Cogn Affect Behav Neurosci (2017) 17:764–783
DOI 10.3758/s13415-017-0511-2

A simple computational algorithm of model-based choice preference

Asako Toyama [1,2,3] · **Kentaro Katahira** [1,2] · **Hideki Ohira** [1,2]

Published online: 1 June 2017

Abstract A broadly used computational framework posits that two learning systems operate in parallel during the learning of choice preferences—namely, the *model-free* and *model-based* reinforcement-learning systems. In this study, we examined another possibility, through which model-free learning is the basic system and model-based information is its modulator. Accordingly, we proposed several modified versions of a temporal-difference learning model to explain the choice-learning process. Using the two-stage decision task developed by Daw, Gershman, Seymour, Dayan, and Dolan (2011), we compared their original computational model, which assumes a parallel learning process, and our proposed models, which assume a sequential learning process. Choice data from 23 participants showed a better fit with the proposed models. More specifically, the proposed eligibility adjustment model, which assumes that the environmental model can weight the degree of the eligibility trace, can explain choices better under both model-free and model-based controls and has a simpler computational algorithm than the original model. In addition, the forgetting learning model and its variation, which assume changes in the values of unchosen actions, substantially improved the fits to the data. Overall, we show that a hybrid computational model best fits the data. The parameters used in this model succeed in capturing individual tendencies with respect to both model use in learning and exploration behavior. This computational model provides novel insights into learning with interacting model-free and model-based components.

Keywords Computational model · Model-free · Model-based · Eligibility trace · Reinforcement learning

One common theoretical framework is that value-based decision-making is realized using two distinct cognitive or learning systems: One is habitual and inflexible and requires little computation, whereas the other is deliberative and accurate and requires heavy computation (Dickinson, 1985; Kahneman, 2010; Redish, Jensen, & Johnson, 2008). In the field of instrumental learning, these two systems correspond to the *model-free* and *model-based* learning systems, respectively (Daw, Niv, & Dayan, 2005; Dolan & Dayan, 2013; Gillan, Otto, Phelps, & Daw, 2015). Prediction that is based on model-free learning is analogous to Thorndike's law of effect, in which a behavior that is followed by a pleasant outcome is likely to be repeated, whereas a behavior that is followed by an unpleasant outcome is likely to be inhibited (Thorndike, 1911). In contrast, the model-based learning system uses the agent's internal model, or *cognitive map* (Tolman, 1948), of a structure in the environment to dynamically change a behavior by propagating information to all states and actions, including those that have not previously been experienced. However, it has yet to be determined how humans and animals shape a preference that is based on these learning systems and how the interaction of these systems is implemented.

Electronic supplementary material The online version of this article (doi:10.3758/s13415-017-0511-2) contains supplementary material, which is available to authorized users.

✉ Asako Toyama
asako.toyama@gmail.com

[1] Department of Psychology, Graduate School of Environmental Studies, Nagoya University, Nagoya, Japan

[2] Department of Psychology, Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8601, Japan

[3] Japan Society for the Promotion of Science, Tokyo, Japan

The two-stage decision task developed by Daw, Gershman, Seymour, Dayan, and Dolan (2011) is well-suited to address these questions. In this task, the model-free and model-based systems make different predictions about the choice, depending on the reward outcome and the transition at the previous trial. The relative contributions of these mechanisms are projected onto a weighting parameter of their computational model. This task has provided interesting findings that show that the bias in one system relates to disorders that involve compulsion (Voon et al., 2015) and alcohol dependence (Sebold et al., 2014), working memory capacity (Otto, Gershman, Markman, & Daw, 2013; Otto, Raio, Chiang, Phelps, & Daw, 2013), and individual traits such as extraversion (Skatova, Chan, & Daw, 2013). In addition, neural substrates that are critical to these learning systems have been searched for under this framework (Gläscher, Daw, Dayan, & O'Doherty, 2010; Smittenaar, FitzGerald, Romei, Wright, & Dolan, 2013; Wunderlich, Smittenaar, & Dolan, 2012).

The original computational model of the task above assumes that the model-free and model-based values are computed in parallel. The model-free values are calculated by a state–action–reward–state–action (SARSA) (λ) temporal-difference (TD) learning rule (Rummery & Niranjan, 1994) using reward prediction errors, whereas the model-based values are calculated using max values of each future state weighted by a transition probability. These two independent values are ultimately combined with a certain weight to produce one net value for a choice. Thus far, however, it remains to be seen whether this type of value-updating structure, which uses two independent learning mechanisms and its combination mechanism, is the best candidate to reflect the decision-making process using the environmental model.

The other possible learning mechanism is based on the model-free learning system. For example, the DYNA architecture proposed by Sutton (1990) supposes that the model-free learning system updates values on the basis of both real experiences and model-based simulated experiences. Although some experiments support this idea (Gershman, Markman, & Otto, 2014), this architecture includes a black-box simulation process and requires offline training of the model-based system. To implement more simple and immediate online control from the environmental model, we focused on the eligibility trace rule often used in TD learning (Sutton & Barto, 1998), which is part of the model-free learning mechanism and can be realized as persistent neural activities modulated by previous actions (Curtis & Lee, 2010). The *eligibility trace* is a temporary record of the occurrence of state–action pairs and expresses how much these past events contributed to the outcome, determining the degree of value updating. We extended this algorithm to reflect the model-based prediction by changing the eligibility with an internal model of the task's transition structure. Thus, the proposed model assumes a sequential value-updating process by the model-free learning system.

We also examined an additional hypothesis in a TD learning process. TD learning generally assumes that only the chosen option value is updated, while the other values remain the same (Sutton & Barto, 1998). However, a variant TD learning model that hypothesizes value changes for unchosen actions has recently succeeded in capturing the choices of both monkeys (Barraclough, Conroy, & Lee, 2004) and rats (Ito & Doya, 2009). In addition, this model has shown some important characteristics of choice behavior, such as a dependence on choice history (Katahira, 2015). However, this mechanism has not yet been examined in Markov decision processes, including state transitions such as the present task. Here we applied this hypothesis to a computational model to determine whether it improves the model's fit to data.

Thus, the purpose of this study was to examine the hypotheses above by using the two-stage decision task (Daw et al., 2011) to find an alternative model with a simpler value-updating rule that both requires parsimonious computation, as compared to the existing model, and satisfactorily predicts the balance of the model-free and model-based effects on choice behavior. The exploration of this topic is important and urgently required because the original computational model is already widely used as a basis for identifying the brain regions involved in model-based decision-making or qualifying the characteristics of some psychiatric diseases, as we mentioned above.

Method

Participants

Twenty-three undergraduate students at Nagoya University (12 males, 11 females; age $M = 19.1$ years, $SD = 0.8$) participated in the experiment. All participants gave informed consent, and the study was approved by the ethics committee at Nagoya University. The participants were paid ¥1,000, with additional monetary rewards between ¥500 and ¥584 that were calculated by multiplying .4 by the money earned in the two-stage decision task. The additional rewards of the two participants who earned less than ¥500 were rounded up to ¥500.

