PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

rstb.royalsocietypublishing.org

Research

Cite this article: Dezfouli A, Lingawi NW, Balleine BW. 2014 Habits as action sequences: hierarchical action control and changes in outcome value. *Phil. Trans. R. Soc. B* **369**: 20130482.
http://dx.doi.org/10.1098/rstb.2013.0482

One contribution of 18 to a Theme Issue 'The principles of goal-directed decision-making: from neural mechanisms to computation and robotics'.

Subject Areas:
computational biology, behaviour, cognition, neuroscience

Keywords:
goal-directed actions, habits, decision-making

Author for correspondence:
Bernard W. Balleine
e-mail: bernard.balleine@sydney.edu.au

Royal Society Publishing

Habits as action sequences: hierarchical action control and changes in outcome value

Amir Dezfouli, Nura W. Lingawi and Bernard W. Balleine

Brain and Mind Research Institute, University of Sydney, 100 Mallett St., Camperdown,
New South Wales 2050, Australia

Goal-directed action involves making high-level choices that are implemented using previously acquired action sequences to attain desired goals. Such a hierarchical schema is necessary for goal-directed actions to be scalable to real-life situations, but results in decision-making that is less flexible than when action sequences are unfolded and the decision-maker deliberates step-by-step over the outcome of each individual action. In particular, from this perspective, the offline revaluation of any outcomes that fall within action sequence boundaries will be invisible to the high-level planner resulting in decisions that are insensitive to such changes. Here, within the context of a two-stage decision-making task, we demonstrate that this property can explain the emergence of habits. Next, we show how this hierarchical account explains the insensitivity of over-trained actions to changes in outcome value. Finally, we provide new data that show that, under extended extinction conditions, habitual behaviour can revert to goal-directed control, presumably as a consequence of decomposing action sequences into single actions. This hierarchical view suggests that the development of action sequences and the insensitivity of actions to changes in outcome value are essentially two sides of the same coin, explaining why these two aspects of automatic behaviour involve a shared neural structure.

1. Introduction

Goal-directed action is a form of decision-making guided by encoding the relationship between actions and their consequences, and the value of those consequences [1–5]. Outcome devaluation studies provide direct evidence that both humans and other animals engage in this form of action control. For example, in a typical experiment, an agent is first trained to perform two different actions that earn different food outcomes. After this training, an outcome devaluation treatment is conducted off baseline or 'offline'; that is, in a situation where the outcome is presented without the action being performed, a treatment that generally involves sating the animals on one of the two outcomes to decrease its value. Subsequently, back 'online', a test is conducted in which choice between the two actions is assessed in the absence of the outcome. Typically, when given this choice, humans and other animals decrease their performance of the action that previously delivered the now devalued outcome, demonstrating that such actions reflect both the relationship to their consequences and the value of those consequences [6].

Extended training makes goal-directed actions habitual or automatic. This automaticity has two manifestations: (i) inflexibility of actions to the offline changes in the value of their outcomes [2,7] and (ii) the concatenation of actions executed together to form action sequences that are then treated as a single response unit [8,9]. These two aspects of automaticity share a similar neural structure (e.g. [10]), however, computationally, they have been attributed to two different models: insensitivity to changes in outcome value has often been interpreted as evidence for a model-free reinforcement learning (RL) account of instrumental conditioning [11,12], whereas the development of action sequences has been linked to hierarchical RL [13]. Here, building on previous work [13–15], we demonstrate that the insensitivity of specific actions to changes in outcome value can be a consequence of developing action sequences and, therefore, that both types of automaticity can be reconciled within hierarchical model-based RL. This account proposes that, early in training, decision-making involves a goal-directed controller that deliberates over the consequences of individual actions in a step-by-step manner. After sufficient training, however, action sequences form and goal-directed actions become hierarchically organized. After this stage, offline changes in the value of outcomes that fall within the boundaries of an action sequence will be invisible to the high level controller and, as such, decisions will appear insensitive to such changes. Here, we elaborate this account of habits within the context of two decision-making tasks in humans and rodents. First, however, we briefly introduce hierarchical decision-making and its properties.

2. Hierarchical decision-making

In many situations, choices are not followed immediately by an outcome and require a number of subsequent choices prior to outcome delivery. In order to make goal-directed decisions, the available options and their consequences need to be considered in addition to the immediate, one-step-ahead consequence of each action. However, this process is computationally effortful as the number of factors involved in the learning and decision-making process grows exponentially with each additional step required to reach the goal. This makes simple goal-directed decision-making unscalable to complex environments [16].

An alternative view of goal-directed action that we have recently developed [14,15] uses *hierarchical decision-making* [13], proposing that, rather than deliberating step-by-step, individual actions are concatenated to form an *action sequence* or *action chunk* [8,9,17–21]. Action sequences are typically studied in motor-skill learning (referred to as motors skills), although they are also observed in other domains [22,23]. Utilization of action sequences improves learning, planning and performance of actions. Learning is enhanced because outcomes are learned for the whole sequence rather than for each of the individual actions. Subsequently, planning becomes faster: instead of deliberating over each individual action at the choice point, the decision-maker needs only look at the final outcome of each action sequence. Thus, the scalability problem is overcome by assuming that actions are structured in the form of action sequences, and the decision-maker has only to choose among those action sequences. As a result, performance of actions will also improve, as actions within the sequence will run off automatically without planning, enabling fast and precise performance of actions. When the execution of an action sequence finishes, the decision-maker makes another choice and this process continues (see also William James notion of habits for a similar conception ([24] ch. IV on *habits*)).

An alternative approach to the limitations of simple goal-directed decision-making is to use model-free RL. This approach proposes that, whenever simple goal-directed decision-making is either not reliable [11,25] or not required [12], actions that have previously procured the most valuable rewards will automatically be selected for execution without consideration of their outcomes. Within this model, habitual actions operate outside the scope of goal-directed decision-making and, therefore, their computational role is not related to making goal-directed decision-making scalable, which is in contrast to the role of habits in hierarchical decision-making. Along the same lines, the algorithmic representation of automatic actions in the model-free RL account involves processing the current situation at each step followed by selecting an action and then updating the value of the selected action. This contrasts with hierarchical decision-making in which learning and action selection do not occur during the performance of habitual actions.

3. Planning mistakes and slips of action in hierarchical decision-making

The efficiency gained using hierarchical decision-making comes at the cost of certain types of error that can be divided into two broad categories: errors in action planning, known as ‘planning mistakes’, and errors in action execution, known as ‘slips of action’ [26,27].

