The Role of Task Representation in Reinforcement Learning Strategies

Thanaphat Thongpaibool

Msc Brain and Cognitive Sciences, University of Amsterdam, The Netherlands

SUMMARY

In reinforcement learning (RL) framework, abundant neural and behavioral evidence suggest that the brain uses multiple, distinct learning strategies. One is computationally characterized as model-free which simply learns reward associations. Another is model-based strategy which learns and takes into account the interactions between rewards and state transitions. In human, many RL research attempting to understand how the brain arbitrates between these two strategies use an abstract two-stage Markov decision task. Ample evidence from different experimental paradigms (e.g. spatial navigation) have provided some clues that task representation of RL problems might play a crucial role in the extent to which strategy is dominant. Here, we investigated whether embedding a basic spatial feature to a classical, temporal two-stage Markov task could affect learning about reward and state transition contingencies and potentially influence the arbitration between learning strategies. We found that spatial representation significantly improves state transition learning and, under certain circumstances, prompts more reliance on model-based strategy.

INTRODUCTION

The most important premise in reinforcement learning (RL) theory is that agents, including humans and animals, can learn to maximize reward and minimize punishment by using past experiences to account for future decision-making (Sutton & Barto, 1998; Wunderlich, Smittenaar, & Dolan, 2012). The long-standing view in both the fields of psychology and neuroscience is that there are multiple, distinct learning strategies that the brain can use in learning and guiding decisions (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Daw & O'Doherty, 2014). In an influential theoretical framework, a computational dissociation has been drawn between model-based and model-free strategies (Daw et al., 2011; Daw & O'Doherty, 2014; Dayan & Niv, 2008; Doll, Simon, & Daw, 2012; Gläscher, Daw, Dayan, & O'Doherty, 2010).

In model-free strategy, an agent learns action values, by trial and error, reinforcing previously successful actions (Daw et al., 2011; Daw & O'Doherty, 2014). The action values are then computed to make decisions. In other words, an agent learns from past experiences directly through reinforcement to select future actions. Such learning strategy can be used to explain basic animal learning behavior—for example, the prominent behaviorist principle, "law of effect," which has long been studied in animal research (Thorndike, 1933). This principle states that responses that are followed by a satisfying effect (e.g. food pellets) will likely be chosen in the future, while responses that are followed by a discomforting effect (e.g. electric shock) will less likely be repeated. This behavioral pattern resulting from model-free strategy has been observed for decades in animal instrumental training paradigm (see Daw & O'Doherty, 2014). For example, food-deprived mice that are extensively trained to press a lever in order to get food will continue to press the lever with an expectation of obtaining food. In this way, psychologically, model-free strategy is related to Thorndike's law of effect, since by utilizing model-free strategy previously successful actions tend to be repeated.

In principle, an important mechanism of model-free learning strategy is value updating, via reward prediction error which signals the discrepancy between what is expected and what actually happens (Schultz, Dayan, & Montague, 1997; Wunderlich et al., 2012). In neuroscience, many human neuroimaging and electrophysiological studies have linked prediction error signals to dopaminergic system (e.g. Berns, McClure, Pagnoni, & Montague, 2001; Hare, O'Doherty, Camerer, Schultz, & Rangel, 2008; Pagnoni, Zink, Montague, & Berns, 2002; Schultz et al., 1997). Dopamine precursor (e.g. L-DOPA) has also been used to enhance reward prediction error signals in striatum, which in turn increased the likelihood of model-free choice patterns in human instrumental learning paradigm (Pessiglione, Seymour, Flandin, Dolan, & Frith, 2006).

Nonetheless, model-free learning strategy, which is driven by a simple stimulus-response mechanism, cannot account for all behavioral repertoires—for example, adaptive behaviors in absence of reinforcement such as latent learning (see Thistlethwaite, 1951). Likewise, some behavioral responses observed in outcome devaluation paradigms (Daw, Niv, & Dayan, 2005) and more complex behaviors seem to require novel and/or more flexible decision planning (see Daw & O'Doherty, 2014; Simon & Daw, 2011b). In a latent learning paradigm, it has been demonstrated that rodents were able to learn the structure of the mazes dispensable of reinforcement (e.g. Tolman, 1948). When, later, rewards were introduced, animals were able to plan their actions according to the learned geographical structure of the maze to obtain the

rewards. Furthermore, they could choose new routes through a maze that similarly led to the reward even if previously successful pathways were disrupted or when new routes were available. In such novel situations, if rodents only rely on a stimulus-response mechanism, they would not be able to successfully obtain the reward. Instead, they would keep trying to follow the previously successful pathways since future actions could be drawn only from learned associations between rewards and past experiences. Therefore, such flexible behaviors cannot be explained by a simple stimulus-response mechanism.

To account for such sophistication in behaviors, the concept of “cognitive map” was proposed (Tolman, 1948). This concept suggests that agents form an internal representation of environment contingencies—a so-called “world model”—and use this model to consider the future consequences of a decision, resulting in more flexible and complex behaviors. Such advanced learning mechanism is computationally defined as model-based strategy (Daw et al., 2011; Daw & O’Doherty, 2014). In accordance with Tolman’s idea, a model-based agent learns about the task by internally forming a cognitive map, which draws the sequential connections between different states (namely, situations) of the environment. This learned map of the state transition structure, independent of immediate rewards, can then be used to prospectively (or in a forward-planning manner) guide actions that implement reward maximizing policy and actively evaluate which actions are more beneficial. This type of action value computation is analogous to the mental simulation when playing chess (Gläscher et al., 2010). Put it differently, a model-based agent needs to explicitly learn the state transition structure to estimate future available rewards, while a model-free agent needs not, since it instead learns about reward consequences directly from past experience to evaluate and update action values (Doll et al., 2012).

In recent years, much research attention has been focused on investigating the different properties of the two aforementioned learning strategies as well as how agents may implement and integrate these strategies to make better decisions (e.g Gläscher et al., 2010). In humans, most studies (e.g. Daw et al., 2011; Gläscher et al., 2010; Wunderlich et al., 2012) have used a probabilistic and sequential Markov decision task, which usually entails two successive stages of decision choice (or a state), followed by an outcome or a reward (see Figure 1A for an example). In this task, at each stage, subjects have to choose between two options, which are typically abstract symbols or fractal images. On the one hand, the task is probabilistic in a sense that an outcome of a specific choice within a state (consisting of two available options) may be obtained under certain probability. On the other hand, it is sequential in that a first-stage choice

leads predominantly to a particular state of a second-stage decision. The distinctions between two learning strategies in executing decisions have been studied by analyzing subjects' choice behavior. In a two-stage Markov decision task, although utilizing model-based strategy seems to be more optimal for reward maximization, numerous evidence illustrated that, on average, humans exhibit a mixture of both learning strategies (e.g. Daw et al., 2011; Wunderlich et al., 2012).

Much evidence has shown that the extent to which learning strategy is more dominant can be internally manipulated. One study used transcranial magnetic stimulation (TMS) to disrupt right dorsolateral prefrontal cortex while subjects were engaging in a two-stage Markov task (Smittenaar, FitzGerald, Romei, Wright, & Dolan, 2013). The results showed that the TMS disruption drove behavior more towards model-free strategy. Another study demonstrated that by depriving cognitive resource the reliance on model-free strategy was increased (Otto, Gershman, Markman, & Daw, 2013). Furthermore, the effect of dopamine administration was also reported to influence subjects' reinforcement learning strategies, promoting model-based over model-free strategy (Wunderlich et al., 2012). The implication of these studies is that manipulating the arbitration between two reinforcement learning strategies is in fact possible. One strategy may be arbitrated more dominantly than the other under different circumstances.

Alternatively, some researchers have recently argued that the representation of reinforcement learning problems might be critical for understanding the balance between two learning strategies (Botvinick, Weinstein, Solway, & Barto, 2015). Representation entails how subjects perceive the states of environment and the set of feasible actions (Rangel, Camerer, & Montague, 2008). Although the mechanistic implementation of both learning strategies is fundamental, how representation may be internally coded in the brain and involve in reinforcement learning process remains relatively unexplored in human (Botvinick et al., 2015; Rangel et al., 2008). Most classical two-stage Markov tasks (e.g. Daw et al., 2011) that have been used in human reinforcement learning research community only explore a temporal structure of task representation with abstract cues (e.g. abstract symbols or fractal images) of states and state transitions. Although there are some variations in the task's outcome probability of different studies, the spatial structure of the task's representations (e.g. the positions of states and abstract cues on the screen) is relatively the same. While the combination of both learning strategies was shown in a two-stage Markov task, research that used a relatively simpler experimental task—which required subjects to learn the relationship between a particular set of actions and associated outcomes—reported more model-free driven behavior (e.g. Palminteri,

Khamassi, Joffily, & Coricelli, 2015; Pessiglione et al., 2006). This seems to suggest that an inherent structure of the task may play a role in the arbitration between learning strategies.

In fact, similar trend of behavioral and neural evidence demonstrating one dominant learning strategy over the other was reported in different reinforcement learning tasks, such as a spatial navigation task and a sequential finger movement task. In a spatial navigation task, subjects were forced to continuously learn about the configuration of the maze (Simon & Daw, 2011b). The results showed that the choice behavior in such spatial decision task was better explained by model-based strategy. A novel sequential finger movement task also supplied similar evidence of model-based driven behavior (Fermin, Yoshida, Ito, Yoshimoto, & Doya, 2010). In this task, subjects learned to map different buttons to movement directions and, in doing so, selected sequential action based on the learned internal model of state transition. The behavioral results potentially connected back to the concept of Tolman's cognitive map illustrating that subjects could form a world model (e.g. a maze), learn about the structure of perceived states of the environment and state transition, and rapidly evaluate their actions in a forward-planning manner. Interestingly, the shared feature of the two aforementioned tasks is spatiality in their representational structure. This provides evidence that task representation *per se* might indeed dictate the arbitration between different learning strategies.