Two-stage decision task

The two-stage decision task was based on a procedure developed by Daw et al. (2011). The present task consisted of three blocks that contained 101 trials, which were separated by 30-s breaks. Each trial had two stages (Fig. 1A). The participants first chose one of two fractal images in the first stage (state A) by clicking a corresponding mouse button within 2 s. The selected fractal was highlighted with a yellow frame. Each fractal at state A led to states B and C at different rates. One option led to state B at a rate of 70% (*common* transition) and state C at a

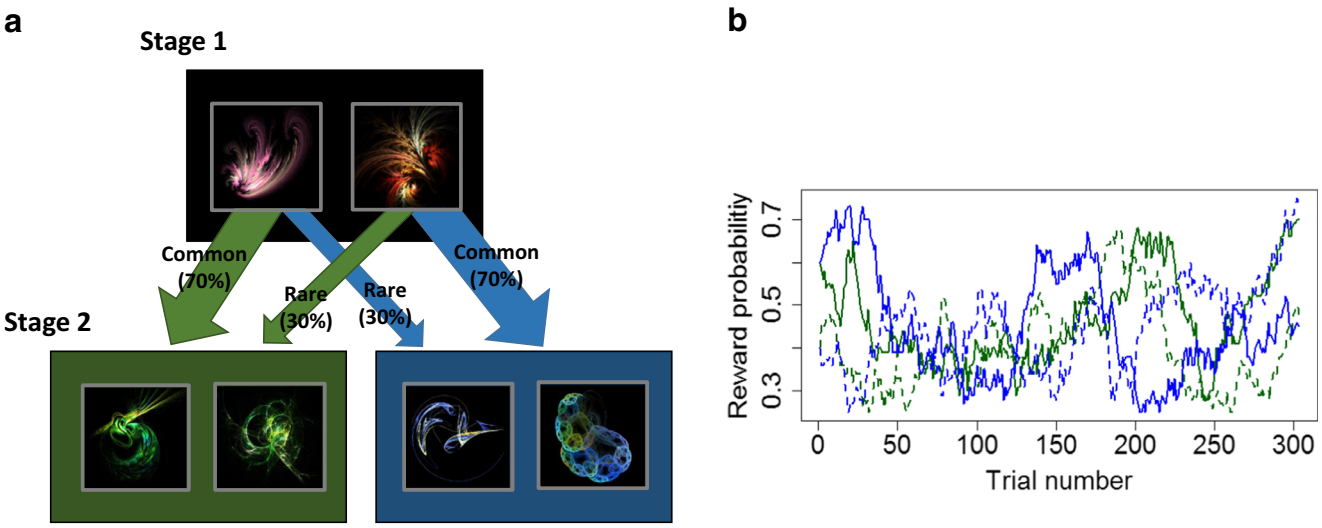

Fig. 1 Two-stage decision task. (A) Each trial involved two sequential stages. The participants chose one of two fractal images. Each of these image had a common transition (70%) and a rare transition (30%), each of which led to a specific state at the second stage. Depending on the second-stage choice between two fractal images, the participants received reward feedback that indicated whether or not they had received ¥10. (B) The reward probabilities for each second-stage image changed slowly and independently according to Gaussian random walks, but they were the same for all participants.

rate of 30% (*rare* transition), whereas the other led to states B and C at the reversed rates. In the second stage, each state had a particular color background and fractals to easily distinguish the two states, and the small fractal image that had been chosen in the first stage was shown at the top of the screen as a reminder. The participants again chose one of two fractals in the second stage within 2 s, and the selected fractal was again highlighted. Depending on this choice, feedback was given for 1 s that indicated whether the participants were rewarded or unrewarded. When the participants were rewarded, a picture of ¥10 was shown, and when they were not rewarded, a red "x" mark was shown. The participants were told in advance that each first-stage fractal tended to lead to a particular state in the second stage, but not which one. They would receive the total earned money that was discounted by a certain rate after the experiment. In both stages, if no response was made within 2 s, a message that said "Too late!!" was presented, and the participants proceeded to the next trial.

The reward probabilities of each second-stage fractal had previously been determined and independently and slowly changed over trials with drifting Gaussian random walks (*SD* = .025) between .25 and .75 (Fig. 1B). The reward probabilities were the same for all participants, to exclude the possibility that individual differences in performance and the computational model parameters' fits to the data arose out of differences in the task-reward condition. Before the experiment, the participants were told that the reward probability of the second-stage fractals would change slowly and independently over time, and they had a ten-trial training session before the task.

The simulated model-free and model-based choice predictions are shown in Fig. 2A. The model-free learning was based on the outcome of the second stage. Therefore, if a choice was rewarded, the next choice at the first stage might have stayed the same as the previous trial, but if it was unrewarded, the next choice might have changed. In contrast, the model-based learning used the information about the transition rate. If a choice was rewarded after a rare transition, the next choice at the first stage might change in order to go to the same second-stage state; if the choice was unrewarded after a rare transition, the next first-stage choice might remain unchanged to go to another second-stage state.

Operation span task and questionnaires

After the two-stage decision task, the operation span task was conducted to estimate working memory capacity. In addition, the participants answered questionnaires on psychological distress, impulsivity, and attention. These data are beyond the purpose of this study, and thus are not reported in this article.

Awareness and intentional use of the model concerning the transition structure

At the end of all procedures, the participants indicated whether they had noticed the particular transition structure, and if they did, whether they made choices based on information about the previously experienced outcome and transition. We asked the participants whether they made choices using information about the previously experienced outcome and transition by giving a concrete example. Specifically, we asked "Did you intend to change the first-stage choice from your choice in the previous trial after you were rewarded with the rare transition, or did you intend to reselect the same first-stage option after you missed the reward with the rare transition?"

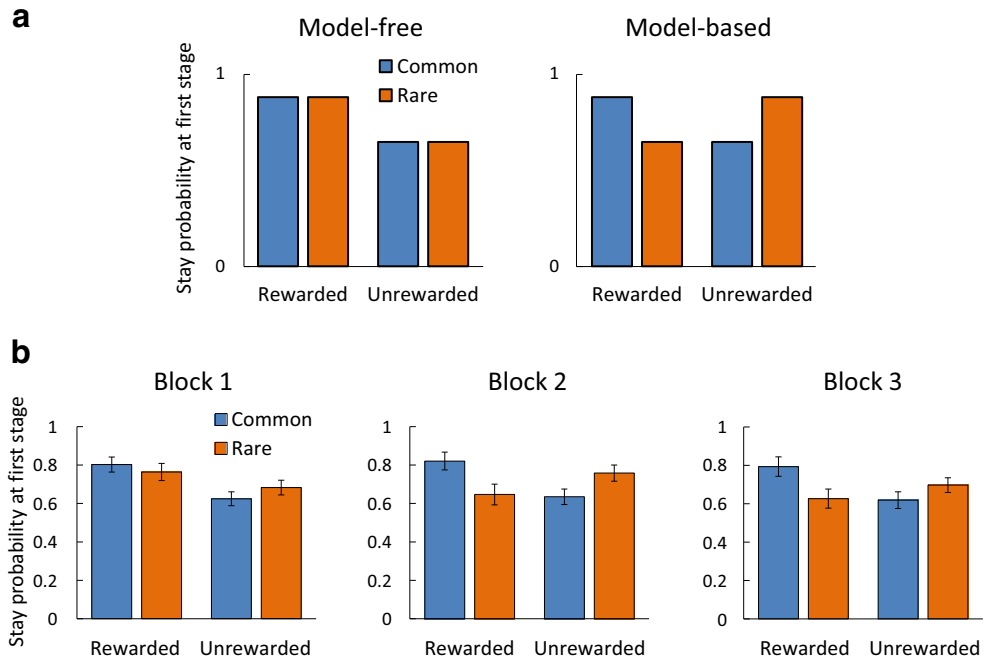

Fig. 2 Effects of the transition and outcome of the previous trial. (A) Graphs of predicted stay probabilities at the first stage for the purely model-free system and the purely model-based system. The model-free system predicts that the participant will reselect the same option after the choice is rewarded and will change after the choice is unrewarded (left graph), but the model-based system predicts an interaction between transition and outcome (right graph). (B) The observed stay probabilities are shown. On average, across participants, model-free control (the main effect of the previous outcome) was observed in Block 1, and model-based control (the interaction between the previous transition and outcome) was seen in Blocks 2 and 3.