(a) Planning mistakes

Although mistakes can take various forms (see for example [28] for different types of optimality in hierarchical RL), an important type of planning mistake relates to the inflexibility of decisions after offline changes in outcome value. One effect of chunking actions together and turning them into a single response unit is that the representation of the action sequence is independent of its embedded individual actions and their outcomes. Although this higher level representation of actions makes decision-making easier and faster, it also renders the evaluation of action sequences insensitive both to offline changes in individual action–outcome contingencies and to changes in the value of any outcomes delivered within the sequence boundaries [15]. After such changes, the planner might, therefore, continue to choose an action sequence even though it is less likely to result in a valued outcome than previously and another action now has a higher value under the new conditions.

(b) Slips of action

Errors associated with slips of action are not an inherent part of hierarchical decision-making but occur because of the way the execution of action sequences is realized at a mechanistic level. Slips of action also take various forms but are generally related to the ‘open-loop’ or ‘feed-forward’ performance of action sequences [8,18,29]; that is, once an action sequence is launched, the actions in the sequence will be executed automatically up to the end of the sequence. As a consequence, the feedback received from the environment after the execution of an action (i.e. the outcome of an action) will not affect which action is taken next because this is determined by the order of actions in the action sequence. This open-loop property is contrasted with ‘closed-loop’ control that is engaged after the execution of each individual action, meaning the planner selects actions according to the outcome of previous actions.

This open-loop property can, however, cause slips of action because the next action in the sequence will be executed automatically even if the outcome indicates another action should have been taken. For example, a monkey trained to press a sequence of buttons to earn a reward will continue to execute the sequence up to the end even if the reward is delivered within the sequence, and the performance of the rest of the sequence is not required [30] (see also [31] for a similar phenomenon in maze navigation). It is worth mentioning that, although action selection is divorced from environmental feedback under these conditions, this does not imply that the execution of actions is feedback-free; action execution can operate in a closed-loop manner by relying on sensory feedback and environmental stimuli. Another slip of action, known as a 'capture error' or 'strong habit intrusion' [26,27], describes a situation in which the decision-maker produces an action that is a part of a well-practiced action sequence resulting in the subsequent action sequence unintentionally running off automatically. The situation 'I only meant to take off my shoes, but took off my socks as well' [26] represents an example of this kind of error.

The model-free approach to automaticity also entails certain types of errors with respect to simple goal-directed decision-making. With respect to the sensitivity of habitual actions to offline changes in outcome values, whereas the hierarchical account predicts that automatic actions will be performed only when the outcome falls within the action sequences (due to a mistake in planning), the model-free account predicts that automatic actions will be executed irrespective of the position of the devalued outcome within or outside a sequence of actions. Furthermore, slips of action are not anticipated by the model-free account of automatic actions.

4. Testing predictions of the hierarchical account: planning mistakes and insensitivity to changes in outcome value

Having introduced the hierarchical account, in this section we describe two experiments testing predictions derived from this account. Evidence from humans and other animals suggests that over-training causes actions to become insensitive to changes in the value of their consequences or outcome. From the perspective of hierarchical decision-making, this insensitivity to outcome revaluation/devaluation is linked to the planning mistakes described above [15]. This connection is more readily visible in the performance of a two-stage task that we recently developed [14] based on that of Daw *et al.* [32]. In this section, we first describe an experimental test of this connection by developing this task within the context of the outcome revaluation/devaluation paradigm (see [14] for detailed methods). We then explore the issue of how the hierarchical account explains variations in the sensitivity of actions to outcome devaluation in instrumental conditioning (see [15] for simulations) and describe a second experiment in which we assess a prediction regarding the effect of extinction on the performance of habit sequences.

(a) Experiment 1: outcome devaluation/revaluation in a two-stage task

(i) Method

In this version of the two-stage task (cf. [14] for details), 15 human subjects were instructed to make a binary choice at stage 1 (i.e. A_1 or A_2), the outcome of which was either O_1 or O_2 (which are two distinct two-armed slot machines). Subjects could then make another binary choice in stage 2, choosing one or other arm (i.e. R_1 or R_2), which had either a rewarding or a neutral result (e.g. $ versus X; figure 1*a*). At the first stage, A_1 typically led to O_1 and A_2 to O_2 (common trials; figure 1*b*). However, on a minority of trials, A_1 led to O_2, and A_2 to O_1 (rare trials; figure 1*b*). Depending on the outcome in stage 1 (O_1 or O_2), the choices in stage 2 could have a rewarding or neutral result. The role of stage 2 choices was, therefore, to manipulate the value of O_1 and O_2 (figure 1*b*). In order to change the value of these outcomes during the session, the probability of a reward for each stage 2 choice increased or decreased randomly on each trial (e.g. an R_1 choice in O_1 that was leading to money, might now lead to an X), which will cause frequent devaluation/revaluation of the O_1 and O_2 outcomes during the task (cf. [14] for details). Each participant completed 270 trials.

Changes in outcome value are usually accomplished in revaluation/devaluation studies by offline treatments such as specific satiety and taste aversion learning [3,33]. In this task to test whether stage 1 actions were guided by the value of their outcomes, a small number of rare trials were inserted among common trials such that a stage 1 choice occasionally led to the outcome of the other choice (e.g. A_2 will lead to O_1 which is usually the outcome of A_1). These rare trials allowed the offline manipulation of outcome values; that is, they allow the value of O_1 to be manipulated (devalued/revalued) without A_1 being chosen at stage 1 (figure 1*c*,*d*).

(ii) Results

Using this design, we found that stage 1 choices were sensitive to the offline revaluation/devaluation of their outcomes confirming that these actions were goal-directed (figure 1*e*). As two steps are required to reach a reward, however, goal-directed action selection can be extended to engage hierarchical decision-making. From this perspective, subjects can combine stage 1 and stage 2 actions and build action sequences A_1R_1, A_1R_2, A_2R_1, A_2R_2, such that, at stage 1, the choice will be between all actions including both the single actions (A_1, A_2, etc.) and action sequences (A_1R_1, A_1R_2, etc.), based on their contingency to reward (figure 2*a*). If subjects are using action sequences then we should expect to observe the open-loop execution of actions. In general, the best action in O_1 is independent from that in O_2, and, therefore, the choice of stage 2 action (R_1 versus R_2) should be based on the outcome of the stage 1 action. We found, however, that when the previous trial was rewarded, and subjects repeated the same stage 1 action (A_1 or A_2), they also tended to repeat the same stage 2 action (R_1 or R_2), irrespective of the outcome of the stage 1 action (figure 2*b*). In these cases, the stage 2 action was determined at stage 1 when the sequence was launched, which is one marker for the development of such sequences (figure 2*c*; see [14] for a similar pattern in reaction times). This observed open-loop execution of actions is not related to the generalization of action values across O_1 and O_2 (e.g. a high value of R_1 in O_1 entails a high value for R_1 in O_2), as subjects repeat the same stage 2 action only if they have executed the same action at stage 1 (figure 2*c*), whereas in the case of generalization of values, we expect subjects to repeat the same stage 2 action even if the same stage 1 action was not chosen. It is also possible that subjects formed longer action sequences (A_1R_1, A_1R_2, …); however, within the current task, the existence of such action sequences cannot be detected either using reaction times, because of the