The fact that learning strategies can be internally arbitrated in the brain (e.g. Smittenaar et al., 2013) and the converging evidence that different learning behaviors may inherently result from the external features of the task' representations (e.g. Simon & Daw, 2011b) question the optimality of a classical Markov decision task currently used in RL research community to investigate human learning process—especially the distinction between model-free and model-based strategies. This is essential because task representation *per se* may potentially define the selection of learning strategies and the possible course of resulting behaviors (Rangel et al., 2008). Nonetheless, fundamentally how task representation (e.g. states and state transitions) affects reinforcement learning and influences the arbitration between learning strategies remain unclear. The current study attempted to answer these issues by investigating how embedding a spatial structure—i.e. a specific location of states—to a classical temporal two-stage Markov task could affect learning about reward and state transition contingencies and potentially influence the arbitration between learning strategies, which can be assessed by observing choice behavior. We hypothesized that a spatial feature would allow subjects to create a spatial mapping of state transitions, similar to the idea of cognitive map, which in turn improves state transition learning. Consequently, embedding a spatial structure to a temporal task would trigger

a more prospective, forward-planning action for reward maximization by drawing on a learned map of state transition structure—in the same way as model-based learning strategy.

EXPERIMENTAL PROCEDURES

Subjects

Twelve healthy subjects (7 females, Age: *M* = 23.18, *SD* = 2.96, one 50-year-old male subject was excluded from age calculation) were recruited to participate in this study. All subjects gave the informed consent before the experiment, and after completing the experiment all were debriefed and received the payoff according to their performance. The study was approved by The Ethics Committee of the Faculty of Social and Behavioral Sciences, the University of Amsterdam.

Behavioral task

To test whether an incremental difference in task's representations affect the arbitration between learning strategies, we designed two two-stage Markov decision tasks—named "Temporal" (*T*) and "Spatial" (*S*)—involving two sequential decision choices, one at each stage between the two options, followed by a rewarded outcome (Figure 1A and 1B). In each stage, there were two possible color states (e.g. red and blue), each with a certain set of two abstract Agathodaimon symbols. In other words, one state consisted of two symbols, a color boundary, and a cross sign. At each stage, a certain color state appeared on the screen and subjects were asked to choose between one of the two different symbols, by pressing either the right or left button at self-paced. The only difference between the two tasks (i.e. Temporal and Spatial) was the position of states. For the Temporal task, like classical two-stage Markov decision tasks (e.g. Daw et al., 2011; Gläscher et al., 2010), color states always situated in the middle of the computer screen (Figure 1A). For the Spatial task, each of the four color states would appear at a fixed position on one of the four corners of the screen throughout the session (Figure 1B). The first-stage states were on either side of the upper part of the screen; the second-stage states were on either side of the bottom part of the screen. This creates a maze-like structure of states and state transitions—starting at the top and finishing at the bottom with the best outcome.

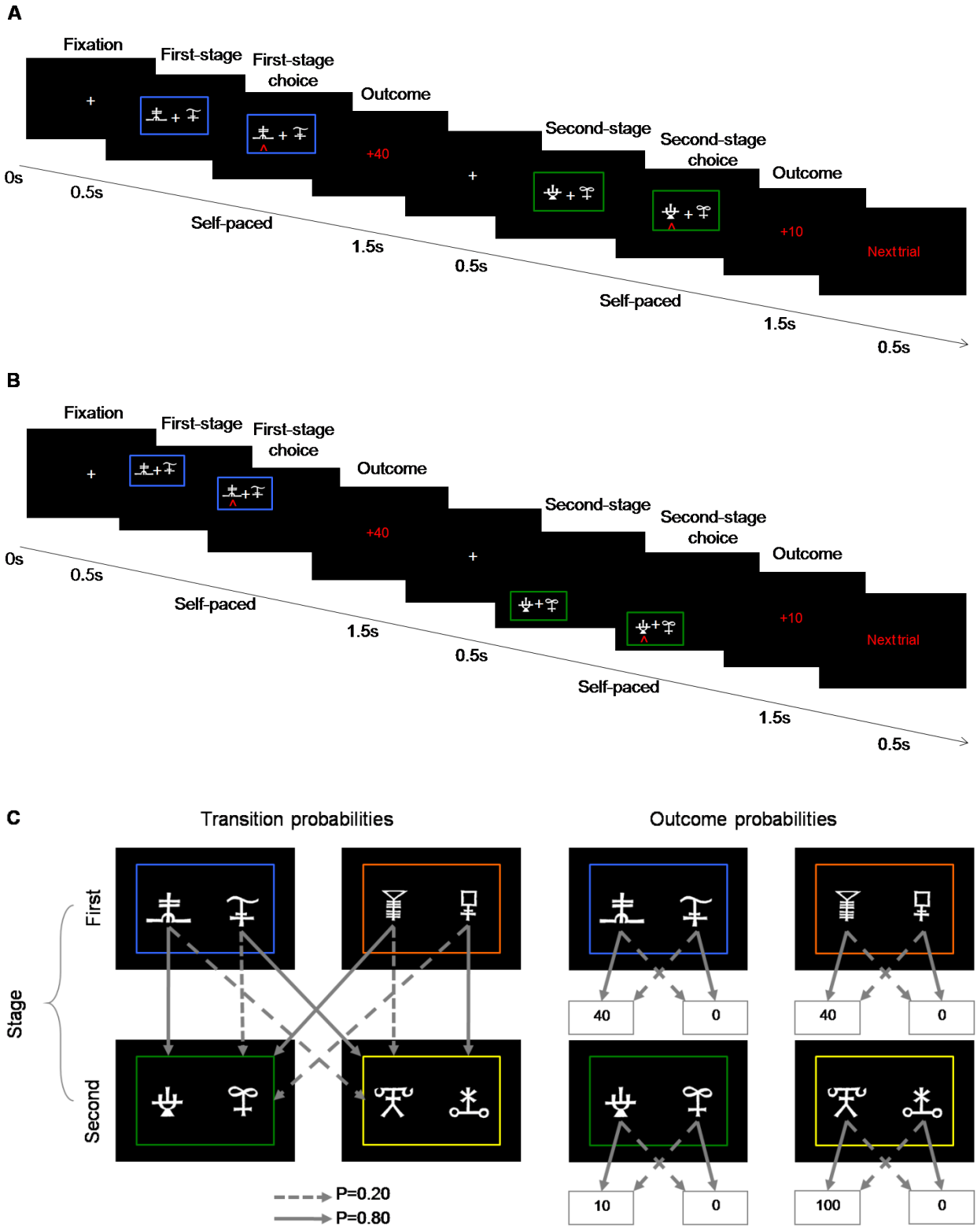


Figure 1. Task design A timeline of events in a single trial for (A) the Temporal task and (B) the Spatial task. In the current study, the experimental task is a two-stage Markov decision task in

which all decision states are represented by a color boundary and two abstract symbols. Within each stage, one of two possible color states appears and subjects have to choose between two symbols in each state. (C) Structure of the outcome and state transition probabilities. With certain probability (0.8/0.2), one particular symbol gives a certain reward (0, 40, or 100 points) and leads predominantly to one second-stage state (e.g. in this case either green or yellow state).

The structure of outcome probability and state transition probability is depicted in Figure 1C. Critically, within the same state, one of the two symbols was on average more rewarded. For both first-stage states, one of the two symbols delivered 40 points at the probability of 0.8. In contrast, for the second-stage decision, one of the two symbols in a certain state gave 10 points, while one symbol in another state yielded 100 points, also at the probability of 0.8. Each symbol in the first stage was associated predominantly with a certain second-stage state, and led there 80 percent of the time. Specifically, the first-stage 40-point symbols had 80 percent chance of transitioning to a second-stage state with a 10-point symbol, while the first-stage 0-point symbols had 80 percent chance of leading to a second-stage state with a 100-point symbol. The first-stage 40-point reward, purposely, creates a tension between obtaining an immediate reward and a possibility to plan forward and acquire a higher reward, i.e. 100 points. The relationships between symbols, color states, and stages were fixed throughout a session but were randomized for different sessions. Importantly, the optimal maximization strategy for our tasks was to try to obtain 100 points, in every trial (see the calculation of expected value in Supplemental Information (SI)).

The main experiment consisted of 4 successive sessions—two consecutive sessions of each task, each with different sets of symbols. We referred to each session of one task as *attempt* (i.e. first and second). For each subject, the order of the tasks was either two sessions of the Temporal task followed by two sessions of the Spatial task (*TS*) or two sessions of the Spatial task followed by two sessions of the Temporal task (*ST*). The order was counter-balanced across subjects. Each session comprised of 80 two-stage choices trials (i.e. 160 choices per session) and 5 extra forced-choice post-test questions which assessed subjects' knowledge of the learning task' s features and probabilistic contingencies (see Table 1). In one session, the assignment of two different symbols to states was randomized but each pair of symbols was always matched with a certain color state throughout the session. The stage to which each color

state belonged to was also randomized. Note that for the Spatial task the position of certain color states varied from session to session.

Table 1. Post-test questions to test subjects' acquired explicit knowledge of task factors

Names	Post-test questions
"Symbol-State Associations"	In which state does this symbol belong to?
"State-Stage Associations"	In which stage does this color state belong to?
"State Transitions"	Which second state does this symbol mostly lead to?
"Reward Contingencies"	Which outcome does this symbol mostly lead to?
"Preferred Strategy"	Which of these two symbols is more advantageous from your experience?

Note. To distinguish subjects' preferred learning strategy, we reasoned that model-free subjects would indicate that 40-point symbols were more advantageous, while model-based subjects would rate 0-point symbols as more advantageous. However, we decided to exclude the post-test question 1 from the main analysis because, from both pilot study and main experiment, most subjects interpreted 'more advantageous' as more rewarded (i.e. more points), and this interpretation was not well matched with their learning strategy, when examined concurrently with their choice behaviors. On top of that, during the informal debriefing, even though the verbal description of reward maximization strategy resembled more model-based learning strategy, most subjects chose 40-point symbols to be more advantageous than 0-point symbols. Therefore, unfortunately, the Preferred Strategy question failed to capture individual difference in reward maximization strategy, and was excluded from the analysis.

Before the beginning of the main experiment, there was a short training session, which consisted of 10 two-stage choices trials (i.e. 20 choices) of the Temporal task with a different set of symbols (Thai alphabets). Subjects' goal was to maximize the number of points (see the Experimental Instruction in SI). In the end, 1600 points were converted to 1 Euros, and subjects' earnings would correspond to the total points earned over four sessions.