Models

Next we searched for a suitable computational model to explain the data we obtained. In the following section, we first introduce the original model proposed by Daw et al. (2011) to explain the data for this task. In this article, we call it the *parallel-learning model*. Then we introduce a model that assumes a different connection mechanism between the model-free and model-based components from the original model: the *eligibility adjustment model* (the EA model). Subsequently, two models are introduced that assume changes in the unchosen action values—namely, the forgetting Q-learning model (the F model) and a variation that hypothesizes regression to a certain default value (the FD model). Finally, we propose four hybrid models: the parallel–F model, the EA–F model, the parallel–FD model, and the EA–FD model.

The parallel-learning model (from Daw et al., 2011) The parallel-learning model (Daw et al., 2011; Daw et al., 2005) hypothesizes that parallel processes calculate the model-free and model-based values; these processes are combined by a weighting parameter w to a net value for a choice.

The model-free learning system uses a SARSA (λ) TD-learning rule (Rummery & Niranjan, 1994) and updates state–action values, $Q_{MF}(s_{i,t}, a_{i,t})$ at each stage i of each trial t. In the present task, there are three types of states (s_A for $s_{1,t}$ and s_B and s_C for $s_{2,t}$). Each state has two available actions, and $a_{i,t} \in \{1, 2\}$ denotes the chosen action. The value of a chosen state–action value is updated as follows:

$$Q_{MF}(s_{i,t}, a_{i,t}) \leftarrow Q_{MF}(s_{i,t}, a_{i,t}) + \alpha_i\left(r_{i,t} + Q_{MF}(s_{i+1,t}, a_{i+1,t}) - Q_{MF}(s_{i,t}, a_{i,t})\right), \quad (1)$$

where α_i is a learning rate parameter that is unique for each stage, and $r_{i,t}$ denotes the rewards at trial t (with $r_{i,t} = 1$ when the reward is given, and $r_{i,t} = 0$ when the reward is not given). Here, $r_{1,t}$ and $Q_{TD}(s_{3,t}, a_{3,t})$ are always 0, because there is no reward at the first stage and no next stage after the second stage. The values of Q_{MF} at the first and second stages are updated as follows:

$$Q_{MF}(s_{1,t}, a_{1,t}) \leftarrow Q_{MF}(s_{1,t}, a_{1,t}) + \alpha_1\left(Q_{MF}(s_{2,t}, a_{2,t}) - Q_{MF}(s_{1,t}, a_{1,t})\right), \quad (2)$$

$$Q_{MF}(s_{2,t}, a_{2,t}) \leftarrow Q_{MF}(s_{2,t}, a_{2,t}) + \alpha_2\left(r_{2,t} - Q_{MF}(s_{2,t}, a_{2,t})\right). \quad (3)$$

At the end of each trial, all first-stage values are again updated according to the second-stage reward prediction error (RPE), the difference between the expected and actual rewards, as follows:

$$Q_{MF}(s_{1,t}, a_{1,t}) \leftarrow Q_{MF}(s_{1,t}, a_{1,t}) + \alpha_1\lambda\left(r_{2,t} - Q_{MF}(s_{2,t}, a_{2,t})\right), \quad (4)$$

where λ denotes the trace decay parameter, which modulates the degree of the effect of the second-stage RPE on the first-stage value (Sutton & Barto, 1998). This type of updating is called the *eligibility trace* rule; $\lambda = 1$ indicates that its effect is maximum, and $\lambda = 0$ indicates no effect.

The first-stage model-based values, Q_{MB}, for each action are updated depending on the transition probability function T, which is a function of the transition probabilities from a first-stage action to the following second-stage states, whereas the second-stage Q_{MB} is equivalent to Q_{MF}, because there is no transition to the next stage. Thus, the values of Q_{MB} for the first-stage actions are as follows:

$$\begin{aligned} Q_{MB}(s_A,\ a_j) &= T\left(s_B\middle|s_A, a_j\right) \max_{a\in\{a_1,\ a_2\}} Q_{MB}(s_B,\ a) \\ &+\left(1 - T\left(s_B\middle|s_A, a_j\right)\right) \max_{a\in\{a_1,\ a_2\}} Q_{MB}(s_C,\ a) \end{aligned}, \quad (5)$$

where $T\left(s_B|s_A, a_j\right)$ is a transition probability function corresponding to the transition probability to the state s_B after taking action a_j at state s_A, which is simply defined as the binary value of .7 or .3, following Daw et al. (2011). $T\left(s_C|s_A, a_j\right)$ equals $\left(1 - T\left(s_B|s_A, a_j\right)\right)$. When there are two actions a_1 and a_2 in the state s_A, $T(s_B|s_A, a_1)$ and $T(s_C|s_A, a_2)$ are set at .7 if the occurrence of s_B following a_1 plus s_C following a_2 is greater than that of s_C following a_1 plus s_B following a_2 at that time; otherwise, they are .3. Finally, the first-stage Q_{MF} and Q_{MB} are integrated to make a net value with a weighting parameter w.

$$wQ_{net}(s_A,\ a_j) = wQ_{MB}(s_A,\ a_j) + (1-\mathrm{w})Q_{MF}(s_A,\ a_j). \quad (6)$$

At the second stage, Q_{net} is the same as Q_{MF} and Q_{MB}. These net values determine the choice probabilities $P\left(a_{i,t} = a|s_{i,t}\right)$ of choosing action i from two candidates, as follows:

$$P\left(a_{i,t} = a\middle|s_{i,t}\right) = \frac{\exp\left[\beta_i\left(Q_{net}\left(s_{i,t},\ a\right) + p\cdot\mathrm{rep}(a)\right)\right]}{\sum_{a'}\exp\left[\beta_i\left(Q_{net}\left(s_{i,t},\ a'\right) + p\cdot\mathrm{rep}(a')\right)\right]}, \quad (7)$$

where β_i is an inverse temperature at each stage and determines the bias between the value dependency and the randomness of choice. The choice trace weight p is a parameter that controls the tendency toward preservation ($p > 0$) or switching ($p < 0$) in the first-stage actions; $rep(a)$ is 1 if a is a first-stage action and is the same as the action that was chosen on the previous trial, and it is 0 otherwise.

In summary, in each trial, the parallel-learning model first updates Q_{MF} and Q_{MB} separately. For Q_{MF}, the chosen state–action values are updated using Eq. 2 at the first stage and Eq. 3 at the second stage, and the first chosen state–action value is again updated by eligibility trace using Eq. 4. For Q_{MB} in the first stage, values are calculated using Eq. 5, whereas at the second-stage Q_{MB} has the same value as the second-stage Q_{MF}. At the end, Q_{MF} and Q_{MB} are combined with Eq. 6. As we explain at the beginning of the next section, we actually used Eq. 8 instead of Eq. 4 for the comparison with other models; this change had no effect on the fit of the parallel-learning model.

Eligibility adjustment model First, we changed the conventional equation of the eligibility trace rule. The eligibility trace typically works by conveying the RPE to the previously experienced eligible state–action pairs, as in Eq. 4. We assumed that the reward itself is used to update the previous state–action pairs as follows:

$$\begin{aligned} Q_{MF}(s_{1,t},\ a_{1,t}) &\leftarrow Q_{MF}(s_{1,t},\ a_{1,t}) \\ &+\ \lambda\left(r_{2,t} - Q_{MF}(s_{1,t},\ a_{1,t})\right). \end{aligned} \quad (8)$$

This change in the equation is merely a reparameterization of the parallel-learning model, but it is easy to adapt to the other models that include the updating of unchosen options, such as the EA model introduced in this section (see Supplementary Text 1). Therefore, it becomes easy to make comparisons between the parallel-learning model and the other models by using Eq. 8 instead of Eq. 4 for all models. In the SARSA(λ) TD learning and parallel-leaning models, there was no difference in fit, regardless of whether Eq. 4 or Eq. 8 was used in the eligibility trace part, because of the reparameterization. Furthermore, although these equations produce different fit results in the other models, the models consistently showed better fits when we used the proposed Eq. 8 than when we used the conventional Eq. 4 (see Supplementary Text 1 and Table S1 for details).