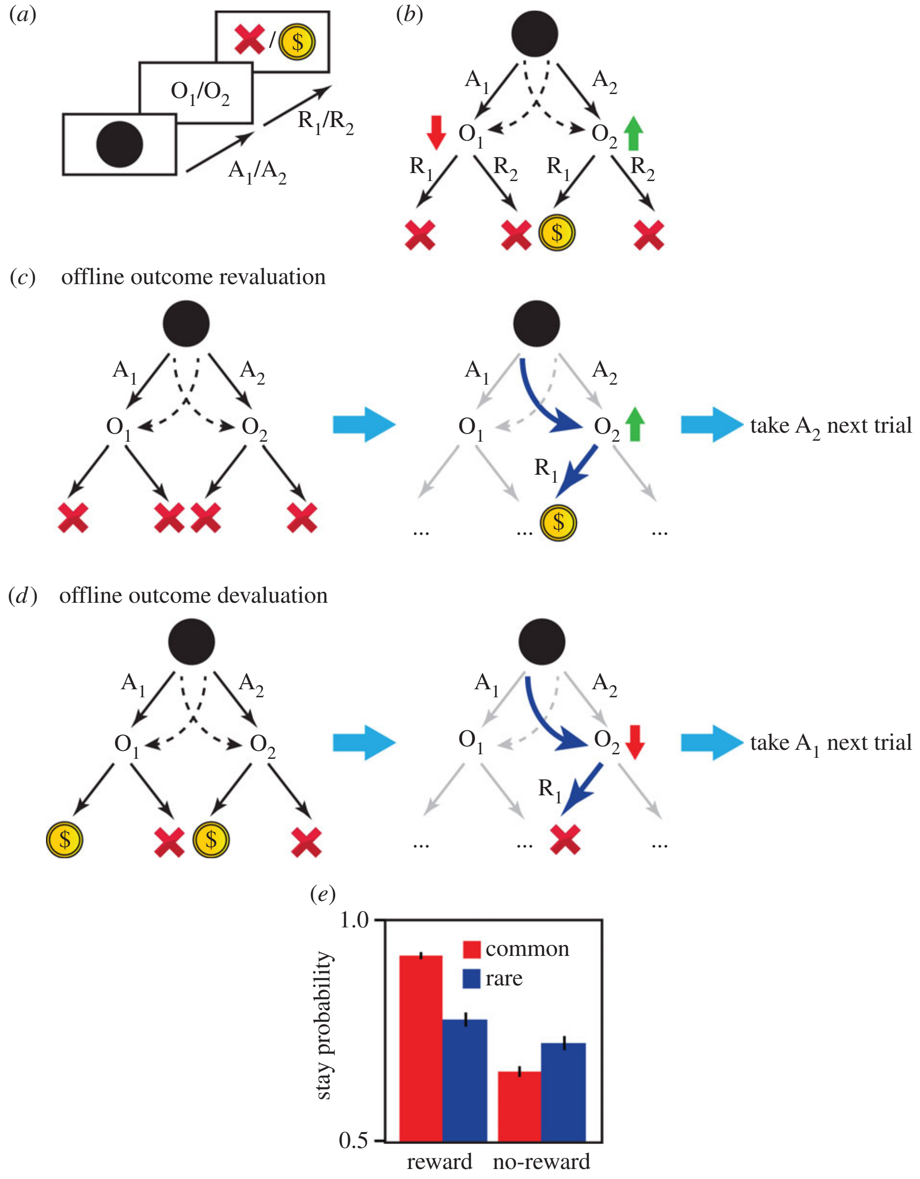

Figure 1. The two-stage task. (*a*) At stage 1, subjects choose between A_1 and A_2, and the outcome can be O_1 or O_2. They then make another choice (R_1 versus R_2), which can have a rewarding or neutral result ($ versus X). (*b*) The outcomes of A_1 and A_2 are commonly O_1 and O_2, respectively. On approximately 30% of trials, however, these relationships switch and A_1 leads to O_2, and A_2 leads to O_1 (dashed arrows). The values of O_1 and O_2 depend on the probability of earning a reward after the stage 2 actions, R_1 or R_2. In the current illustration, actions in O_1 are not rewarded and so O_1 has a low value, whereas O_2 has a high value because an action in O_2 is rewarded. Each stage 2 action will result in a reward with either a high (0.7) or a low probability (0.2) independent of the other actions. In each trial, there is a small chance (1 : 7) that these reward probabilities reset to high or low randomly, which causes frequent devaluation/revaluation of outcomes (O_1 and O_2) across the session. (*c*) Both choices at O_2 have a low value (left), and in the next trial, the reward probability of one of the actions becomes high (right). In a rare trial, the subject executes A_1 and receives O_2 instead of O_1, and then the action in O_2 becomes rewarded (blue arrows indicate executed actions), which causes offline revaluation of O_2. Thus A_2 should be taken in the subsequent trials to reach O_2. (*d*) O_2 has a high value (left), and on the next trial, the action that was rewarded previously in O_2 is not rewarded (right). In a rare trial, the subject chooses A_1 and receives O_2 instead of O_1, after which the action in O_2 is not rewarded. This causes the offline devaluation of O_2 and, in subsequent trials, A_1 should be chosen so as to avoid O_2. (*e*) The probability of selecting the same stage 1 action on the next trial as a function of whether the previous trial was rewarded, and whether it was a common or rare trial (mixed-effect logistic regression analysis with all coefficients treated as random effects across subjects; 'reward' × 'transition type' interaction: coefficient estimate $= 0.41$; s.e. $= 0.11$; $p < 5 \times 10^{-4}$). Based on (*c*,*d*), when rewarded after a rare trial, a different stage 1 action should be taken, whereas when unrewarded after a rare trial, the same stage 1 action should be taken. This pattern is reversed if the previous trial is common: the same stage 1 action should be taken if the previous trial is rewarded, and a different stage 1 action should be taken if it is unrewarded. This stay/switch pattern predicts an interaction between reward and transition type if stage 1 actions are guided by their outcomes values, which is consistent with the behavioural results (see [14]). Error bars, 1 s.e.m.

inter-trial intervals, or using the open-loop property of sequences, because the initial state of the task (the black screen in figure 1*a*,*b*) is the same for both actions.

