effect on movement selection of an evolving sensory representation: a multiple controller model of skill acquisition. *Brain Res.*, **1299**, 55–73. Elsevier B.V.
Shull, R.L. & Grimes, J.A. (2003) Bouts of responding from variable-interval reinforcement of lever pressing by rats. *J. Exp. Anal. Behav.*, **80**, 159–171.
Shull, R.L., Gaynor, S.T. & Grimes, J.A. (2002) Response rate viewed as engagement bouts: resistance to extinction. *J. Exp. Anal. Behav.*, **77**, 211–231.
Skinner, B.F. (1938) *The Behavior of Organisms*. Appleton-Century-Crofts, New York.
Spence, K.W. (1956) *Behavior Theory and Conditioning*. Yale University Press, New Haven.
Sutton, R.S. & Barto, A.G. (1998) *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.
Tang, C., Pawlak, A.P., Prokopenko, V. & West, M.O. (2007) Changes in activity of the striatum during formation of a motor habit. *Eur. J. Neurosci.*, **25**, 1212–1227.
Terrace, H.S. (1991) Chunking during serial learning by a pigeon: I. Basic evidence. *J. Exp. Psychol. Anim. Behav. Process.*, **17**, 81–93.
Tolman, E.C., Ritchie, B.F. & Kalish, D. (1946) Studies in spatial learning: II. Place learning versus response learning. *J. Exp. Psychol.*, **36**, 221–229.
Tsitsiklis, J.N. & Roy, B.V. (1999) Average cost temporal-difference learning. *Automatica*, **35**, 1799–1808.
Turner, R.S. & Desmurget, M. (2010) Basal ganglia contributions to motor control: a vigorous tutor. *Curr. Opin. Neurobiol.*, **20**, 704–716.
Watkins, C.J.C.H. (1989) *Learning from Delayed Rewards*. Cambridge University, Cambridge.
Willingham, D.B. (1998) A neuropsychological theory of motor skill learning. *Psychol. Rev.*, **105**, 558–584.
Willingham, D.B., Nissen, M.J. & Bullemer, P. (1989) On the development of procedural knowledge. *J. Exp. Psychol. Learn. Mem. Cogn.*, **15**, 1047–1060.
Wolpert, D.M. & Flanagan, J.R. (2010) Motor learning. *Curr. Biol.*, **20**, R467–R472.
Wolpert, D.M., Ghahramani, Z. & Flanagan, J.R. (2001) Perspectives and problems in motor learning. *Trends Cogn. Sci.*, **5**, 487–494.
Yin, H.H. & Knowlton, B.J. (2004) Contributions of striatal subregions to place and response learning. *Learn. Mem.*, **11**, 459–463.
Yin, H.H., Knowlton, B.J. & Balleine, B.W. (2004) Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *Eur. J. Neurosci.*, **19**, 181–189.
Yin, H.H., Knowlton, B.J. & Balleine, B.W. (2005) Blockade of NMDA receptors in the dorsomedial striatum prevents action-outcome learning in instrumental conditioning. *Eur. J. Neurosci.*, **22**, 505–512.
Yin, H.H., Knowlton, B.J. & Balleine, B.W. (2006) Inactivation of dorsolateral striatum enhances sensitivity to changes in the action-outcome contingency in instrumental conditioning. *Behav. Brain Res.*, **166**, 189–196.
Yin, H.H., Ostlund, S.B. & Balleine, B.W. (2008) Reward-guided learning beyond dopamine in the nucleus accumbens: the integrative functions of cortico-basal ganglia networks. *Eur. J. Neurosci.*, **28**, 1437–1448.

Appendix

Estimating the cost of sequence-based action selection

We aim to estimate $C(s, a, a')$ by the samples of the error signal experienced by taking action a', after action a in state s.

$C(s, a, a')$ is defined as:

$$C(s,a,a') = E[\delta(s',a')|s,a] \tag{A1}$$

We maintain that:

$$C(s,a,a') = E\left[\frac{P(a'|s,a)}{P(a'|s')}\delta(s',a')|s,a,a'\right] \tag{A2}$$

This follows from:

$$\begin{aligned} C(s,a,a') &= E\left[\frac{P(a'|s,a)}{P(a'|s')}\delta(s',a')|s,a,a'\right] \\ &= \sum_{s'}\frac{P(a'|s,a)P(s'|s,a,a')}{P(a'|s')}\delta(s',a') \\ &= \sum_{s'}P(s'|s,a)\delta(s',a') \\ &= E[\delta(s',a')|s,a] \end{aligned} \tag{A3}$$

$P(a'|s, a)$ and $P(a'|s')$ can be estimated directly by counting number of times action a' has been taken after s, a and after s', respectively. Given these, $C(s, a, a')$ can be estimated by averaging over the samples of the error signal multiplied by the factor α_t:

$$\alpha_t = \frac{P(a_t|s_{t-1}, a_{t-1})}{P(a_t|s_t)} \tag{A4}$$

α_t is in fact the percentage of taking action a' after s and a is due to being in state s'. If action a' is not available in state s', we assume this factor is infinity. Given α_t, the cost function is estimated using Eqn 8.

Implementation details

The implementation of the mixed architecture is similar to model-based hierarchical reinforcement learning (RL). After execution of a macro action finished, the characteristics of the underlying semi-Markov decision process (MDP) are updated. That is, the total reward obtained through executing the macro action, and the total time spent for executing the macro action, are used for updating the reward and transition delay functions. Here, because we assumed that reward function depends on the states, and not state–action pairs, the total reward obtained through the macro action cannot be assigned to the state in which the macro action was launched. This is because the total reward obtained through the macro action can be different from the reward of the state. For addressing this problem, when an action sequence was formed, a temporary extended auxiliary state is added to the MDP, to which the agent enters when the macro action starts, and exits when the macro action finished. The transition delay and reward function of this auxiliary state is updated using the rewards obtained through executing the macro action, and the time spent to take the macro action.

Because according to the Bellman equation for average reward semi-Markov RL, the Q-values satisfy Eqn 3 (Puterman, 1994), when a non-exploratory action is taken (the action with highest value is taken), we update the average reward as follows:

$$R_{t+1} = (1 - \sigma)R_t + \sigma\left[\frac{r_t + V(s_{t+1}) - V(s_t)}{d_t}\right] \tag{A5}$$

where σ is the learning rate of the average reward. For the action selection (in the model-based system), soft-max rule is used:

$$P(s|a) = \frac{e^{\beta Q(s,a)}}{\sum_{a'} e^{\beta Q(s,a')}} \tag{A6}$$

where parameter β determines the rate of exploration. For computing model-based values, a tree-search algorithm was applied with the depth of search of three levels. After this level, goal-directed estimations are replaced by model-free estimations.

Due to the fluctuations of $C(s, a, a')$ and $\bar{R}\tau$ at the point they meet, a series of sequence formation/decomposition may happen before the two curves become separated. To address this issue, an asymmetric rule for sequence decomposition is used, and a sequence decomposes only if $-C(s, a, a') > 1.6\bar{R}\tau$. Also, for simplicity, we assumed that macro actions could not be cyclic.

Internal parameters of the model are shown in Table A1.

TABLE A1. Free parameters of the model, and their assigned values

Parameter	Value
Update rate of the reward function	0.05
Update rate of the average reward (σ)	0.002
Update rate of the cost of sequence-based control (η_C)	0.001
Initial value of the cost of sequence-based control	−2
Rate of exploration (β)	4