In addition, we hypothesized that the environmental model (transition probabilities) is used in the update rule of the eligibility trace, and we proposed a new model called the EA model. The critical characteristic of this model is that one value is assigned to one state–action pair. This is different from the parallel-learning model, which assumes two values for each state–action pair. Thus, the values of the EA model are simply noted as Q, and Q_{MF} in the other models shall be replaced with Q when they are combined with the EA model.

In the EA model, the model-based component is combined in the eligibility trace equation. The fully model-free eligibility trace is the same as Eq. 8, where the reward is used to update the previously chosen state–action pairs in that trial. However, we theorized that fully model-based eligibility tracing is performed proportional to the probabilities from the first-stage action to the second-stage state. Under this model-based system, the unchosen action value in the first stage is also updated on the basis of the subjective transition

probability, because it also had a possibility of reaching the currently visited second-stage state in that probability if it were chosen. Thus, in the EA model, eligibility trace updating for the chosen action $a_{1,t}$ includes mixed effects of the model-free and model-based systems on a weighting parameter, and the unchosen action $\overline{a_{1,t}}$ is also updated under the model-based system as follows:

$$Q(s_{1,t},\ a_{1,t}) \leftarrow Q(s_{1,t},\ a_{1,t}) + \lambda\left[w{\cdot}T\left(s_{2,t}\middle|s_{1,t},a_{1,t}\right)\left(r_{2,t}-Q(s_{1,t},\ a_{1,t})\right) + (1-w)\left(r_{2,t}-Q(s_{1,t},\ a_{1,t})\right)\right], \tag{9}$$

$$Q\left(s_{1,t},\ \overline{a_{1,t}}\right) \leftarrow Q\left(s_{1,t},\ \overline{a_{1,t}}\right) + \lambda w{\cdot}T\left(s_{2,t}\middle|s_{1,t},\ \overline{a_{1,t}}\right)\left(r_{2,t}-Q\left(s_{1,t}\ \overline{a_{1,t}}\right)\right), \tag{10}$$

where $T\left(s_{2,t}|s_{1,t},a_{1,t}\right)$ is simply defined as the binary value of .7 or .3 introduced in the explanation for Eq. 5. Here, w $(0\leq w\leq 1)$ controls the balance of the model-free and model-based effects in the eligibility trace. If this parameter equals 0, the eligibility trace is totally performed in the model-free manner, whereas if this parameter equals 1, the eligibility trace is totally performed in the model-based manner. The unchosen action value $\overline{a_{1,t}}$ is also updated under the model-based system. In the long run, this type of model-based eligibility trace may reflect the neural networks, which are activated depending on the frequencies of the experienced transitions and propagate the reward information to the eligible actions in proportion to their activation.

In the original model (Daw et al., 2011), the transition probability function T is used when calculating the model-based value (Q_{MB}). In contrast, in this EA model Q_{MB} is not calculated, but the information concerning the transition probability is used to adjust the model-based degree of the eligibility trace. This EA model thus requires a lower cost than the original model, which calculates and then combines two types of values.

In summary, on each trial under the EA model, the chosen state–action values are first updated using Eq. 2 in the first stage and Eq. 3 in the second stage. At the end of the trial, both of the first-stage values are updated by adjusting the eligibility traces: The chosen state–action value is determined by Eq. 9, and the unchosen state–action value is determined by Eq. 10.

The forgetting learning model and its variation (the F and FD models) Typical TD learning assumes that the value of the chosen option is updated by the outcome, although the values of the unchosen options remain unchanged. However, this assumption is unnatural when considering that memory decays over time. To recover from this limitation, some researchers have introduced new mechanisms to take into account decay in the values of unchosen options and have achieved better fits using real data (Barraclough et al., 2004; Ito & Doya, 2009). Accordingly, the unchosen action values, including the action values of the unvisited state, are updated as follows:

$$Q_{MF}\left(s_{i,t},\ \overline{a_{i,t}}\right) \leftarrow Q_{MF}\left(s_{i,t},\ \overline{a_{i,t}}\right) - \alpha_F Q_{MF}\left(s_{i,t},\ \overline{a_{i,t}}\right), \tag{11}$$

where α_F is called a *forgetting* parameter. The model that uses this rule is called the *forgetting learning model* (F model). In the model fitting, we assumed the special case that α_F is the same as the learning rate in the first and second stages. Although we also tried other models that assumed that α_F is different from the learning rate, the best-fitting model in this study supported the present assumption.

Equation 11 predicts that the value of unchosen options is close to 0 if they are not chosen for a long time. However, it is again unnatural to assume this result, particularly when considering the behavioral tendency known as exploration. This behavior is partly caused by an expectation that rarely-chosen uncertain options may have advantages over often-chosen certain options. Therefore, the value of the unchosen options might not be monotonically devalued, but instead might be regressed to a certain default expected value. Particularly in this task, because the participants were first instructed that the rewarded probabilities of the second-stage options would continue to change, it was natural to think that uncertain options would have some expected value. To include this idea in the model, we added a default-value parameter, μ $(0\leq\mu\leq 1)$, and formalized that the values of unchosen options were regressed to the default value μ.

$$Q_{MF}\left(s_{i,t},\ \overline{a_{i,t}}\right) \leftarrow Q_{MF}\left(s_{i,t},\ \overline{a_{i,t}}\right) + \alpha_F\left(\mu - Q_{MF}\left(s_{i,t},\ \overline{a_{i,t}}\right)\right) \tag{12}$$

Under this rule, all initial Q values are equal to μ. As a predicted tendency, a μ value smaller than the recently chosen option value promotes the avoidance of unchosen options, whereas a μ greater than the recently chosen option value promotes active exploration. Thus, the μ parameter becomes an index that determines the exploration tendency. We call the model that includes this updating rule the *forgetting-to-default learning model* (FD model).

Thus, in each trial, the chosen state–action values are first updated using Eq. 2 in the first stage and Eq. 3 in the second stage, and the unchosen and unvisited state–action values are updated using Eq. 11 in the F model and Eq. 12 in the FD model. At the end of trial, the first chosen state–action value is again updated by eligibility trace using Eq. 8.

Hybrid models (parallel or EA models with a forgetting mechanism) To combine the assumptions concerning the integration mechanism of the model-free and model-based components and the updating rule of the model-free component, we proposed four hybrid models. Two models are hybrids of the F and the parallel and EA models (the parallel–F and EA–F models); the other two models are hybrids of the FD and the parallel and EA models (the parallel–FD and EA–FD models).

The concrete updating procedures of each model are as follows. In the parallel–F model, Q_{MF} for the chosen state–action are updated using Eq. 2 at the first stage and Eq. 3 at the second stage, and the remaining, unchosen state–action values are updated using Eq. 11. Q_{MF} for the chosen state–action at the first stage are also updated using the eligibility trace rule of Eq. 8. In parallel, Q_{MB} at the first stage are calculated using Eq. 5, whereas the second-stage Q_{MB} are the same as the second-stage Q_{MF}. Ultimately, Q_{MF} and Q_{MB} are combined using Eq. 6. In the EA–F model, the Q values for the chosen state–actions are first updated using Eq. 2 at the first stage and Eq. 3 at the second stage, and the remaining Q, for the unchosen state–actions, are updated using Eq. 11. At the end of the trial, the first-stage Q are again updated by adjusting the eligibility traces: The chosen action value is determined by Eq. 9, and the unchosen action value is determined by Eq. 10. The value-updating processes of the parallel–FD and EA–FD models are the same as those of the parallel–F and EA–F models, respectively, except that Eq. 11 is replaced with Eq. 12.