Based on our observation of action sequences, we predicted that we would also observe mistakes in planning. It is important to note that the outcomes of the stage 1 choices (i.e. O_1 and O_2) fall within the boundaries of the action sequences, implying that these sequences were not revalued by offline changes in the value of O_1 and O_2 during rare trials (figure 3). This failure to adjust the value of these sequences predicts, given that action sequences enter the action selection process at stage 1, that subjects will make systematic mistakes in choices after offline outcome devaluation/re-evaluation. For example, in a rare trial if the performance of a sequence of actions is

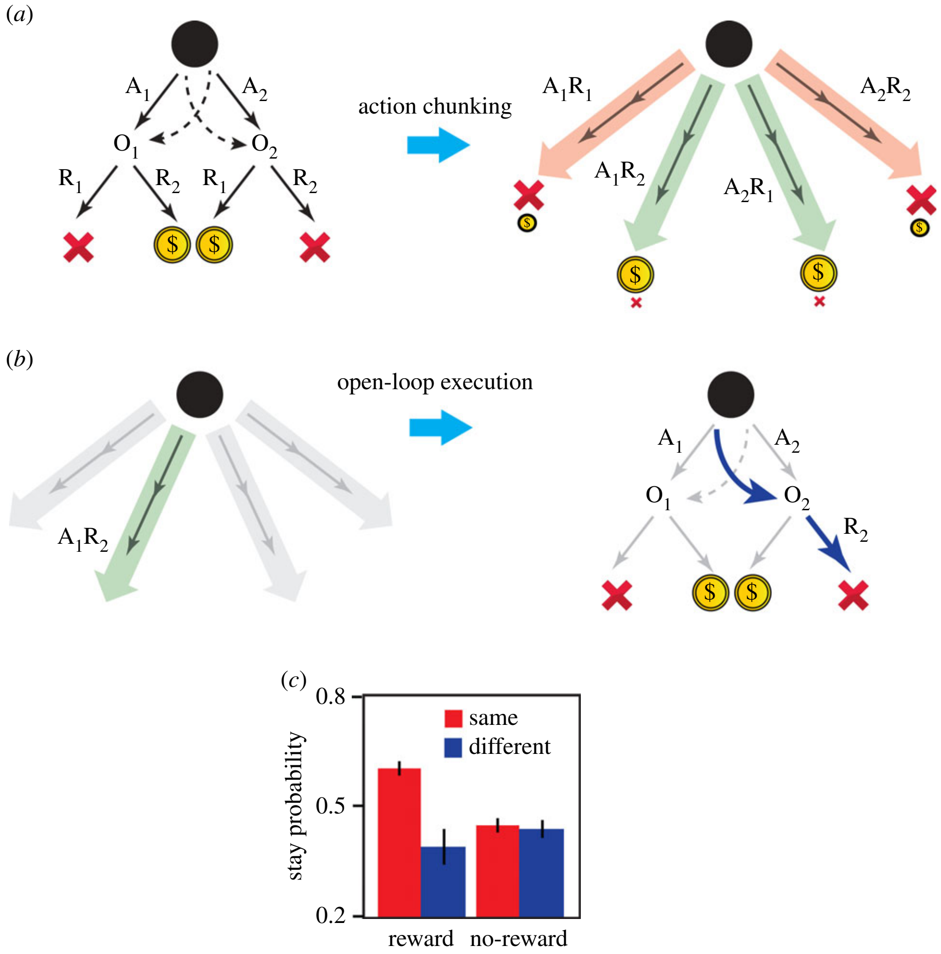

Figure 2. Evidence stage 1 actions are chunked with stage 2 actions to build action sequences. (*a*) Action sequences are evaluated based on their contingency to reward. In this illustration, A_1R_1 has a low value because most of the time it is not rewarded (it is occasionally rewarded on rare trials). A_1R_2 has a high value because it is rewarded most of the time (it is occasionally not rewarded on rare trials). At stage 1, the subject chooses between all the actions, including action sequences (e.g. A_1R_2) and single actions (e.g. A_1) based on their contingency to reward. (*b*) Execution of action sequences is open-loop. For example, the subject chooses the A_1R_2 sequence for execution (left) such that, after A_1, even if the trial is a rare transition to O_2, the subject still executes R_2, although clearly R_1 should be executed in O_2 (right). (*c*) The probability of selecting the same stage 2 action when the outcome of the stage 1 action was different from the previous trial. When the previous trial was rewarded, and the subject took the same first level action ('same'), they executed the previous action sequence and stayed on the same stage 2 action (mixed-effect logistic regression analysis with all coefficients treated as random effects across subjects; 'reward' × 'same action in stage 1' interaction: coefficient estimate = 1.02; s.e. = 0.38; $p < 0.008$). This occurred even if the outcome was different from the previous trial indicating that another stage 2 action should be taken. The main effect of reward was not significant ($p > 0.05$) implying that, when the subject took a different stage 1 action ('different'), the previous action sequence was not repeated, and thus the same stage 2 action was not repeated if the outcome of the stage 1 action was different from the previous trial (see [14]).

rewarded, subjects tend to repeat the same action sequence in the next trial (and thus the same stage 1 action), ignoring that the outcome of the other stage 1 action is revalued. Consistent with this, we found that being rewarded in the previous trial increased the likelihood of repeating the same stage 1 action irrespective of which outcome and its associated action (A_1 or A_2) were revalued/devalued (figure 3*b* and see [14] for simulations). In addition to this effect, because primitive actions also enter the action selection process at stage 1 and are sensitive to outcome devaluation/re-evaluation, the choices also exhibited sensitivity to outcome revaluation leading to a mixture of actions including those guided by outcome value (when primitive actions are selected; figure 1*e*), and those that are not (when action sequences are selected; figure 3).

(iii) Discussion

Within the context of this task we found, as detected via their open-loop property, that action sequences were performed and that, as a consequence, decisions were insensitive to a change in outcome value. These two aspects of performance are both addressed within a hierarchical account of instrumental conditioning, whereas the model-free account of automaticity only provides an account of the insensitivity of actions to outcome devaluation and cannot explain the slips of action at stage 2. In addition to the behavioural analysis provided here it can be shown, by directly comparing a family of hierarchical and model-free models on a trial-by-trial basis, that the hierarchical models provide a better fit to the subjects' choices than the model-free account [14].