Behavioral analyses

The extent to which model-based and model-free strategies can be separated depends on details of the task and hypotheses on how the two strategies interact with each other to implement behavior (O'Doherty, Lee, & McNamee, 2015). In our experimental paradigm, the main questions are whether task representation has an effect on state transition learning and on the arbitration between model-free and model-based strategies. To dissociate learning strategies, we considered subjects' first-stage choice behavior. The logic of the task was that reliance on model-based or model-free learning strategies would produce different behavioral patterns. Model-free strategy predicts that the first-stage 40-point symbols are likely to be chosen regardless of the second-stage state the symbols mostly lead to, which for 80 percent of the time is the state containing a 10-point symbol. This is because model-free agents do not consider the structure of the task but rather learn about reward consequences from previous trials; hence, immediate rewarded choices are more likely to be repeated without a consideration of state transitions. In contrast, model-based strategy predicts that the propensity of choosing the first-stage 0-point symbols, instead of 40-point symbols, will be high, since the 0-point symbols substantially lead to the state with a 100-point symbol, which in the long run is considerably more rewarded. Under model-based strategy, subjects learn the structure of state transitions and use the mapping of states to prospectively plan and make choices. In this case, choosing 0-point symbols indicates a more forward-planning, model-based action. Therefore, to analyze subjects' choice behavior, we computed the percentage of 0-point symbols being chosen as dependent variable.

Our main analyses concerned the effect of task representation (i.e. embedded spatial feature versus classical temporal structure) on the arbitration of learning strategies and on learning about reward and state transition contingencies. We first assessed the learning rate by extracting the average percentage of 0-point symbols being chosen over every 8 trials (i.e. 10 time points) for each task. Using a 2 (task representation: Temporal versus Spatial) x 10 (time point: 1^{st}-10^{th} time points) repeated-measures analysis of variance (ANOVA), the difference in learning rates between the two tasks was examined by looking whether there was an interaction between task representations and time points. Quantitatively, increased propensity to choose 0-point symbols signifies a transition to model-based strategy, whereas decreased tendency of choosing 0-point symbols indicates more reliance on model-free strategy. Furthermore, because of non-stochastic nature of our tasks' reward and state transition probabilistic structures, our assumption was that the difference in subjects' choice behaviors between two learning

strategies would be more pronounced in the second half of the session (41st to 80th trial), converging to one dominant learning strategy. From this, we computed the percentages of 0-point symbols being chosen in the second half of the session and entered the values as a dependent variable into two-way repeated-measures ANOVA with the task representation (T versus S) and attempt (first versus second) as within-subject factors. To assess subject's acquired task knowledge, we used a paired-wise t-test on the percentage of correct answers to all post-test questions (with the exception of Preferred Strategy, see above) to compare the effect of task representation. In addition, a one-tailed one-sample t-test was conducted to check whether subjects' knowledge was above the 50-percent chance level.

RESULTS

Rate of learning

Figure 2A illustrates differences in the learning rates between the Temporal and Spatial tasks. A repeated-measures ANOVA was used to analyze the learning rate. Mauchly's test indicated that the assumption of sphericity in a task representation x time points interaction had been violated ($p < 0.002$), therefore degree of freedom was corrected using Greenhouse-Giesser estimates of sphericity ($\varepsilon = 0.279$). The analysis revealed significant main effects of task representation, $F(1,11) = 6.66$, $p = 0.026$, suggesting different levels of arbitration between two learning strategies for the Temporal and Spatial tasks, and main effect of time points, $F(9,99) = 2.92$, $p = 0.004$, showing evidence of a learning process during the tasks. However, there was no interaction between task representations and time points, $F(2.51,27.62) = 2.92$, $p = 0.477$, suggesting that there was no difference between the learning rates of the two tasks. Quantitatively, the results from both tasks supported evidence that a combination of both model-free and model-based learning strategies contribute to human choice behavior (choosing both 40-point and 0-point symbols). Interestingly, our results demonstrated higher reliance on model-free in general as the percentage of choosing 0-point symbols was relatively below fifty percent for both tasks. Nonetheless, dependency on model-based strategy seemed to be more eminent in the Spatial task, as compared to the Temporal task.

Choice behavior in the second half of the session

Assuming that subjects learned about symbol-reward associations and state transition structure prominently during the first half of the session, we speculated that in the second half the effect of task representation would elicit more supremacy of model-based strategy in the Spatial task. To our surprise, the analysis revealed only a trend for the effect of task representation, $F(1,11) = 3.77$, $p = 0.078$, implying that the spatial feature might not be a sole factor triggering model-based learning strategy (Figure 2B). There was also neither main effect of attempt, $F(1,11) = 2.85$, $p = 0.122$, nor interaction between task represenation x attempt, $F(1,11) = 0.16$, $p = 0.695$, indicating no repetition effect on choice behavior.

Subjects' knowledge of task's features and contingencies

We found significant evidence for better state transition learning in the Spatial task ($M = 73.96$, $SEM = 5.21$), compared to the Temporal task ($M = 56.25$, $SEM = 4.74$); $t(11) = -4.93$, $p < 0.001$ (Figure 2C). There was no significant effect of task representation on learning about symbol-state associations, state-stage associations, and reward contingencies (see Table 2). Subjects' percentage of correct answers was also higher than 50-percent chance level for Symbol-State associations, State-Stage associations, and Reward Contingencies post-test questions (see Table 3). The results indicated that subjects successfully learned the associations between task's features (i.e. relationships between symbols, states, and stages) and symbol-reward contingencies. Interestingly, the percentage of correct answers for State Transitions was significantly higher than chance level only in the Spatial task, $t(11) = 4.60$, $p < 0.001$, but not in the Temporal task, $t(11) = 1.32$, $p < 0.107$, suggesting that subjects successfully learned state transition structure only in the Spatial task.

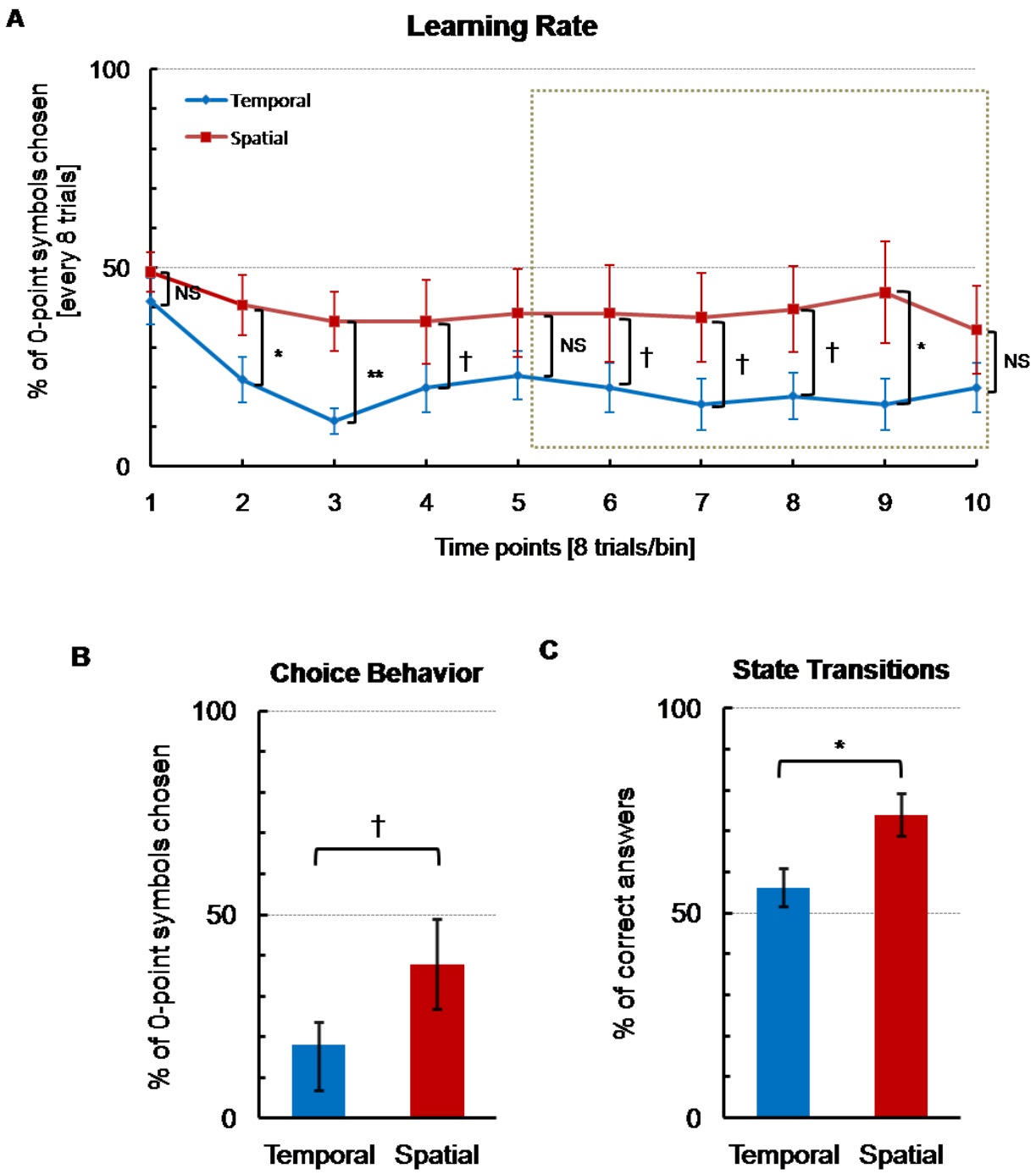


Figure 2. Results comparing between the Temporal and Spatial tasks (A) Percentage of 0-point symbols being chosen for each task on average of every eight trials throughout the session. A paired-wise t-test was conducted on each data point (see Table S1 in Supplemental Information). (B) Average percentage of 0-point symbols being chosen for each task in the second half of the session. (C) Average percentage of correct answers to State Transitions

question. Error bars represent intersubject SEM. Note. † = p < .10, *= p < .05, **= p < .01, ***= p < .001. *N* = 12.