Measures of model fitting and selection criteria

We used the R function "solnp" in the Rsolnp package (Ghalanos & Theuss, 2015) to estimate the fitting parameters. For a comparison of these models, we computed the Akaike information criterion (AIC; Akaike, 1974) and the Bayesian information criterion (BIC; Schwarz, 1978). These values were given by

$$\mathrm{AIC} = -2LL + 2k, \tag{13}$$

$$\mathrm{BIC} = -2LL + k\cdot\log(n), \tag{14}$$

where LL is the log likelihood, k is the number of free parameters, and n is the total number of choices. The model with a smaller value is considered the preferred model.

Logistic regression analysis

To examine the individual action tendencies, logistic regression analyses were carried out with the R function "glm" separately on the data for each participant. The model tested the effects of previous outcome (coded as rewarded [.5] and unrewarded [–.5]) and previous transition (coded as common [.5] and rare [–.5]) on the stay/switch actions at the first stage (coded as stay [1] and switch [0]).

Results

All participants performed 303 trials. In the following analyses, the data from the uncompleted trials, in which a choice was not made within 2 s, were excluded. We also excluded the data from trials in which the response time was less than 100 ms, because these were considered anticipated responses that did not reflect the stimulus types. An average of 2.96 (SD = 3.06) trials were excluded per participant.

Stay probability at the first stage depending on the previous trial's transition and outcome

The effect of the previous outcome, or the preference for an option followed by a reward, accords with model-free prediction, whereas interaction of the previous outcome and the previous transition accords with the model-based prediction (Fig. 2A). To examine this point, we conducted an analysis of variance (ANOVA) on the stay probabilities at the first stage, with Block (1, 2, and 3), Previous Outcome (rewarded and unrewarded), and Previous Transition (common and rare) as within-subjects factors (Fig. 2B).

We observed a significant main effect of previous outcome [$F(2, 44) = 12.37$, $p = .002$, $\eta_p^2 = .36$], showing that the mean stay probability was higher after the rewarded trials than after the unrewarded trials [rewarded, .74 (SD = .24); unrewarded, .67 (SD = .19)]. Additionally, significant interactions were found between block and previous outcome, previous outcome and previous transition, and all three of the factors [$F(2, 44) = 3.39$, $p = .043$, $\eta_p^2 = .13$; $F(1, 22) = 16.80$, $p < .001$, $\eta_p^2 = .43$; $F(2, 44) = 4.95$, $p = .012$, $\eta_p^2 = .18$, respectively]. There were no other significant effects (all ps > .18). Because of the three-way interaction, we conducted post-hoc analyses to examine the effects of previous outcome and previous transition separately in each block (Table 1). This analysis revealed that in Block 1 participants showed only the significant main effect of previous outcome ($p < .001$). Conversely, in Blocks 2 and 3, participants showed significant interactions between previous outcome and previous transition (both ps < .001), although there were no or only moderate effects of previous outcome (respectively, $p = .22$, $p = .08$). These results showed that the choices were produced by both model-free and model-based systems. The former effect was found mainly in the first block, whereas the latter was mainly in the latter blocks.

Table 1 Main effects and interactions in three blocks

		F	*p* Value
Block 1	**Outcome**	**15.87**	**.0006*****
	Transition	0.25	.620
	Outcome × Transition	2.87	.105
Block 2	Outcome	1.62	.217
	Transition	1.06	.314
	Outcome × Transition	**16.69**	**.0005*****
Block 3	Outcome	3.27	.0843†
	Transition	2.72	.114
	Outcome × Transition	**15.25**	**.0008*****

Significant effects are in bold. $^{\dagger}p < .10$, $^{***}p < .001$

Examination of the forgetting mechanism

In the present task, if the forgetting mechanism worked as intended in the F and FD models, we predicted that the second-stage choice would become noisier in trials that did not include the same second-stage state as the previous trial (*different* condition) than in trials that included the same second-stage state as the previous trial (*same* condition), because the values of both unvisited states would always approach 0 in the F model and μ in the FD model, and the difference between the values would always diminish. To examine this point, we calculated (separately for each condition) the ratio of the second-stage stay probability after being rewarded to that after being unrewarded for the trial on which the same second-stage state was last visited. The result of a paired *t* test revealed that this ratio was bigger in the *same* condition ($M = 1.7$, $SD = 0.6$) than in the *different* condition ($M = 1.4$, $SD = 0.3$) [$t(22) = 3.15$, $p < .001$, $d = 0.66$]. This result supported the existence of a forgetting mechanism. A considerable confounding factor of this effect was that the *different* condition included a larger ratio of rare-transition trials than did the *same* condition [on average, .40 for the *different* condition and .26 for the *same* condition; $t(22) = 5.67$, $p < .001$]. The rare transitions might cause noisy actions relative to the common transitions, and might lead to the results above. To exclude this possibility, we conducted the same analysis but restricted it to the data of the common-transition trials. This analysis resulted in the same effect, showing that the choices in the *same* condition were more sensitive to the previous outcome [$t(22) = 2.80$, $p = .010$, $d = 0.58$].

In addition, to gain further confidence in the prediction above regarding the forgetting mechanism, we conducted the same analyses on synthetic datasets. We first generated 200 datasets for each model, including the SARSA (λ) TD learning, F, FD, parallel-learning, and EA models. Each dataset included 23 simulated data points generated using the best-fitting parameter combinations for each participant in the experiment. In each model, 100 datasets were generated under the same reward probabilities used in the experiment, and the other 100 had newly generated reward probabilities that slowly changed over the course of the 303 trials according to Gaussian random walks ($SD = .025$) with reflecting bounds at .25 and .75. For these datasets, the same *t* tests were conducted, and the number of results significant at the 5% level out of the 100 tests was counted in each condition. The results are shown in Table 2. As we expected, only the models that included the forgetting mechanism showed that the choices in the *same* condition were more sensitive to the previous outcome than those in the *different* condition, whereas the other models showed a trivial number of significant results that would be expected from the significance level. It seems that the FD model showed a stronger forgetting mechanism than the F model. This difference may be caused by the closer distances between a default value μ estimated in the FD model and the options' estimated values, in comparison with the distance between 0 and the options' estimated values.

Awareness of the transition model and its intentional use

We were also interested in whether the participants were truly aware of the transition structure of the task and whether they intentionally used it in a manner that corresponded to the interaction between the transition and outcome. Table 3 shows the participants' answers to these questions after the experiment. Almost all participants, except two, noticed the transition bias of the options at the first stage (Q1 of Table 3), but only one-third used the bias after a rare transition when making their choice at the first stage (Q3 of Table 3).

Comparison of the models

The negative log likelihood (–*LL*), AIC, and BIC were calculated for each model introduced in Models section and, for comparison, a model using standard SARSA (λ) TD learning (see Table 4). All models in the Models section were developed by adding other mechanisms to SARSA (λ) TD learning, and these mechanisms are roughly divided into two. One is a mechanism to make model-free learning more efficient, which is addressed by the F and FD models. The second is a mechanism to incorporate model-based influence into the value updating, which is addressed with by the parallel-learning and EA models. Table 4 also shows the results of the hybrid models.