Over and above these theoretical differences, at the application level the hierarchical account provides a basis for measuring the degree of competition and cooperation between goal-directed and habitual actions. Previous work (e.g. [11]) has viewed the interaction between these actions solely in terms of competition, whereas, in the current framework, the processes both cooperate and compete. At the point of initiation of action sequences, the goal-directed controller launches

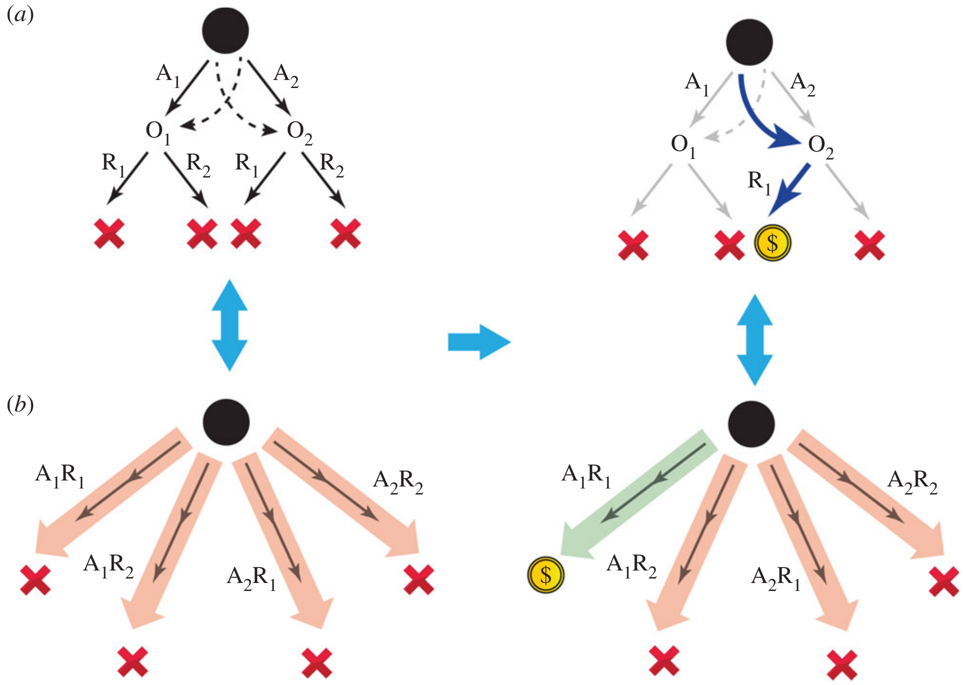

Figure 3. Insensitivity to outcome devaluation as a consequence of action sequences. (*a*) After a change in reward probabilities, the value of O_2 increases (right). On a rare trial, the subject is rewarded for taking R_1 at O_2, and thus O_2 revalues offline and on the next trial, according to simple goal-directed behaviour, A_2 should be taken. (*b*) Before the change in reward probabilities, all of the action sequences have low values (left). By receiving reward on a rare trial, the value of the executed action sequence increases (A_1R_1), whereas the rest of action sequences retain their previous value as the outcome (O_2) falls within their boundaries. Therefore, on the next trial, the probability that the subject will take the A_1R_1 action sequence is increased, in contrast to a simple goal-directed choice that indicates A_2 should be taken. After receiving a reward, there is tendency for the subject to take the same stage 1 action if they were rewarded on the previous trial. This is true even if the outcome of the other action is revalued, which predicts a main effect of reward on the probability of staying on the same stage 1 action, consistent with the behavioural results (see [14]; refer figure 1*e*; (analysis similar to figure 1; main effect of reward: coefficient estimate = 0.61; s.e. = 0.09; $p < 3 \times 10^{-11}$)).

habits (cooperation), whereas during the execution of action sequences, the planner can inhibit an ongoing action sequence to regain control (competition). This allows for the measurement of the degree of inhibition, as well as cooperation, between these two processes. For example, within the context of the two-stage task, the degree of cooperation between the processes can be measured by the proportion of times that action sequences are selected in stage 1 (e.g. using computational modelling), and the degree of inhibition between the processes can be measured by looking at the number of trials in which the subject inhibits an ongoing action sequence when an irrelevant outcome occurs at stage 2.

(b) Experiment 2: outcome devaluation and instrumental extinction

A number of studies in rodents have demonstrated that, under certain conditions, instrumental actions are divorced from the value of their consequences. For example, in a typical outcome devaluation experiment, animals are first trained to press a lever for a food reward. For some of the rats, the food is then devalued by pairing its consumption with gastric malaise, which can be induced by injections of lithium chloride. This process of outcome devaluation is known as conditioned taste aversion. Results show that rats receiving moderate instrumental training (around 250 lever press–outcome pairings) reduce lever press responding after outcome devaluation in subsequent extinction tests. However, after extended training (around 500 lever press–outcome pairings), lever pressing in rats is impervious to outcome devaluation; that is, the rats continue to respond on the lever even if the outcome it previously delivered has subsequently been paired with illness and the food reward is no longer desirable [2].

Behavioural evidence from motor skill learning experiments suggests that, when a sequence of actions is fixed, actions concatenate together over the course of training to form a single response. This is indicated by a reduction in the reaction times between actions [34,35]. Within this context, the training phase of a typical rodent instrumental conditioning experiment arranges a sequential organization between lever pressing (LP) and the magazine entry (EM) response required to reach the outcome, followed by more lever presses to earn the next outcome (figure 4*a*). This implies that after extended training, lever-press and magazine entry responses could become concatenated to form both LP → EM and EM → LP action sequences, presumably in a circular manner (figure 4*b*). In this situation, the outcome falls within the EM → LP action sequence, and, as such, this sequence should be predicted to retain its value despite any offline outcome devaluation, leading to a mistake in planning [15].

Assuming that, owing to inflexibility in decision-making, habits impose a cost on the decision-maker (e.g. by inducing planning mistakes and slips of action), from a normative perspective using habits can only be justified when their benefits, such as faster planning and execution of actions, exceeds their cost [12,15]. The advantage of being faster is determined by the value of time, which is usually measured in terms of the reward rate per unit time [36]. This view predicts, therefore, that, as the value of time drops, the control of an action should revert from being habitual to being goal-directed. In fact, outcome devaluation tests are usually conducted under extinction conditions in which animals do

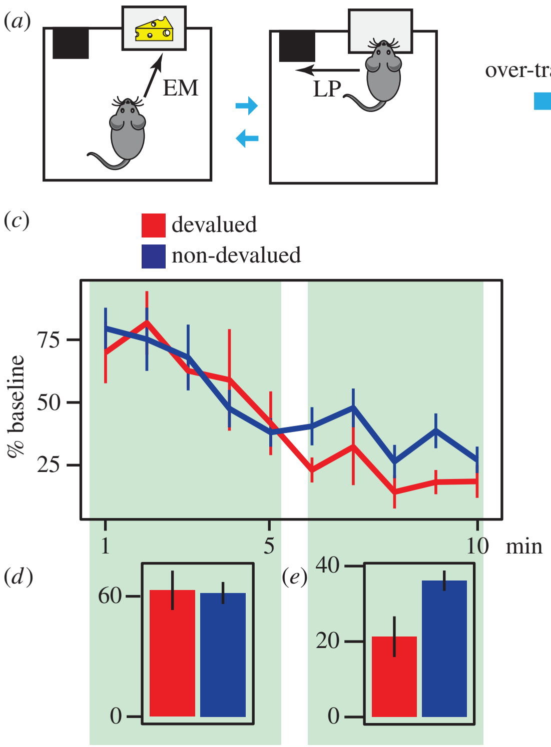

Figure 4. The sensitivity of actions to outcome devaluation in instrumental conditioning. (*a*) Animals make an 'enter the magazine' (EM) response to consume the outcome, and then make another lever-press response (LP) to earn the next outcome. (*b*) By over-training, the EM and LP responses combine to make the action sequences EM → LP and LP → EM, which presumably can be circular. The outcome falls within the boundaries of the EM → LP action sequence and, as a consequence, after offline devaluation of outcome, the sequence retains its previous value. Consequently during the test, animals for which the outcome is devalued press the lever at the same rate as those for which the outcome was not devalued. (*c*) Number of responses as a percentage of baseline response rates (calculated for each subject as an average of their four RI60 s sessions) in each minute of the extinction test conducted after conditioned taste aversion. (*d*) Number of lever presses during first 5 min of the extinction test averaged over subjects. (*e*) Number of lever presses in the last 5 min of the extinction test. Error bars, 1 s.e.m.