Table 2. Statistical comparison of the percentage of correct post-test answers between the Temporal and Spatial tasks

Post-test questions	Task		*t*(11)	p
	Spatial	Temporal		
Symbol-State Associations	65.63 ± 7.24	76.56 ± 4.22	-1.61	0.136
State-Stage Associations	77.08 ± 5.29	69.79 ± 6.05	0.86	0.41
State Transitions	73.96 ± 5.21	56.25 ± 4.74	4.93	< 0.001***
Reward Contingencies	80.73 ± 4.52	83.33 ± 3.38	-0.77	0.46

Note. Data are expressed as mean ± SEM. Paired-wise t-test, † = p < .10, *= p < .05, **= p < .01, ***= p < .001. *N* = 12.

Table 3. Statistical analysis of the percentage of correct post-test answers above 50-percent chance level

Post-test questions	Spatial		Temporal	
	t(11)	p	*t*(11)	p
Symbol-State Associations	2.16	0.027*	6.29	< 0.001***
State-Stage Associations	5.12	< 0.001***	3.27	0.004**
State Transitions	4.6	< 0.001***	1.32	0.107
Reward Contingencies	6.80	< 0.001***	9.85	< 0.001***

Note. One-tailed, one-sample t-test, † = p < .10, *= p < .05, **= p < .01, ***= p < .001. *N* = 12.

The effect of task order on choice behavior

We conducted further tests to examine other possible factors, in addition to task representation, that could influence the arbritation of learning strategies, which may help us improve task design

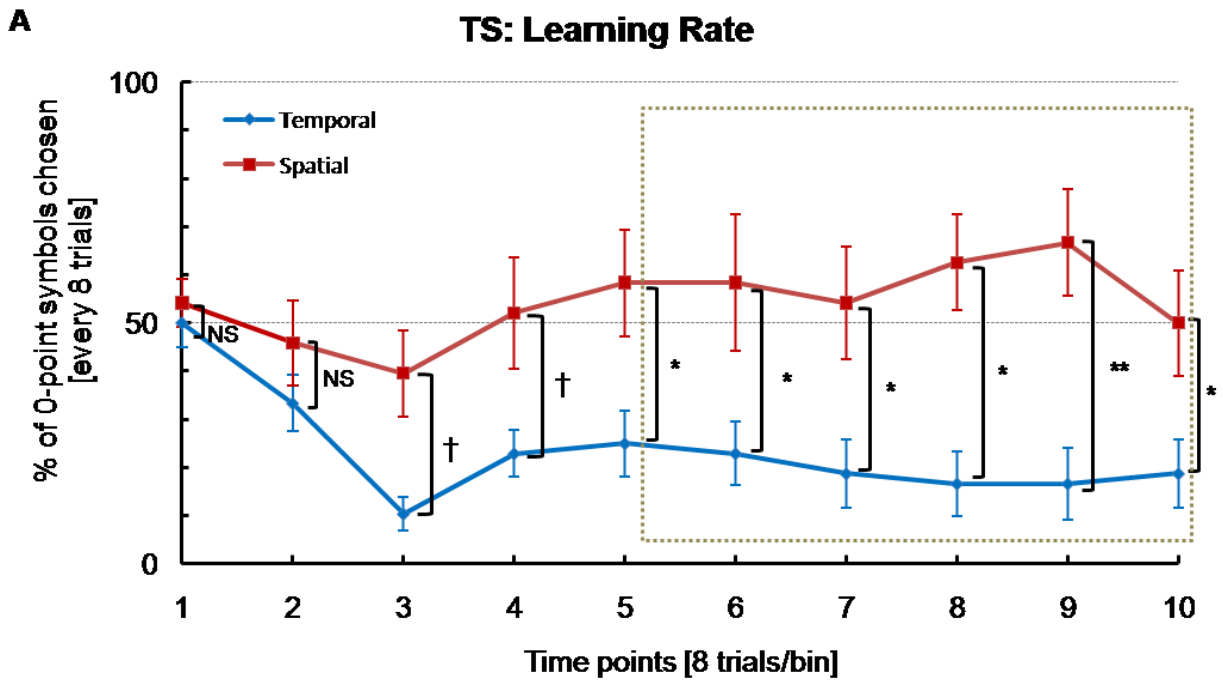
in future studies. In comparison with previous works that reported more influence of model-based strategy (Daw et al., 2011; Gläscher et al., 2010), our tasks had a relatively small number of trials for each session and subjects were not pre-trained to understand task contingencies. From this, we suspected that in general subjects' propensity to switch to model-based strategy might also be depedent on their experience with the tasks, which may result from the effect of task order. If so, there should be an interaction between task order (i.e. TS and ST) and task representation. Accordingly, we conducted a mixed ANOVA on the percentage of choosing 0-point symbols in the second half of the session with task representation and attempt as within-subject factors and, in addition, task order as a between-subject factor (Table S2). Critically, having also accounted for order conditions, we observed a significant main effect of task presentation, $F(1,10) = 5.00$, $p = 0.049$, and a trend for a task representation x task order interaction, $F(1,10) = 4.59$, $p = 0.058$, implying that the percentage of choosing 0-point symbols between the two tasks might be different in the two order conditions. We did not observe main effect of order, $F(1,10) = 2.38$, $p = 0.154$, meaning no direct influence of order conditions on learning strategy.

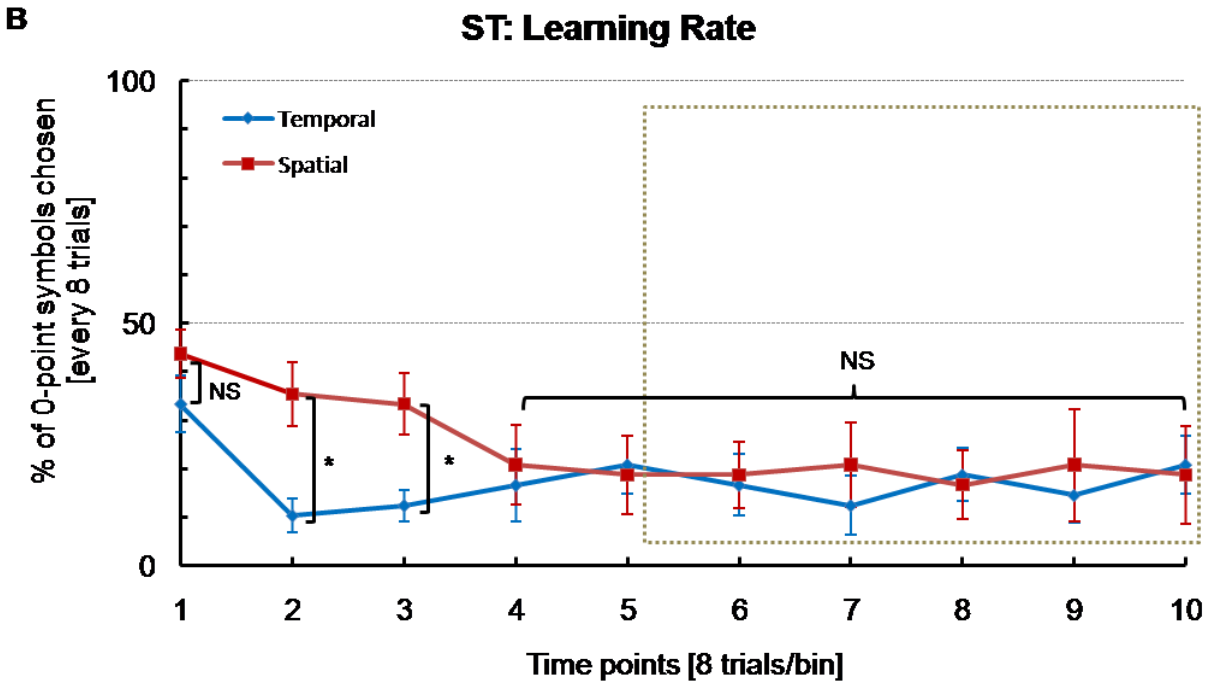
If experience with the tasks actually interplays with influence on model-based strategy, a shift from model-free to model-based strategy should be more notable in the TS than in the ST condition, since in the TS condition the Spatial task was performed at later sessions. Consequently, we extracted the data according to the order of the tasks (i.e. TS and ST) and performed a t-test on choice behavior between the two tasks in each condition. Indeed, in the Spatial task, there was a significant shift towards more model-based strategy in the TS condition (Figure 3C; $t(5) = -3.98$, $p = 0.011$), but not the ST condition (Figure 3D; $t(5) = -0.06$, $p = 0.957$). Quantitatively, reliance on model-free strategy was prevalent in the Temporal task for both conditions; only in the TS condition model-based strategy was more dominant during the Spatial task (Figure 3C and 3D). Furthermore, we also tested whether pure learning experience impacts a tendency to rely on model-based strategy. We sorted the percentage of choosing 0-point symbols by sessions and extracted them according to two order conditions. Using one-way repeated-measures ANOVA with sessions as an independent factor, the main effect of session did not reach a significance level for both conditions (TS: $F(3,20) = 2.80$, $p = 0.067$; ST: $F(3,20) = 0.030$, $p = 0.993$), showing that pure learning experience was not the main factor for more reliance on model-based strategy. The results seem to suggest that the interaction between spatial representation and task experience are necessary for subjects to adapt model-based strategy at later sessions. Given a small sample size, more behavioral data has to be collected

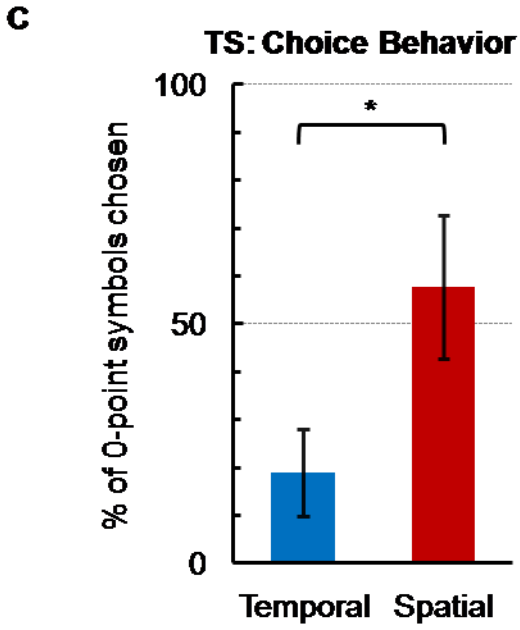
to confirm th effect of spatial representation and its interaction with task experience on the arbitration between model-free and model-based strategies.