Of the two mechanisms above, the second is especially important in the present research, because our primary aim was to develop a model to capture the balance of the model-free and model-based systems using the two-stage decision task. Therefore, we first focused on the comparison between the parallel-learning and EA models. At the population level,

Table 2 Numbers of significant results (5% level) out of 100 simulations in the analyses of the forgetting mechanism

	Same Reward Probabilities		New Reward Probabilities	
Model	All	Common	All	Common
SARSA(λ) TD	3 [d = –0.02]	3 [d = –0.05]	4 [d = –0.03]	4 [d = –0.03]
F	52 [d = 0.56]	35 [d = 0.46]	30 [d = 0.49]	18 [d = 0.38]
FD	96 [d = 0.67]	76 [d = 0.58]	87 [d = 0.61]	66 [d = 0.53]
Parallel-learning	7 [d = 0.06]	1 [d < 0.01]	5 [d = 0.02]	4 [d = –0.02]
EA	1 [d = –0.03]	2 [d = –0.03]	5 [d = –0.02]	5 [d = –0.03]

Each cell shows the number of significant results from paired t tests of 100 synthetic datasets regarding the second-stage stay probabilities between the *same* and *different* conditions. These datasets were generated using either the same reward probabilities as in the present experiment or new reward probabilities for each model with the best-fitting parameter combinations in the experiment. Furthermore, the t tests were conducted using all trials or using only common-transition trials. The average effect sizes are shown in brackets. Importantly, all significant results for the F and FD models were obtained with positive t values, indicating that the choices in the *same* condition were more sensitive to the previous outcome than those in the *different* condition

both the AIC and BIC favored the EA model over the parallel-learning model (Table 4). We also conducted a paired t test to examine whether the BIC differences between two models per participant were significantly different from zero. This again revealed that the EA model was significantly better than the parallel-learning model [$t(22) = 3.48$, $p = .002$, $d = 0.73$]. At the individual level, 18 of the 23 participants supported the EA model, as is shown in Fig. 3, which presents the AIC/BIC difference between the two models for each participant.

In addition, to show that the EA model, similarly to the parallel-learning model, can produce both model-based-like and model-free-like behavior, we simulated the predicted stay probabilities at the first stage using the same reward probability conditions used in Fig. 1B. Each bar graph in Figs. 4B and C shows the results of 20 simulations using the best-fitting parameters, where the parameter w was set to equal the estimated one, 0 (model-free), 1 (model-based), or .5 (a mix of model-free and model-based systems). This simulation confirmed that the EA model can produce model-based-like and model-free-like behavior, depending on the value of the parameter w. In addition, both models can produce choice tendencies similar to those in the real data of the present experiment (Fig. 4A) when the estimated w is used. For a comparison, we also show the results of simulation using the F model, in which we added a parameter for the forgetting rate (Fig. 4D). As we hypothesized, this model could not predict model-based behavior, because it lacks the computational structure to incorporate the transition model. We will detail the structural differences between the parallel-learning model and the EA model in the Detailed Comparison Between the Parallel-Learning and EA Models.

Next we focused on the first mechanism. This analysis confirmed that both the F and FD models improved the fits as compared to the model using standard SARSA (λ) TD learning for any criteria (Table 4). Subsequently, we compared the F and FD models to examine the effect of the default value μ, because the F model is a special case of the FD model for which $\mu = 0$. At the population level, the FD model was favored over the F model by the AIC and BIC criteria. The likelihood ratio test also significantly favored the FD model [χ^2 (22) = 178.49, $p < .001$]. At the individual level, a moderate number of participants favored the FD model (11 of 23 participants according to AIC, 8 of 23 according to BIC, and 8 of 23 according to the likelihood ratio test, at $p < .001$).

We then focused on the four hybrid models, which combine both mechanisms discussed above; these models include the parallel–F model, the EA–F model, the parallel–FD model, and the EA–FD model. All hybrid models except the parallel–F model were better than all of the models with only one of the two mechanisms, and the EA–FD model was the best

Table 3 Summary of responses to the posttask questionnaires

About the task's transition structure	Number of participants		
Q1. Did you notice the bias?	Yes: 21		No: 2
Q2. What was the ratio?	"More extreme than 7:3": 6	"7:3": 11	"More moderate than 7:3": 4
Q3. Did you use a transition structure to make the choice, particularly after a rare transition?	Yes: 7	Never: 16	

Almost all participants were aware of the transition bias. However, only seven participants reported that they intentionally used the rare-transition probabilities to control their choices.

Table 4 Information concerning the nine models compared on the basis of their fit to the choices of 23 participants

Model name	Description	Additional free parameters	# of free parameters	-*LL*	AIC	BIC
Model-free learning only						
SARSA (λ) TD		-	6	319.0 (69)	649.9 (138)	676.3 (138)
F	Update unchosen action values using forgetting parameter	-	6	312.8 (75)	637.7 (149)	664.1 (149)
FD	The F model with regression to the default value for unchosen actions	μ	7	309.0 (74)	631.9 (147)	662.7 (147)
Model-free and model-based components						
Parallel learning	Independently calculate the model-free and model-based values	*w*	7	314.0 (69)	641.9 (139)	672.7 (139)
EA	Use the environmental model in the eligibility trace updating	*w*	7	312.1 (69)	638.1 (139)	668.9 (139)
Hybrid models						
Parallel-F	The hybrid of the Parallel-learning and F models	*w*	7	309.3 (74)	632.5 (149)	663.3 (149)
EA-F	The hybrid of the EA and F models	*w*	7	306.8 (75)	627.7 (149)	658.4 (149)
Parallel-FD	The hybrid of the Parallel-learning and FD models	*w*, μ	8	303.5 (73)	623.0 (146)	658.2 (146)
EA-FD	The hybrid of the EA and FD models	*w*, μ	8	302.3 (74)	620.6 (147)	655.8 (147)

This list provides the mean values and standard errors across participants regarding the negative log likelihood (–*LL*), the Akaike information criterion (AIC), and the Bayesian information criterion (BIC) for each model.

model with respect to both the AIC and BIC criteria, followed by the parallel–FD model (Table 4). These two best-fitting models were directly compared using a paired *t* test, which examined whether the BIC differences between the two models were significantly different from zero. We determined that the EA–FD model was significantly better than the parallel–FD model [$t(22) = 2.56$, $p = .018$, $d = 0.53$]. In addition, 18 of 23 participants showed a better fit for the EA–FD model than for the parallel–FD model according to AIC/BIC.

Finally, to confirm the improvement of this best-fitting EA–FD model relative to the parallel-learning model, a similar *t* test was conducted. This showed that the EA–FD model was significantly better than the parallel-learning model [$t(22) = 3.77$, $p = .001$, $d = 0.79$]. The comparisons at the individual level are also shown in Fig. 5. Almost all participants favored the proposed EA–FD model. Furthermore, the EA–FD model was most favored in all three blocks according to AIC (Table S3), although the choice tendency changed across blocks, as is shown in Table 1.

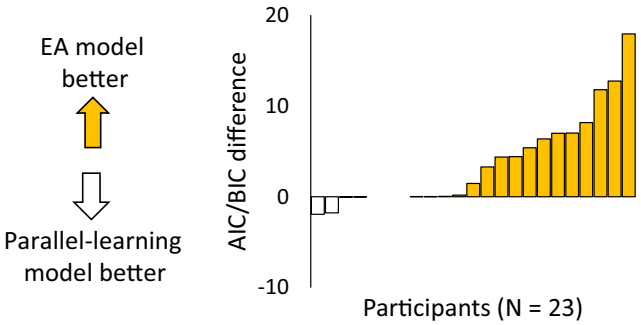

Fig. 3 Differences in the Akaike information criterion (AIC) scores or Bayesian information criterion (BIC) scores between the eligibility adjustment (EA) model and the parallel-learning model for each participant. The results of the two criteria are the same because the same number of free parameters are used in both models. Color bars favor the EA model, and white bars favor the parallel-learning model.