not receive any reward and, therefore, the value of time drops gradually during the test, predicting that animals should tend to revert to goal-directed control over time [12,15]. To test this prediction, we first over-trained animals and then gave them an extinction test as follows.

(i) Method

We trained 20 male Long-Evans rats to press a lever for 20% sucrose solution on a continuous reinforcement schedule for 3 days. Subsequently, rats received four sessions of 15 s random interval (RI15 s), followed by four sessions of random interval 30 s (RI30 s) and four sessions of random interval 60 s (RI60 s). Rats were given one session per day, but animals slow to acquire were given remedial sessions at the end of the day. Sessions terminated when 30 outcomes had been earned, or after 60 min, resulting in approximately 450 outcome presentations per animal by the end of training. The day after the last lever-press training session, the sucrose solution was devalued by conditioned taste aversion. All rats were given ad libitum access to sucrose solution for 30 min each day for 3 days. On each day, half of the animals received an intraperitoneal injection of lithium chloride (LiCl; 0.15 M; Sigma Aldrich, Castle Hill, New South Wales, Australia; 20 ml kg^{-1}); the other half received saline injections (20 ml kg^{-1}).

On the day following outcome devaluation, rats received a 10 min extinction test. Rats could respond on the lever, though no outcomes were delivered. The number of lever presses and magazine entries were recorded.

(ii) Results

There was a significant group × time (first versus second half of the test) interaction (repeated measure ANOVA; $F_{1,18} = 5.80$; $p = 0.026$), suggesting that outcome devaluation had a different effect in the first and the second half of the test. Responding during the first 5 min of extinction phase does not differ between groups (Welch's *t*-test; $t_{15.8} = 0.11$, $p = 0.91$); both the animals that had the food paired with gastric malaise (devalued) and those that did not receive the food-illness pairings (non-devalued) responded similarly in the first 5 min of extinction (figure 4*c*,*d*). However, by the last 5 min of the test, there was a significant difference in responding when the devalued and non-devalued groups were compared (Welch's *t*-test; $t_{11.56} = 2.81$, $p = 0.016$), suggesting that the goal-directed control of actions was restored in these rats (figure 4*c*,*e*). Data from the first half of the extinction test deviated from normal distribution significantly (Shapiro–Wilk normality test; $p = 0.008$) and thus a logarithmic transformation was applied to the data.

(iii) Discussion

Consistent with our prior prediction, and presumably as a result of the breaking down of the EM → LP and LP → EM action sequences, the rats started to show an increase in sensitivity to outcome devaluation over the course of extinction. This behavioural observation is consistent with a report showing that the pattern of neuronal activity, within dorsolateral striatum that marks the beginning and end of the action sequences during training, is diminished when the reward is removed during extinction [37]. It should be mentioned that this effect of extinction on the recovery of goal-directed actions can also be attributed to contextual change, or environmental volatility [38], derived from the absence of outcomes during the test. In regard to this factor, further modelling and experimental studies will be required to decouple the value of time, environmental statistics and context changes on goal-directed actions.

The formation of action sequences requires a set of actions to be executed in a fixed sequence. When the order of actions is variable, action sequences do not form [34]. The conditions depicted in figure 4*a*,*b* for instrumental conditioning are examples of fixed sequences of actions; the same lever needs to be pressed after a magazine entry response to earn the next outcome [15], allowing for the formation of an EM → LP action sequence. By contrast, when the action to be executed after EM is variable, for example, if the animal should sometimes press the left lever and sometimes the right lever to earn the outcome, then the EM response should not concatenate with the next action and the outcome will remain outside the action sequences. This account predicts, therefore, that when two different levers can be pressed to earn an outcome (making the response after EM variable), action sequences will not form and actions will remain sensitive to outcome devaluation even after extended training. This prediction is consistent with previous behavioural findings [7,39].

The differential sensitivity of one-lever and two-lever training to outcome devaluation both can also be addressed using the model-free account but based on a different logic: from the model-free perspective, the reason that habits are insensitive to outcome devaluation is that they are selected based on their reward history, and this history should not be altered by an 'offline' devaluation/revaluation treatment [11,12]. To account for why decisions remain goal-directed when two actions are available, the model-free account postulates that when an environment is non-stationary, and thus action values are subject to change, choosing between two closely valued actions (i.e. two levers in instrumental conditioning) requires persistent engagement of model-based processes in order to track which action is better at each point in time [12]. As a consequence, decisions remain goal-directed despite over-training. This explanation is based, therefore, on the assumption that certainty about the value of actions cannot exceed a certain threshold due to the environment being non-stationary. This is in contrast to the hierarchical account according to which the persistent engagement of goal-directed processes is due to sequence variability, and the same pattern of behaviour is predicted whether the environment is assumed to be stationary or not.

5. The neural bases of action sequences

Various models of learning and performance processes that mediate action sequences have been proposed (see [40,41] for a review). For example, we have recently advanced a normative model for learning action sequences [15] based on the open-loop property of action sequences and that implies, when two actions concatenate to form an action sequence, the second action will be executed irrespective of the outcome of the first. We previously demonstrated that the amount of reward loss due to this open-loop execution of actions is equal to the average of the prediction error generated by the second action [15]. When this reward loss is high, it implies that the second action should be chosen based on the specific outcome of the first action, and the actions do not concatenate. Conversely, a low reward loss implies safe open-loop execution allowing the action sequence to form.

It has been suggested that reward prediction errors are coded by midbrain dopamine neurons [42,43], and, as such, the above framework connects dopamine to action sequence formation. Consistent with this proposal, evidence suggests that the administration of a dopamine antagonist disrupts the chunking of movements into well-integrated sequences in capuchin monkeys [44], which can be reversed by co-administration of a dopamine agonist [45]. In addition, motor chunking appears not to occur in Parkinsons patients [46] due to a loss of dopaminergic activity in the sensorimotor putamen, which can be restored in patients on L-DOPA [47]. Similarly, insensitivity to outcome devaluation induced by over-training has been shown to depend on the ascending nigrostriatal dopamine pathway in rats [48], and the expression of NMDA receptors on dopamine neurons in mice [49].