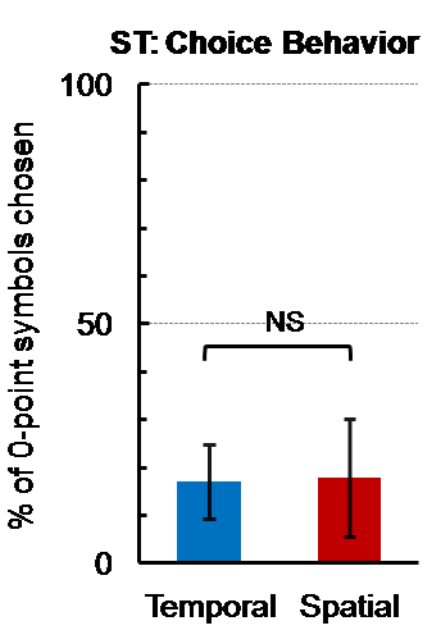
The effect of task order on subjects' knowledge of task's state transition structure

In addition to choice behavior, we also tested whether state transition learning was different in two order conditions (Figure 3E and 3F). A mixed ANOVA revealed a main effect of task representation, $F(1,10) = 27.26$, $p < 0.001$, but no task representation x order interaction, $F(1,10) = 2.358$, $p < 0.156$, suggesting no difference in state transition learning between the two order conditions (Table S3). The results demonstrate that spatial representation allows subjects to learn state transitions better, but only in certain circumstances (namely extened experience) subjects use state knowledge to implement model-based strategy.




B
ST: Learning Rate
100
50
0
% of 0-point symbols chosen
[every 8 trials]
Temporal
Spatial
NS
*
*
NS
1
2
3
4
5
6
7
8
9
10
Time points [8 trials/bin]


C
TS: Choice Behavior
100
50
0
% of 0-point symbols chosen
*
Temporal
Spatial


D
ST: Choice Behavior
100
50
0
% of 0-point symbols chosen
NS
Temporal
Spatial

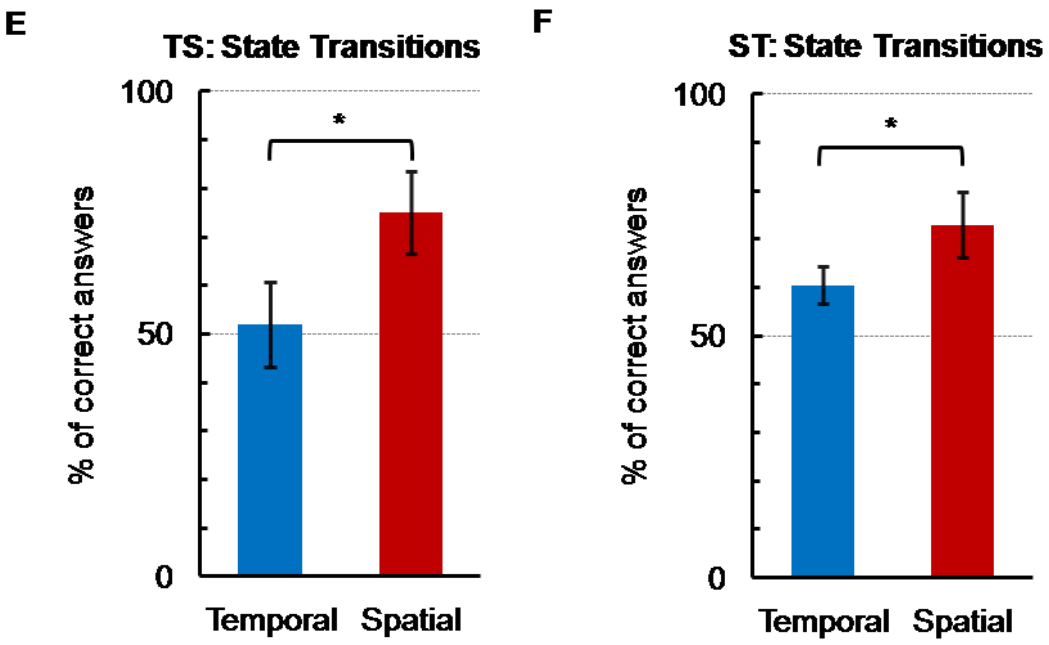


Figure 3. Results comparison according to the order of tasks between the Temporal and Spatial tasks (A) TS: Percentage of 0-point symbols being chosen for each task on average of every eight trials throughout the session. (B) ST: Percentage of 0-point symbols being chosen for each task on average of every eight trials throughout the session. A paired-wise t-test was conducted on each data point (Table S1). In the TS condition, the Temporal task was played in the first and second sessions, followed by the Spatial task in the third and forth sessions. The order was the opposite for the ST condition. (C) TS: Average percentage of 0-point symbols being chosen for each task in the second half of the session. (D) ST: Average percentage of 0-point symbols being chosen for each task in the second half of the session. (E) TS: Average percentage of correct answers to State Transitions question. (F) ST: Average percentage of correct answers to State Transitions question. Error bars represent intersubject SEM. Note. † = $p < .10$, *= $p < .05$, **= $p < .01$, ***= $p < .001$. $n = 6$.

DISCUSSION

The existence of both model-free and model-based learning strategies has been shown to influence human choice behaviors in a two-stage Markov decision task. Although the probabilistic and sequential structure of a two-stage Markov task is rudimentary, to our knowledge, no reinforcement learning research has paid attention to differences in elemental

features of task representation. In this preliminary study, we investigated a fundamental role of task representation in human reinforcement learning. We designed two two-stage Markov tasks, referred to as Temporal and Spatial. The probabilistic and sequential structure of both tasks was similar to previous work's (Gläscher et al., 2010), except that there was a possibility of obtaining immediate rewards (40 points) after the first-stage choice. Importantly, the only difference in representational structure between the two tasks was the position of decision states. Throughout each session, in the Temporal task all states always situated in the middle of the screen, while in the Spatial task the position of each of four color states was located on a specific corner of the screen. Albeit a small sample size (N=12), our results show that embedding a spatial feature into task representation significantly improves subject's learning about the structure of state transitions. However, only after having an extended experience, an embedded spatial feature elicits more dependence on model-based learning strategy.

Comparison with related studies

In human, our findings may relate to the behavioral results of one computational and neuroimaging study using the Markov task, similar to our Temporal task (Gläscher et al., 2010). The main finding of this study shows that there are two different signals in the brain that dissociates between state and reward prediction errors (SPEs and RPEs). Unlike our experimental design, this study resembles animal latent learning experiment in that during the first experimental session subjects are free to explore the non-spatial task and learn about state transitions in absence of rewards, and afterwards in the second session reward contingencies are introduced. In comparison, in our tasks subjects have to concurrently learn symbol-reward associations and state transition structure. Accordingly, some aspects of behavioral results are notable for a discussion in order to guide future studies that aim to investigate the arbitration process between learning strategies. Although their task was non-spatial, the neural and behavioral analysis demonstrated that subjects could learn about state transitions in absence of rewards. On the contrary, our behavioral results supported that subjects were able to learn state transition structure (substantially above the chance level) only in the Spatial task. There might be some explanations for the discrepancy—though it is important to note that the purpose of the two studies is not the same.

First of all, the behavioral analyses were approached differently. While we assessed subjects' explicit knowledge about rewards and state contingencies directly after the task, the previous

study concluded that subjects successfully learned state transitions based on the logic that the first choice made by most subjects (13 out of 18) in the first state of the second session was an optimal choice. Although they did not report a possible inconsistency of first-state choices throughout the session, the state-learning signal was detected on a trial-by-trial basis. This might be because the task statistics of rewards in a latent learning session permits more contribution of model-based learning strategy (Simon & Daw, 2011a). Without rewards (low noise in the statistics of reward), model-based learning is at advantageous, providing subjects an opportunity to learn pure-state transition contingencies. Thus, albeit in a non-spatial task, state transition learning might occur in their case. On the contrary to previous studies, our tasks had high noise in general, since subjects had to simultaneously learn both the symbol-reward associations and state transition structures. The result from Reward Contingencies question confirmed that subjects were able to successfully learn symbol-reward associations, regardless of task representation types. Since the explicit goal of the tasks is to obtain rewards, learning about reward associations may be prioritized. In this way, our spatial feature may accommodate the possibility of state transition learning, in addition to reward association learning, which explains the discrepancy between the results.

Furthermore, as opposed to Gläscher and colleagues (2010)'s, our experimental design does not allow us to distinguish pure state representation learning from value-based representation learning since there is a chance of confounded association between first-stage rewards (i.e. 0 or 40 points) and state transitions. More specifically, first-stage actions (i.e. choosing a certain symbol) may be represented by their associated rewards (i.e. either 0 or 40 points), which are then mapped with their transitions to second-stage states, instead of a direct mapping between choosing a symbol and state transitions. Nonetheless, our finding provides clear evidence that task representation potentially impacts state transition learning. Spatial representation significantly enhances subjects' capability to learn state transition structure.

Theoretical explanations: task designs, features, and statistics

Despite the improvement in state transition learning, the behavioral implementation of model-based strategy was not always prominent in the Spatial task. Albeit plausible variability between subjects' individual dominant strategy, Preferred Strategy question could not successfully separate between the two learning strategies. Since our experiment uses a within-subject design, it is unlikely that individual differences—such as working memory capacity (e.g.

Eppinger, Walter, Heekeren, & Li, 2013; Smittenaar et al., 2013) or risk aversion (Dayan & Niv, 2008)—are the major cause. In contrast with previous studies which suggest more supremacy of model-based strategy (e.g. Daw et al., 2011; Gläscher et al., 2010), model-free seems to be prevalent learning strategy in our experimental paradigm. Possible reasons for the difference in behavioral results could potentially be accounted by scrutinizing task designs as well as task features and task statistics.