Parameters of the EA–FD model

The lower part of Table 5 shows the estimated parameter values of the EA–FD model. Information about the parallel-learning model is also shown, for reference, in the upper part of Table 5 . The parameter *w* captures the model-based effects in choice behavior. In the EA–FD model, this parameter value was almost zero for six of the 23 participants (<.002), but the remaining participants showed dispersion from .04 to almost 1. Fig. 6A shows the correspondence between this parameter and the participants' intentional model use. The gray bars indicate the participants' reported model use (see also Q3 of Table 3). Although one participant (no. 2) who did not report model use showed a high *w* value, his trace decay parameter λ was extremely low ($\lambda < 10^{-46}$) as compared to all the other participants ($\lambda > 0.03$), which causes his *w* to become unconstrained. This finding indicates that for this participant, the eligibility trace itself had no effect, and the value of *w* was meaningless. For the purposes of comparison, we also show the same graph of *w* for the other models (Fig. 6, B C, and D). The correspondence of the parameter and the participants' intentional model use appears high and occurs in the order EA–FD model > EA model > parallel–FD model > parallel-learning model. Although there are no large differences among the first three models, this order supports the

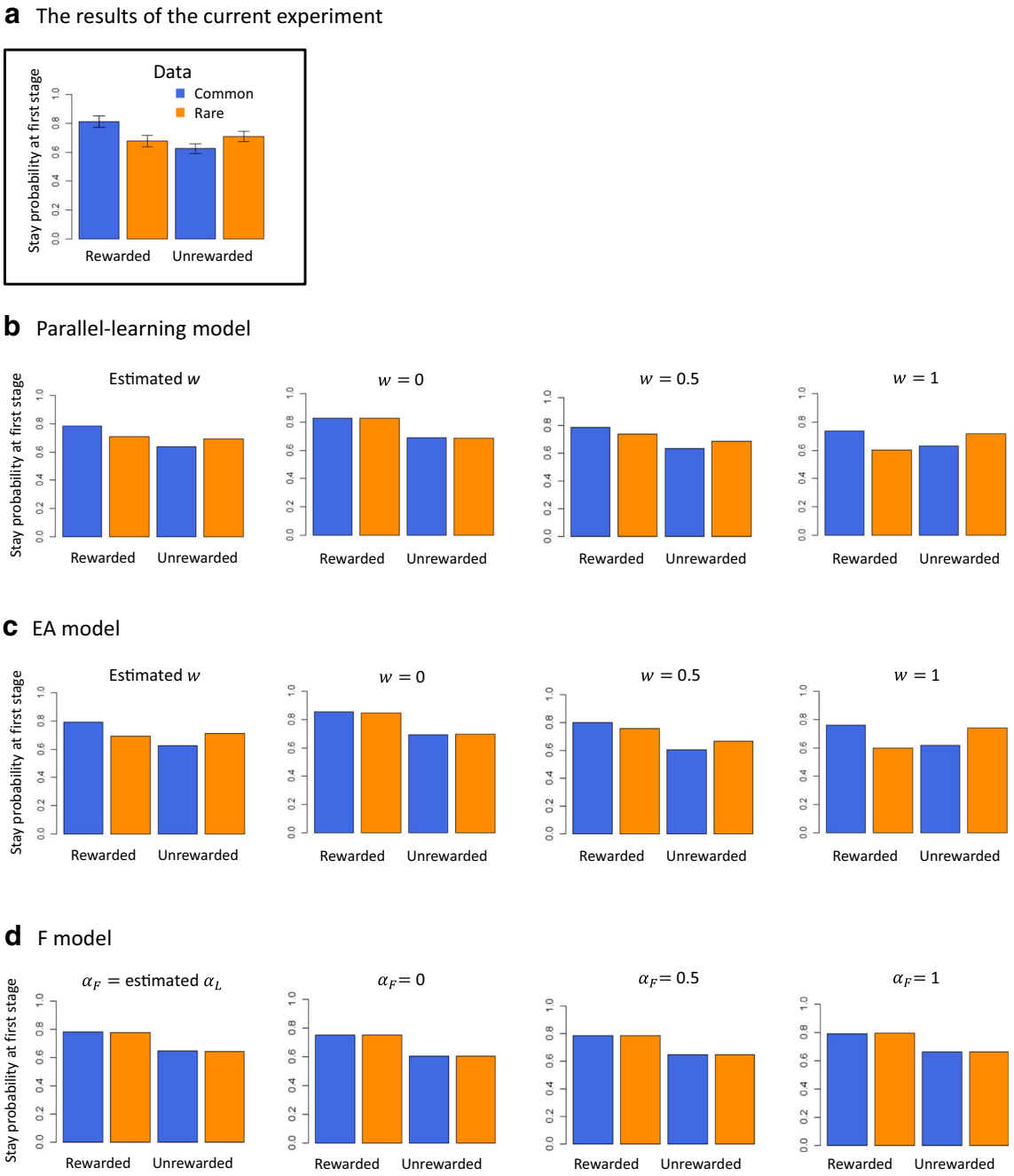

Fig. 4 Observed and simulated results depending on the parameter value. (A) Observed stay probabilities at the first stage. (B–D) Results of simulations using the best-fitting parameters of the parallel-learning model (B), EA model (C), and forgetting F model (D). In each graph, one parameter is attenuated. The simulated results of the parallel-learning and EA models using the estimated parameters are similar to the observed pattern. In addition, both models can produce model-free-like and model-based-like behavioral patterns when w = 0 and w = 1, respectively. In the F model, the forgetting rate α_F was assumed to be the same as the learning rate α_L in the parameter estimation. The F model cannot represent model-based-like behavior for any value of α_F.

usefulness of the parameter w in the EA models. Although the other parameter, λ, in the eligibility trace models determines the overall degree of the eligibility trace, we confirmed that w is a better predictor of model use than λ, by confirming the correlations between individual values of the parameters and reported model use or the results of individual-based logistic regressions (Table S2).

In addition, as expected, another new parameter μ, which denotes the default value, was negatively correlated with stay probability (r = –.56, p = .005; see Fig. 7). In other words, the

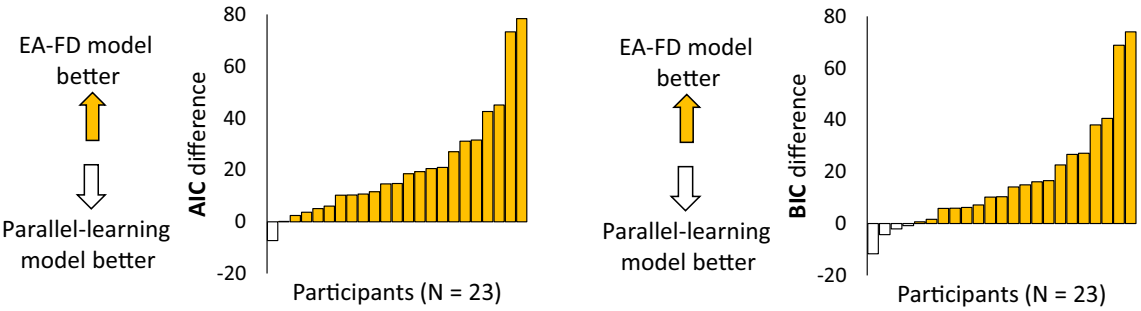

Fig. 5 Differences in Akaike information criterion (AIC) scores (left panel) and Bayesian information criterion (BIC) scores (right panel) between the hybrid EA–forgetting model with default-value parameter (EA–FD model) and the parallel-learning model for each participant. Color bars favor the EA–FD model, and white bars favor the parallel-learning model.

higher default value prompted an exploration tendency as a result of enhanced expectations of uncertain options.