Along similar lines, the hypothesis that insensitivity to devaluation is rooted in the formation of action sequences is consistent with a body of evidence demonstrating that a similar neural substrate mediates both habit development, as measured by outcome devaluation, and action sequence learning (for a review, see [10,15]). For example, lesion or inactivation of sensorimotor striatum restores sensitivity to outcome devaluation in over-trained animals [50,51]. Similarly, inactivation of sensorimotor striatum disrupts the expression of previously learned motor sequences [52] (but see [53]), and in humans, the blood-oxygen-level-dependent activity in sensorimotor putamen is correlated with the concatenation of action sequences [54]. Neural firing patterns recorded in the rat sensorimotor striatum have been reported to mark the start and end of action sequences in T-maze navigation [55], and sequences of lever presses [35,56]. Furthermore, it is reported that most of the striatal neurons that were more active during performance of a learned action sequence are in sensorimotor striatum, whereas neurons in the associative striatum more strongly responded to the performance of a new action sequence [57].

In summary, unlike the model-free account of habitual actions, which is silent with regard to the role of dopamine and the involvement of the sensorimotor striatum in action sequence learning, the hierarchical framework proposed here assigns a computational role to dopamine for learning action sequences, and is consistent with the shared neural structure between action sequence learning and insensitivity to outcome devaluation.

6. Conclusion

We have argued that insensitivity to outcome devaluation, one marker of habits, can occur as a by-product of hierarchical decision-making. This hypothetical link between habits and hierarchical decisions is readily distinguished from previous computational frameworks that view habits in terms of model-free decision-making. From the perspective of the current approach, however, the model-free account is not sufficient to explain automatic behaviour because it does not predict the existence of action sequences, which is inconsistent with recent data (see [14] for a comparison between these accounts). Although it can be extended to accommodate action sequences, this will render current model-free RL explanations of outcome devaluation theoretically redundant. This does not imply, however, that model-free RL is absent in instrumental conditioning, and further experimental work will be required to study whether a model-free valuation system coexists within the hierarchical framework of instrumental conditioning.

Funding statement. The research described here and the preparation of this paper was supported by a Laureate Fellowship from the Australian Research Council, no. FL0992409, and a grant from the National Institute of Mental Health, no. MH56446, to B.W.B.

References

1. Tolman EC. 1948 Cognitive maps in rats and men. *Psychol. Rev.* **55**, 189–208. (doi:10.1037/h0061626)
2. Adams CD. 1982 Variations in the sensitivity of instrumental responding to reinforcer devaluation. *Q. J. Exp. Psychol. B* **34**, 77–98. (doi:10.1080/14640748208400878)
3. Dickinson A. 1994 Instrumental conditioning. In *Animal cognition and learning* (ed. NJ Mackintosh), pp. 4–79. London, UK: Academic Press.
4. Dickinson A, Squire S, Varga Z, Smith JW. 1998 Omission learning after instrumental pretraining. *Q. J. Exp. Psychol. B* **51**, 271–286. (doi:10.1080/713932679)
5. Balleine BW, O'Doherty JP. 2010 Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology* **35**, 48–69. (doi:10.1038/npp.2009.131)
6. Colwill RM, Rescorla RA. 1984 Postconditioning devaluation of a reinforcer affects instrumental responding. *J. Exp. Psychol. Anim. Behav. Process.* **11**, 120–132. (doi:10.1037/0097-7403.11.1.120)
7. Kosaki Y, Dickinson A. 2010 Choice and contingency in the development of behavioral autonomy during instrumental conditioning. *J. Exp. Psychol. Anim. Behav. Process.* **36**, 334–432. (doi:10.1037/a0016887)
8. Lashley KS. 1951 The problem of serial order in behavior. In *Cerebral mechanisms in behavior* (ed. LA Jeffress), pp. 112–136. New York, NY: Wiley.
9. Book W. 1908 *The psychology of skill*. Missoula, MT: Montana Press.
10. Ashby FG, Turner BO, Horvitz JC. 2010 Cortical and basal ganglia contributions to habit learning and automaticity. *Trends Cogn. Sci.* **14**, 208–215. (doi:10.1016/j.tics.2010.02.001)
11. Daw ND, Niv Y, Dayan P. 2005 Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711. (doi:10.1038/nn1560)
12. Keramati M, Dezfouli A, Piray P. 2011 Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Comput. Biol.* **7**, e1002055. (doi:10.1371/journal.pcbi.1002055)
13. Botvinick MM, Niv Y, Barto AG. 2009 Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition* **113**, 262–280. (doi:10.1016/j.cognition.2008.08.011)
14. Dezfouli A, Balleine BW. 2013 Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized. *PLoS Comput. Biol.* **9**, e1003364. (doi:10.1371/journal.pcbi.1003364)
15. Dezfouli A, Balleine BW. 2012 Habits, action sequences and reinforcement learning. *Eur. J. Neurosci.* **35**, 1036–1051. (doi:10.1111/j.1460-9568.2012.08050.x)
16. Russell S, Norvig P. 1995 *Artificial intelligence: a modern approach*. Englewood Cliffs, NJ: Prentice-Hall.
17. Miller GA, Galanter E, Pribram KH. 1960 *Plans and the structure of behavior*. New York, NY: Holt, Rinehart and Winston.
18. Pew RW. 1966 Acquisition of hierarchical control over the temporal organization of a skill. *J. Exp. Psychol.* **71**, 764–771. (doi:10.1037/h0023100)
19. Newell A, Simon HA. 1963 GPS, a program that simulates human thought. In *Computers and thought* (eds EA Feigenbaum, J Feldman), pp. 279–293. New York, NY: McGraw-Hill.
20. Rosenbaum DA. 2009 *Human motor control*. Amsterdam, The Netherlands: Elsevier Science.
21. Abrahamse EL, Ruitenberg MFL, De Kleine E, Verwey WB. 2013 Control of automated behavior: insights from the discrete sequence production task. *Front. Hum. Neurosci.* **7**, 82. (doi:10.3389/fnhum.2013.00082)
22. Gobet F, Simon HA. 1998 Expert chess memory: revisiting the chunking hypothesis. *Memory* **6**, 225–255. (doi:10.1080/741942359)
23. Ericcson KA, Chase WG, Faloon S. 1980 Acquisition of a memory skill. *Science* **208**, 1181–1182. (doi:10.1126/science.7375930)
24. James W. 1890 *The principles of psychology*, vol 1. New York, NY: Holt.
25. Lee SW, Shimojo S, O'Doherty JP. 2014 Neural computations underlying arbitration between model-based and model-free learning. *Neuron* **81**, 687–699. (doi:10.1016/j.neuron.2013.11.028)
26. Reason J. 1990 *Human error*. Cambridge, UK: Cambridge University Press.
27. Norman DA. 1981 Categorization of action slips. *Psychol. Rev.* **88**, 1–15. (doi:10.1037/0033-295X.88.1.1)
28. Dietterich TG. 2000 Hierarchical reinforcement learning with the MAXQ value function decomposition. *J. Artif. Intell. Res.* **13**, 227–303.
29. Keele SW. 1968 Movement control in skilled motor performance. *Psychol. Bull.* **70**, 387–403. (doi:10.1037/h0026739)
30. Matsumoto N, Hanakawa T, Maki S, Graybiel AM, Kimura M. 1999 Nigrostriatal dopamine system in learning to perform sequential motor tasks in a predictive manner. *J. Neurophysiol.* **82**, 978–998.
31. Carr H, Watson JB. 1908 Orientation in the white rat. *J. Comp. Neurol. Psychol.* **18**, 27–44. (doi:10.1002/cne.920180103)
32. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. 2011 Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215. (doi:10.1016/j.neuron.2011.02.027)
33. Balleine BW. 2011 Sensation, incentive learning, and the motivational control of goal-directed action. In *Neurobiology of sensation and reward* (ed. JA Gottfried), pp. 287–310. Oxford, UK: Taylor and Francis.
34. Nissen MJ, Bullemer P. 1987 Attentional requirements of learning: evidence from performance measures. *Cogn. Psychol.* **19**, 1–32. (doi:10.1016/0010-0285(87)90002-8)
35. Jin X, Tecuapetla F, Costa RM. 2014 Basal ganglia subcircuits distinctively encode the parsing and concatenation of action sequences. *Nat. Neurosci.* **17**, 423–430. (doi:10.1038/nn.3632)
36. Niv Y, Daw ND, Joel D, Dayan P. 2007 Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology* (*Berlin*) **191**, 507–520. (doi:10.1007/s00213-006-0502-4)
37. Barnes TD, Kubota Y, Hu D, Jin DZ, Graybiel AM. 2005 Activity of striatal neurons reflects dynamic encoding and recoding of procedural memories. *Nature* **437**, 1158–1161. (doi:10.1038/nature04053)
38. Simon DA, Daw ND. 2011 Environmental statistics and the trade-off between model-based and TD learning in humans. In *Advances in neural information processing systems* (eds J Shawe-Taylor, RS Zemel, PL Bartlett, F Pereira, KQ Weinberger), pp. 127–135. Red Hook, NY: Curran Associates Inc.
39. Colwill RM, Rescorla RA. 1988 The role of response-reinforcer associations increases throughout extended instrumental training. *Anim. Learn. Behav* **16**, 105–111. (doi:10.3758/BF03209051)
40. Rhodes BJ, Bullock D, Verwey WB, Averbeck BB, Page MPA. 2004 Learning and production of movement sequences: behavioral, neurophysiological, and modeling perspectives. *Hum. Mov. Sci.* **23**, 699–746. (doi:10.1016/j.humov.2004.10.008)
41. Grossberg S, Pearson LR. 2008 Laminar cortical dynamics of cognitive and motor working memory, sequence learning and performance: toward a unified theory of how the cerebral cortex works. *Psychol. Rev.* **115**, 677–732. (doi:10.1037/a0012618)