Similar task design that minimizes learning about reward contingencies—comparable to Gläscher and colleagues (2010)'s—has also been used to isolate state transition learning (Daw et al., 2011). The fact that subjects have been pre-trained to understand the task structure, such as state transitions, possibly primes the use of model-based strategy (Akam, Costa, & Dayan, 2015). In this way, the observable choice behaviors naturally favor dependency of model-based learning strategy. However, in such experimental paradigms it is also questionable whether the resulting choice behaviors may be ascribed to the process of learning or the implementation of learned evaluations during the basic computational processes in decision-making (see for a review Rangel et al., 2008). Our task design reduces this ambiguity by having subjects concurrently learned the task structure—i.e. reward and state transition contingencies—ensuring that our behavioral observation is mainly due to the reinforcement learning process.

Furthermore, theoretically, a predominance of each learning strategy can also be explained based on task's features and statistics (Botvinick et al., 2015). One influential computational theory suggests that the degree of reliance on each learning strategy depends on their relative uncertainties, which quantifies ignorance about the long-term true values of actions (Daw et al., 2005). In principle, uncertainty about actions has an impact on the degree of exploration, in which a decision choice accounts not only for the advantage of expected future rewards but also the beneficial prospect of learning about the unknown (for a review, see Dayan & Niv, 2008). In this respect, uncertainty tends to urge more model-based strategy, driving agents to explore the unknown parts of task structure. Within this framework, a shift from model-free to model-based strategy also depends on the complexity of the task and the contiguity of actions to rewards. The fact that our task design has an immediate reward (40 points at the first stage) may increase the relative value of first-stage choices in terms of the action-reward proximity, compared to second-stage choices. On top of that, with a low and limited number of trials, this effect might, to some extent, prevent subjects from exploring the task and realizing that first-stage 0-point symbols in the long run could lead to the optimal reward (i.e. 100 points). Therefore, in our tasks, this feature possibly primes more model-free learning strategy. In

consequence, some subjects kept choosing 40-point symbols once they learned the probable association between symbols and rewards, providing a plausible reason for more reliance on model-free learning strategy in general.

Within the uncertainty-based framework, other causes for the supremacy of model-free strategy may be the specificity of a goal and the reliability of prediction errors (see Lee, Shimojo, & O'Doherty, 2014; O'Doherty et al., 2015). Having specific goals, in contrast to flexible goals, prompted more forward-planning, mode-based strategy. Our experimental instruction states that subject's goal is to get as many points as possible—without revealing that obtaining 100 points is the optimal strategy. This produces flexibility in interpretations and approaches to achieving such goal. In fact, during the informal debriefing after the end of experiment, some subjects reported that their reward maximization strategy was to 'try to choose only symbols that gave points,' even though they realized that 100 was the maximum point in both tasks. In this particular case, the goal then was only to always get points, regardless of the magnitude. Although model-free and model-based strategies are not synonymously translatable to habitual and goal-directed behaviors, the computational process of each learning strategy can capture the essential characteristics of each respective type of behavior (Daw & O'Doherty, 2014; Dayan & Niv, 2008; Smittenaar et al., 2013). However, goal-directed behavior can also interact with habitual behavior in a hierarchical manner, by which the goal-directed system determines a goal and selects a set of actions that can then be executed more efficiently (using less cognitive resources) with model-free strategy (Dezfouli & Balleine, 2013). This might be the reason why in our tasks subject's choice behavior, despite being goal-directed, signifies prevalence of model-free learning strategy.

As an extension to the uncertainty-based framework, the arbitration between the two learning strategies also depends on the competition between the degree of reliabilities of RPEs and SPEs in the brain (see O'Doherty et al., 2015). Low reward and state transition uncertainties (i.e. high probability of 0.8 in our tasks) reflect high reliabilities of RPE and SPE, respectively. A combination of high reliabilities of both prediction errors predicts more model-free strategy (see Supplemental Information in Lee et al., 2014). Therefore, according to the model, this competition between highly reliable prediction errors favors the employment of model-free strategy, providing a reason for prevalence of model-free strategy in our tasks. Nonetheless, these explanations could not clarify why the effect of task representation on forward-planning, model-based learning strategy was significantly higher only during the Spatial task (later in the third and forth sessions) in the TS condition.

Many research have shown that model-based strategy tends to transition to model-free strategy over time (Daw et al., 2005; Keramati, Dezfouli, & Piray, 2011; Yin, Knowlton, & Balleine, 2004), but model-based strategy can assume control in unpredicted circumstances (Isoda & Hikosaka, 2011; Norman & Shallice, 2000 as cited in Wunderlich et al., 2012). When performing the Spatial task in the third and forth sessions, the structure of state transitions may become more apparent. In consequence, this novel learning produces a higher level of volatility present in the environment, since new information about state transitions is incorporated (O'Doherty et al., 2015). Based on the uncertainty framework, increase in volatility could also trigger more model-based learning strategy, which might explain why there was a significant shift from model-free to model-based strategy only in the TS condition's Spatial task.

In addition, given a trend of a transition to model-based strategy in the TS condition, but not in the ST condition, a propensity to realize model-based strategy might depend not only on properties of the tasks—i.e. representation, features and statistics—but to some extent also on learning experience. In our tasks, the overall structures of both tasks (i.e. reward and state transition contingencies) were samilar. For example, symbols that typically gave 0 points usually led to the second-stage state with a 100-point symbol for both Temporal and Spatial tasks. Due to a small number of trails per session (80 trials) compared previous studies (around 200 trials), subjects might learn about this regularity of task structure over time, and synthesize their learning strategy across sessions (Akam et al., 2015). In this way, the number of observations (trials) can affect the arbitration of learning strategies (Lee et al., 2014). Under model-based strategy, the reliabiliy of prediction errors increases with accumulating evidence. Thus, it is also likely that model-based strategy may take over model-free strategy, as subjects have a priori knowledge about state transition structures and gain more experience with the task. Nonetheless, no strong explanation can be concluded since there was no significant interaction between task representations and order conditions in our current results.

Despite aforementioned limitations, our finding that spatial representation improves state transition learning corroborates a conceptual account that the brain can create an abstract model mapping the perceived states of the environment, inspired by the concept of cognitive map. A neurophysiological evidence for the cognitive aspect of cognitive map intuitively stems from neural substrate of place cells (or hippocampal pyramidal neurons) in hippocampus (e.g. Johnson & Redish, 2007; Wikenheiser & Redish, 2015). Place cells exhibit high level of spatial selectivity, forming representations of spatial locations in the environment. Having a geographical representation of states possibly makes the task's attribute of state transitions

more salient (O'Doherty et al., 2015). In this respect, our Spatial task might accommodate more ecological solicitation of neural resources from hippocampal areas, enabling subjects to learn state transitions better by configuring a mental map of states and state transitions (Gibson, 2014). In fact, the wealth of spatial navigation work focusing on investigating viewpoint-specific representations—though not akin to our study of reinforcement learning—provide evidence for neural underpinnings of allocentric representations in human's parahippocampal regions (Burgess, 2006; Ekstrom et al., 2003; Iglói, Zaoui, Berthoz, & Rondi-Reig, 2009). Potentially, our results may provide clues linking human spatial memory with reinforcement learning process—bridging the gap between different experimental paradigms and drawing on the broader picture of mechanisms underlying learning and decision-making.

Future research

Overall, our study champions the fact that fundamentally task representation plays an essential role in studying human reinforcement learning and decision-making (Botvinick et al., 2015; Rangel et al., 2008). Embedding spatial feature into task representation effectively improves state transition learning. However, only after having extended experience with the task, can spatial representation trigger higher reliance on forward-planning, model-based strategy. In principle, task features and task statistics may interactively influence the arbitration between two learning strategies, by modulating task's uncertainties such as noise, volatility, and reliability of prediction errors. In addition, task design can also influence subjects' understanding of task structure and experience with the task, which consequentially may impact the balance of learning strategies. Due to a small sample size and some insignificant results, more evidence is required before making any definite claims.

Henceforth, under the current task design, further investigation may consider collecting more data up to 22 subjects to account for an effect size. The approximated number of subjects was calculated using a priori power analysis of the program *G*Power* (Cohen, 1988; Faul, Erdfelder, Lang, & Buchner, 2007) with d = 0.56; statistical power = 0.8 as input parameters. Also, since our results seem to point out the influence of order conditions, using a between-subject design with approximately 36 subjects for each task (i.e. Temporal and Spatial), d = 0.6 and statistical power = 0.8, might be advantageous in order to control for the order effect. Furthermore, it would be interesting to incorporate different methodologies (e.g. computational modeling, fMRI, electroencephalogram (EEG) and eye-tracking) to synchronously investigate both behavioral

responses and related neural activities. For instance, a combination between EEG and eye-tracking would allow us to verify whether subjects implement forward-planning model-based or repeating model-free strategy, by detecting event-related negativity (e.g. Holroyd, Nieuwenhuis, Yeung, & Cohen, 2003; Yeung, Botvinick, & Cohen, 2004) and tracking the point of gaze after decisions (Duchowski, 2007). In addition, depending on the research goal, future studies may consider adjusting/adding some changes in the task design. For example, studies that aim to explore model-based learning strategy may consider including more trials to prompt more model-based strategy and/or implementing a latent learning session before introducing reward contingencies—instead of having two attempts for each task (which did not provide a significant effect)—to elicit pure state transition learning. In this way, two different types of associative learning (reward versus state transition) can be unambiguously dissociated.

By taking an incremental yet critical step, our results provide empirical evidence that task presentation influences state transition learning and, under certain circumstance, the arbitration between model-free and model-based learning strategies. All in all, this study provides a promising behavioral account to further understand the importance of task representation in human reinforcement learning.

SUPPLEMENTAL INFORMATION

Supplemental Information for this study includes the calculation of expected values, the experimental instruction, one figure for the average percentages of correct answers to Symbol-State associations, State-Stage associations, and Reward Contingencies post-test questions, and three statistical analysis tables.

ACKNOWLEDGEMENTS

Many thank to Dr. Maël Lebreton for his comments on previous versions of this manuscript and his dedicated supervision throughout this study.

REFERENCES

Akam, T., Costa, R., & Dayan, P. (2015). Simple Plans or Sophisticated Habits? State, Transition and Learning Interactions in the Two-Step Task. *PLoS Comput Biol*, *11*(12), e1004648.

Berns, G. S., McClure, S. M., Pagnoni, G., & Montague, P. R. (2001). Predictability modulates human brain response to reward. *The Journal of Neuroscience*, *21*(8), 2793–2798.

Botvinick, M., Weinstein, A., Solway, A., & Barto, A. (2015). Reinforcement learning, efficient coding, and the statistics of natural tasks. *Current Opinion in Behavioral Sciences*, *5*, 71–77. http://doi.org/10.1016/j.cobeha.2015.08.009

Burgess, N. (2006). Spatial memory: how egocentric and allocentric combine. *Trends in Cognitive Sciences*, *10*(12), 551–557. http://doi.org/10.1016/j.tics.2006.10.005

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences . Hilsdale*. Nj: Erlbaum.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron*, *69*(6), 1204–1215. http://doi.org/10.1016/j.neuron.2011.02.027

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711. http://doi.org/10.1038/nn1560

Daw, N. D., & O'Doherty, J. P. (2014). Multiple Systems for Value Learning. In *Neuroeconomics* (pp. 393–410). Elsevier. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/B9780124160088000218

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The Good, The Bad and The Ugly. *Current Opinion in Neurobiology*, *18*(2), 185–196. http://doi.org/10.1016/j.conb.2008.08.003

Dezfouli, A., & Balleine, B. W. (2013). Actions, Action Sequences and Habits: Evidence That Goal-Directed and Habitual Action Control Are Hierarchically Organized. *PLoS Computational Biology*, *9*(12), e1003364. http://doi.org/10.1371/journal.pcbi.1003364

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, *22*(6), 1075–1081. http://doi.org/10.1016/j.conb.2012.08.003

Duchowski, A. (2007). *Eye tracking methodology: Theory and practice* (Vol. 373). Springer Science & Business Media. Retrieved from https://books.google.nl/books?hl=en&lr=&id=WtvVdNESRyIC&oi=fnd&pg=PR15&dq=eye+tracking+spatial+navigation&ots=8lxe2uEN8y&sig=bh5V375OYzIQpdZkZEgf6fJ1w1o

Ekstrom, A. D., Kahana, M. J., Caplan, J. B., Fields, T. A., Isham, E. A., Newman, E. L., & Fried, I. (2003). Cellular networks underlying human spatial navigation. *Nature*, *425*(6954), 184–188.

Eppinger, B., Walter, M., Heekeren, H. R., & Li, S.-C. (2013). Of goals and habits: age-related and individual differences in goal-directed decision-making. *Frontiers in Neuroscience*, *7*. http://doi.org/10.3389/fnins.2013.00253

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Fermin, A., Yoshida, T., Ito, M., Yoshimoto, J., & Doya, K. (2010). Evidence for model-based action planning in a sequential finger movement task. *Journal of Motor Behavior*, *42*(6), 371–379.

Gibson, J. J. (2014). The Theory of Affordances. *The People, Place, and Space Reader*, 56.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free

Reinforcement Learning. *Neuron*, *66*(4), 585–595. http://doi.org/10.1016/j.neuron.2010.04.016

Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the Role of the Orbitofrontal Cortex and the Striatum in the Computation of Goal Values and Prediction Errors. *Journal of Neuroscience*, *28*(22), 5623–5630. http://doi.org/10.1523/JNEUROSCI.1309-08.2008

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., & Cohen, J. D. (2003). Errors in reward prediction are reflected in the event-related brain potential. *Neuroreport*, *14*(18), 2481–2484.

Iglói, K., Zaoui, M., Berthoz, A., & Rondi-Reig, L. (2009). Sequential egocentric strategy is acquired as early as allocentric strategy: Parallel acquisition of these two navigation strategies. *Hippocampus*, *19*(12), 1199–1211. http://doi.org/10.1002/hipo.20595

Isoda, M., & Hikosaka, O. (2011). Cortico-basal ganglia mechanisms for overcoming innate, habitual and motivational behaviors. *European Journal of Neuroscience*, *33*(11), 2058–2069.

Johnson, A., & Redish, A. D. (2007). Neural Ensembles in CA3 Transiently Encode Paths Forward of the Animal at a Decision Point. *Journal of Neuroscience*, *27*(45), 12176–12189. http://doi.org/10.1523/JNEUROSCI.3761-07.2007

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/Accuracy Trade-Off between the Habitual and the Goal-Directed Processes. *PLoS Computational Biology*, *7*(5), e1002055. http://doi.org/10.1371/journal.pcbi.1002055

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural Computations Underlying Arbitration between Model-Based and Model-free Learning. *Neuron*, *81*(3), 687–699. http://doi.org/10.1016/j.neuron.2013.11.028

Norman, D. A., & Shallice, T. (2000). Attention to action: Willed and automatic control of behavior. *Cognitive Neuroscience: A Reader*, 376–390.

O'Doherty, J. P., Lee, S. W., & McNamee, D. (2015). The structure of reinforcement-learning mechanisms in the human brain. *Current Opinion in Behavioral Sciences*, *1*, 94–100. http://doi.org/10.1016/j.cobeha.2014.10.004

Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013). The curse of planning dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological Science*, 956797612463080.

Pagnoni, G., Zink, C. F., Montague, P. R., & Berns, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience*, *5*(2), 97–98. http://doi.org/10.1038/nn802

Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature Communications*, *6*, 8096. http://doi.org/10.1038/ncomms9096

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*(7106), 1042–1045.

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, *9*(7), 545–556.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Simon, D. A., & Daw, N. D. (2011). Environmental statistics and the trade-off between model-based and TD learning in humans. In *Advances in neural information processing systems* (pp. 127–135). Retrieved from http://papers.nips.cc/paper/4243-environmental-statistics-and-the-trade-off-between-model-based-and-td-learning-in-humans

Simon, D. A., & Daw, N. D. (2011). Neural Correlates of Forward Planning in a Spatial Decision Task in Humans. *Journal of Neuroscience*, *31*(14), 5526–5539. http://doi.org/10.1523/JNEUROSCI.4647-10.2011

Smittenaar, P., FitzGerald, T. H. B., Romei, V., Wright, N. D., & Dolan, R. J. (2013). Disruption of Dorsolateral Prefrontal Cortex Decreases Model-Based in Favor of Model-free Control in Humans. *Neuron*, *80*(4), 914–919. http://doi.org/10.1016/j.neuron.2013.08.009

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning* (Vol. 135). MIT Press Cambridge. Retrieved from http://www.cs.utexas.edu/sites/default/files/research/documents/1%20intro%20up%20to%20RL%3ATD.pdf

Thistlethwaite, D. (1951). A critical review of latent learning and related experiments. *Psychological Bulletin*, *48*(2), 97.

Thorndike, E. L. (1933). A proof of the law of effect. *Science*. Retrieved from http://psycnet.apa.org/psycinfo/1933-01793-001

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189.

Wikenheiser, A. M., & Redish, A. D. (2015). Decoding the cognitive map: ensemble hippocampal sequences and decision making. *Current Opinion in Neurobiology*, *32*, 8–15. http://doi.org/10.1016/j.conb.2014.10.002

Wunderlich, K., Smittenaar, P., & Dolan, R. J. (2012). Dopamine Enhances Model-Based over Model-Free Choice Behavior. *Neuron*, *75*(3), 418–424. http://doi.org/10.1016/j.neuron.2012.03.042

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychological Review*, *111*(4), 931.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience*, *19*(1), 181–189.

Supplemental Information

The Role of Task Representation in Reinforcement Learning Strategies

Thanaphat Thongpaibool

Calculation of expected values for choice decisions

Under the assumption that subjects already learned task's reward contingencies, we calculated the expected values (EVs) for two sets of decision driven by model-free and model-based learning strategies according to our task probabilistic structure (Figure 1C). Overall, the probability of a certain symbol giving a certain reward is 0.8.

Without considering state transition structure, model-free strategy predicts that subjects are likely to choose symbols that frequently result in points. Thus, in the first-stage decision, 40-point symbols will be repeatedly chosen and, in doing so, 40-point symbols will lead to the state with a 10-point symbol for 80 percent of the time and to the state with a 100-point symbol for 20 percent of the time. Assuming that subjects always choose symbols that mostly give points in the second-stage decision, the expected value of model-free choice behavior can be calculated as followed:

First-stage decision: choosing 40-point symbols

Second-stage decision: choosing symbols with points

$$\mathbf{EVmfree} = \overbrace{[0 \times 0.2 + 40 \times 0.8]}^{\text{First-stage decision: choosing 40-point symbols}} + \overbrace{\{0.8 \times [10 \times 0.8 + 0 \times 0.2] + 0.2 \times [100 \times 0.8 + 0 \times 0.2]\}}^{\text{Second-stage decision: choosing symbols with points}}$$

$$= 54.4$$

In contrast, model-based strategy predicts that subjects will likely choose first-stage 0-point symbols in order to get to the second-stage state that has a 100-point symbol, which occurs 80 percent of the time. Similarly, in the second stage subjects always choose symbols that mostly give points. The expected value of model-based choice behavior can be computed as followed:

First-stage decision: choosing 0-point symbols

Second-stage decision: choosing symbols with points

$$\mathbf{EVmbased} = \overbrace{[0 \times 0.8 + 40 \times 0.2]}^{\text{First-stage decision: choosing 0-point symbols}} + \overbrace{\{0.8 \times [100 \times 0.8 + 0 \times 0.2] + 0.2 \times [10 \times 0.8 + 0 \times 0.2]\}}^{\text{Second-stage decision: choosing symbols with points}}$$

$$= 73.6$$

Therefore, the optimal reward maximization strategy is to consider both reward contingencies and state transition structure—implementing model-based learning strategy (i.e. choosing 0-point symbols in the first stage).