42. Montague PR, Dayan P, Sejnowski TJ. 1996 A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16**, 1936–1947.

43. Sutton RS, Barto AG. 1998 *Reinforcement learning: an introduction*. Cambridge, MA: MIT Press.

44. Levesque M, Bedard MA, Courtemanche R, Tremblay PL, Scherzer P, Blanchet PJ. 2007 Raclopride-induced motor consolidation impairment in primates: role of the dopamine type-2 receptor in movement chunking into integrated sequences. *Exp. Brain. Res.* **182**, 499–508. (doi:10.1007/s00221-007-1010-4)

45. Tremblay P-L, Bedard M-A, Levesque M, Chebli M, Parent M, Courtemanche R, Blanchet PJ. 2009 Motor sequence learning in primate: role of the D2 receptor in movement chunking during consolidation. *Behav. Brain Res.* **198**, 231–239. (doi:10.1016/j.bbr.2008.11.002)

46. Benecke R, Rothwell JC, Dick JP, Day BL, Marsden CD. 1987 Disturbance of sequential movements in patients with Parkinson's disease. *Brain* **110**, 361–379. (doi:10.1093/brain/110.2.361)

47. Tremblay P-L, Bedard M-A, Langlois D, Blanchet PJ, Lemay M, Parent M. 2010 Movement chunking during sequence learning is a dopamine-dependant process: a study conducted in Parkinson's disease. *Exp. Brain Res.* **205**, 375–385. (doi:10.1007/s00221-010-2372-6)

48. 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. (doi:10.1523/JNEUROSCI.3894-04.2005)

49. Wang LP, Li F, Wang D, Xie K, Wang D, Shen X, Tsien JZ. 2011 NMDA receptors in dopaminergic neurons are crucial for habit learning. *Neuron* **72**, 1055–1066. (doi:10.1016/j.neuron.2011.10.019)

50. Yin HH, Knowlton BJ, Balleine BW. 2004 Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *Eur. J. Neurosci.* **19**, 181–189. (doi:10.1111/j.1460-9568.2004.03095.x)

51. Yin HH, Knowlton BJ, Balleine BW. 2006 Inactivation of dorsolateral striatum enhances sensitivity to changes in the action-outcome contingency in instrumental conditioning. *Behav. Brain Res.* **166**, 189–196. (doi:10.1016/j.bbr.2005.07.012)

52. Miyachi S, Hikosaka O, Miyashita K, Kárádi Z, Rand MK. 1997 Differential roles of monkey striatum in learning of sequential hand movement. *Exp. Brain Res.* **115**, 1–5. (doi:10.1007/PL00005669)

53. Desmurget M, Turner RS. 2010 Motor sequences and the basal ganglia: kinematics, not habits. *J. Neurosci.* **30**, 7685–7690. (doi:10.1523/JNEUROSCI.0163-10.2010)

54. Wymbs NF, Bassett DS, Mucha PJ, Porter MA, Grafton ST. 2012 Differential recruitment of the sensorimotor putamen and frontoparietal cortex during motor chunking in humans. *Neuron* **74**, 936–946. (doi:10.1016/j.neuron.2012.03.038)

55. Thorn CA, Atallah H, Howe M, Graybiel AM. 2010 Differential dynamics of activity changes in dorsolateral and dorsomedial striatal loops during learning. *Neuron* **66**, 781–795. (doi:10.1016/j.neuron.2010.04.036)

56. Jin X, Costa RM. 2010 Start/stop signals emerge in nigrostriatal circuits during sequence learning. *Nature* **466**, 457–462. (doi:10.1038/nature09263)

57. 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. (doi:10.1007/s00221-002-1213-7)