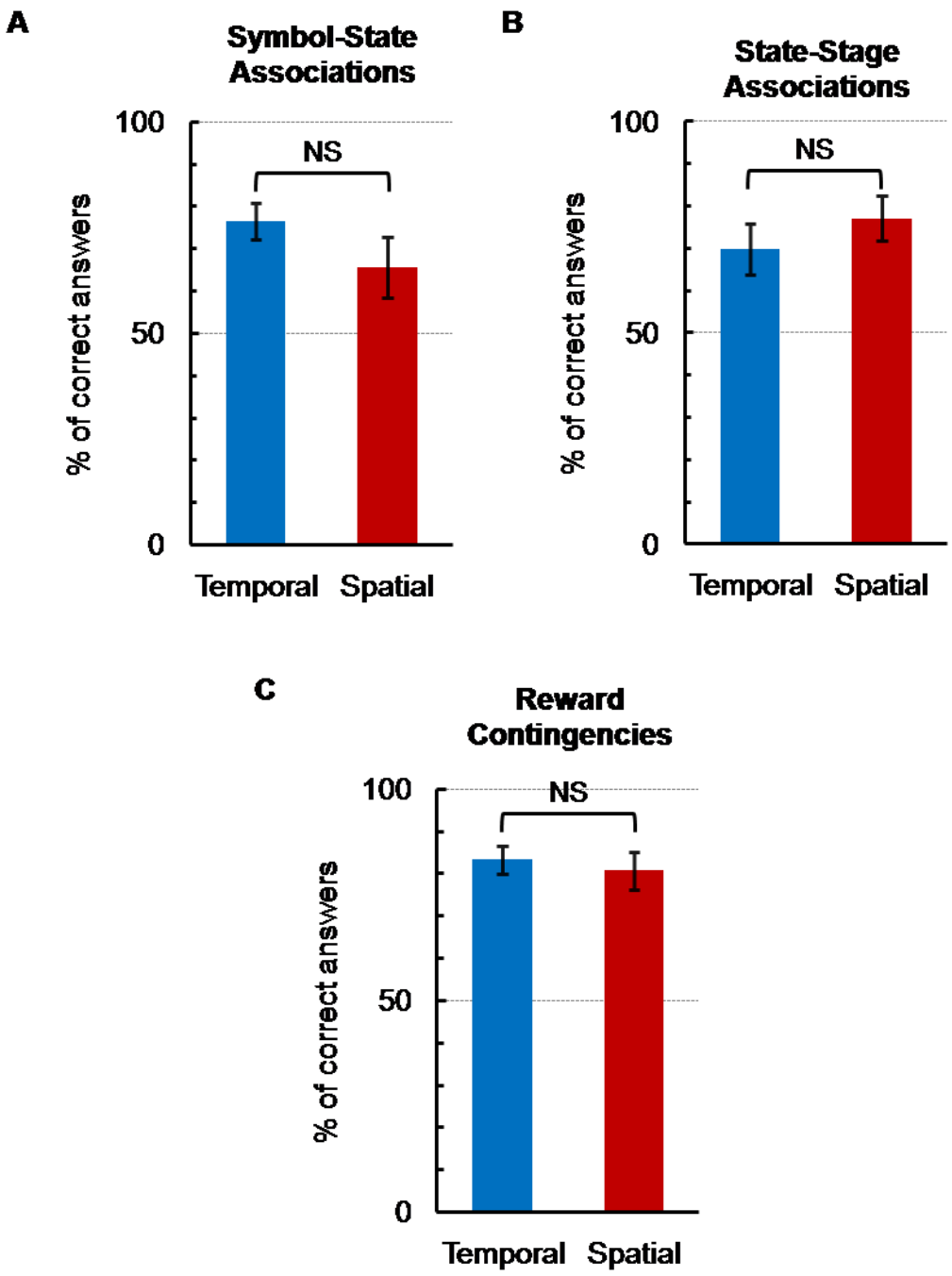


Figure S1. Results comparison between the Temporal and Spatial tasks. Average percentage of correct answers to (A) Symbol-State Associations, (B) State-Stage Associations, and (C) Reward Contingencies. Error bars represent intersubject SEM.

Table S1. A paired-wise t-test of the learning rate on each data points for overall, TS, and ST conditions (related to Figure 3A, 4A, and 4B, respectively)

Time point	#Trial	Overall		TS		ST	
		***t*(11)**	**p**	***t(5)***	**p**	***t(5)***	**p**
1	1-8	0.90	0.3852	0.35	0.741	0.88	0.419
2	9-16	2.57	0.026*	1.07	0.332	2.74	0.041*
3	17-24	3.32	0.007**	2.09	0.091†	2.99	0.031*
4	25-32	1.80	0.099†	2.15	0.084†	0.36	0.732
5	33-40	1.79	0.100	4	0.010*	-0.18	0.867
6	41-48	1.83	0.095†	2.89	0.034*	0.15	0.889
7	49-56	2.01	0.070†	3.25	0.023*	0.46	0.666
8	57-64	1.97	0.075†	3.48	0.018*	-0.18	0.867
9	65-72	2.38	0.037*	4.30	0.008**	0.37	0.728
10	73-80	1.56	0.147	4.04	0.010*	-0.14	0.892

Note. † = p < .10, *= p < .05, **= p < .01, ***= p < .001. *N* = 12.

Table S2. A mixed ANOVA on choice behavior with task representation (Temporal versus Spatial) and attempt (First versus Second) as within-subject factors and order condition (TS versus ST) as a between-subject factor

Within-subject effects	**12 subjects**				
	SS	**df**	**MS**	**F(1,10)**	**p**
Task representation	4700.52	1	4700.52	5.00	0.049*
Task representation x Order	4313.02	1	4313.02	4.59	0.058†
Attempt	1813.02	1	1813.02	2.86	0.122
Attempt x Order	1354.69	1	1354.69	2.13	0.175
Task representation x Attempt	168.75	1	168.75	0.15	0.709
Task representation x Attempt x Order	2.08	1	2.08	0.00	0.967
Between-subject effect					
Order	5208	1	5208	2.384	0.154

Note. † = p < .10, *= p < .05, **= p < .01, ***= p < .001, TS: *n = 6,* ST: *n = 6.*

Table S3. A mixed ANOVA on state transition learning with task representation (Temporal versus Spatial) and attempt (First versus Second) as within-subject factors and order condition (TS versus ST) as a between-subject factor

Within-subject effects	**12 subjects**				
	SS	**df**	**MS**	**F(1,10)**	**p**
Task representation	3763.02	1	3763.02	27.26	< 0.001***
Task representation x Order	325.52	1	325.52	2.36	0.156
Attempt	13.02	1	13.02	0.03	0.862
Attempt x Order	117.19	1	117.19	0.29	0.604
Task representation x Attempt	117.19	1	117.19	0.17	0.693
Task representation x Attempt x Order	117.19	1	117.19	0.17	0.693
Between-subject effect					
Order	117.2	1	117.2	0.10	0.754

Note. † = p < .10, *= p < .05, **= p < .01, ***= p < .001, TS: *n = 6,* ST: *n = 6.*

The Experimental Instruction

Learning to maximize outcome in a two-stage decision task.

Instruction

The goal of this experiment is to earn as many points as you can.

EXPERIMENT DESIGN: The experiment consists of a succession of computerized tasks, implementing two-stage choices (see next paragraph for details). There are 4 sessions in total. Each session comprises a main task of 80 2-stage choices trials (i.e. 160 choices per session) and 5 extra post-test questions which assess your understanding of the task. You will earn points from making choices in the main task, which in the end will be translated to actual money. Overall one session takes about 15 minutes. Before the beginning of the main experiment, there will be a short training session, which consists of 10 2-stage choices trials (i.e. 20 choices).

MAIN TASK STRUCTURE: Each trial consists of **two decision-stages**. In each stage there are two possible color states (e.g. red and blue), each with a certain set of two abstract symbols. This means one state consists of two symbols, a color boundary, and a cross sign, and that a pair of symbol always belongs to the same state (Figure 1).

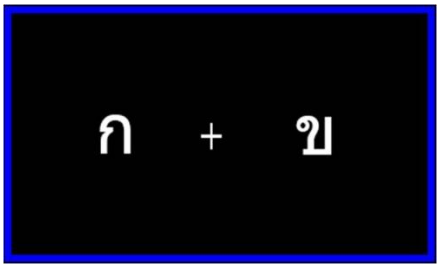


Figure 1. An example of a state with two specific symbols.

In one session, there are 2 possible color states per stage (hence 4 in total), each with a specific pair of symbols. In each stage of each trial, you are asked to choose between one of the two different symbols displayed on either side of a cross sign inside the colored-boundary state. Critically, within the same state, the two symbols are not equivalently **advantageous**. One of the two symbols is on average more **rewarded**, i.e. you can win more points compared to the other. At the same time, each symbol in the first stage will **lead predominantly on average to a certain second-stage** state. Then, how **advantageous** is a first stage symbol, in the long-run, depends both on how rewarded it is and how profitable the second-stage it most often leads to. Therefore, since the task is probabilistic, your goal is to progressively learn by trial-and-errors and by exploring the different options, to maximize the number of points, by choosing what appears to be the most advantageous symbols.

At the end of each session, you will receive your total points for that session. Your final payment will correspond to the sum of the points you earn from all four sessions.

MAIN TASK IMPLEMENTATION: To choose one of the two symbols, press either the right or left button. You can choose each option at your own pace. **Note that:** You cannot change your choice after you press the button. After you have chosen your choice, there will be an arrow indicating your choice, and the outcome will be displayed for 1.5 seconds. Again, for each trial, you have to make two choices—one for each stage.

At the end of each trial, the word 'next trial' will appear shortly, and the next trial will start.

POST-TESTS STRUCTURE: For each session, there are 5 post-test questions (listed according to their order below):

1. Which of these two stimuli is more advantageous from your experience?
2. In which state does this symbol belong to?
3. In which stage does this color state belong to?
4. Which second state does this symbol mostly lead to?
5. Which outcome does this symbol mostly lead to?

For Post-test questions, you can answer at your own pace by pressing the keyboard number to choose the answer and then spacebar to confirm. **Note that:** You can change your post-test answer by pressing different number key before you continue.

If you have any questions about the experiment and the tasks, please ask the experimenters for clarification before starting the experiment.