Detailed comparison between the parallel-learning and EA models

In this section, we ascertain which part of the updating processes accounted for the fitting difference between the parallel-learning model and the EA model, and then we investigate which aspects led to the improvement of the fitting in the proposed EA model by comparing the structural differences between the two models in detail.

First, to observe the fitting difference between the parallel-learning and EA models in the first and second stages, we calculated the means of the estimated choice probabilities for the action chosen in each trial for each stage of each model (Fig. 8A). Here, geometric means, which exponentiate the average log probabilities, were calculated, because they directly reflect the difference in log likelihoods. An ANOVA was conducted on the means using Model (parallel-learning and EA) and Stage (first and second) as within-subjects factors. There was a significant effect of model, showing that the EA model was better than the parallel-learning model [$F(1, 22) = 11.29$, $p = .003$, $\eta_p^2 = .34$]. In addition, we found a significant interaction between model and stage [$F(1, 22) = 10.62$, $p = .004$, $\eta_p^2 = .33$]. This revealed that the better choice predictability of the EA model was significant only in the first stage ($p = .003$) and was marginally significant in the second stage ($p = .051$). This larger difference in the first than in the second stage was predictable, because the two models use different equations to update the first-stage values but exactly the same equation in the second stage. The reason why the second-stage choice probability was also marginally better in the EA model is that the EA model estimates the second-stage choice probabilities relatively independently from the estimation of the first-stage choice probability, as compared with the parallel-learning model. Although both models use a chosen second-stage value in the first-stage value update by SARSA (see Eq. 2), the parallel-learning model gives an additional role to the second-stage values in the estimation of the first-stage model-based values (see Eq. 5). Thus, the parameters used for the second-stage choice probabilities must be adjusted more strongly to fit the first-stage choice probabilities in the parallel-learning model than in the EA model. As a result, the second-stage choice probabilities deteriorated in the parallel-learning model compared with the EA model.

Table 5 Estimated parameter values and negative log likelihood (–*LL*) for the parallel-learning model and the EA–forgetting model with default-value parameter (EA–FD model) model

Percentile	$a1$	$a2$	$\beta1$	$\beta2$	λ	p	w	μ	$-LL$
Parallel-Learning Model									
25	.00	.12	2.26	2.26	.14	.04	.04		275.0
50	**.22**	**.50**	**4.93**	**2.84**	**.35**	**.13**	**.52**		**320.8**
75	.47	.74	8.26	4.88	.70	.25	.70		358.5
EA–FD Model									
25	.01	.19	2.40	4.87	.19	.03	.02	.00	262.3
50	**.14**	**.26**	**5.84**	**7.47**	**.41**	**.11**	.60	**.11**	**316.5**
75	.29	.38	9.98	10.31	.89	.17	.88	.31	349.4

Here, we examine which aspects of the EA model caused the better fits in the first stage. For this purpose, it is useful to reduce the multiple updating equations regarding the first-stage value update to one equation. For the sake of simplicity, $a1$ refers to the chosen action, and $a2$ refers to the unchosen action in the first stage ($s1$). In the second stage ($s2$), $a1$ refers to the chosen action, $a2$ refers to the unchosen action, and $a3$ and $a4$ refer to the actions in the unvisited state. In both models, the chosen actions are first updated by Eq. 2 in the first stage and by Eq. 3 in the second stage. Hereafter, the bold letters $\boldsymbol{Q}(\boldsymbol{s}1, \boldsymbol{a}1)$ and $\boldsymbol{Q}(\boldsymbol{s}2, \boldsymbol{a}1)$ refer to the model-free values already updated by these equations. After this update, in the parallel-learning model, Eqs. 8, 5, and 6 are used to update the chosen action value, and Eqs. 5 and 6 are used to

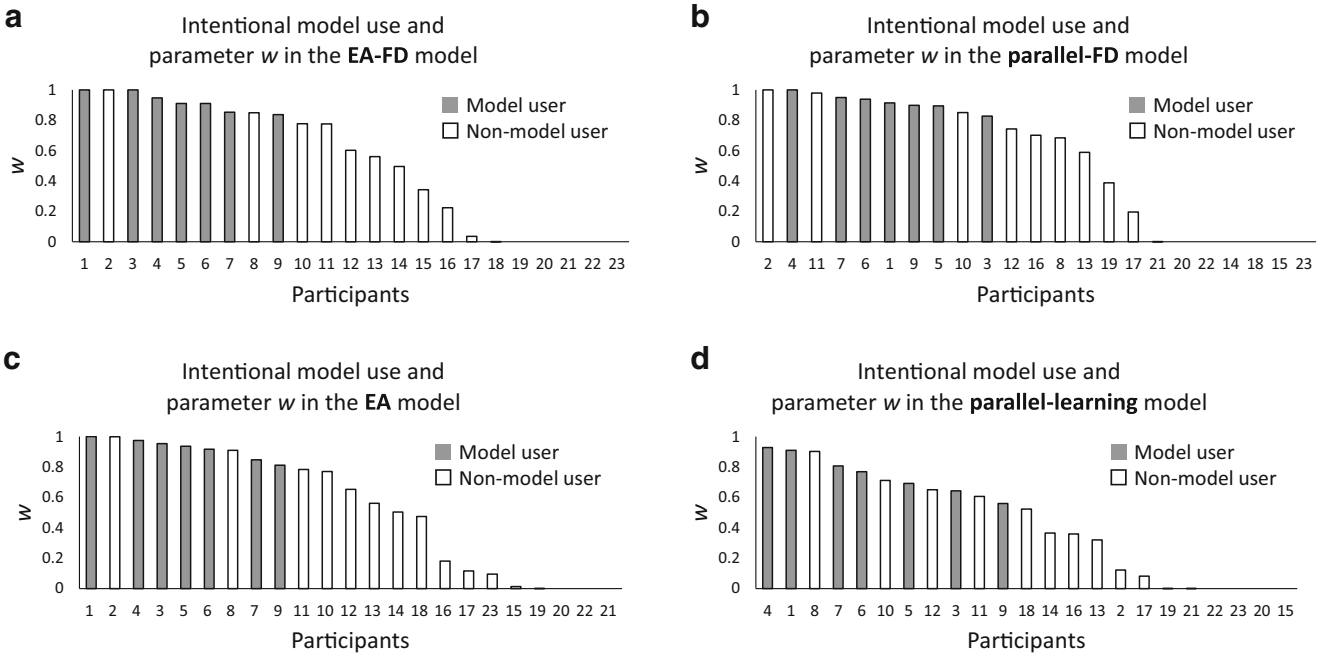

Fig. 6 Relations between the participants' intentional model use and a parameter that relates to the model-based contributions in the four models: (A) w in the EA–FD model, (B) w in the parallel–FD model, (C) w in the EA model, and (D) w in the parallel-learning model are shown. The participants are ordered by parameter value, and the numbers of participants in panels B, C, and D correspond to the number in panel A. The gray bars indicate that the participants reported that they used the transition model ("Model user"), and the white bars indicate that participants reported that they did not use the transition model ("Non-model user").

update the unchosen action value. These equations are reduced as follows:

$$Q_{\mathrm{NET}}(s1,\ a1) \leftarrow w[T\cdot\max(\boldsymbol{Q}(\boldsymbol{s}2,\ \boldsymbol{a}1),\ Q(s2,\ a2)) + (1-T)\cdot\max(Q(s2,\ a3),\ Q(s2,\ a4))] + (1-w)[\boldsymbol{Q}(\boldsymbol{s}1,\ \boldsymbol{a}1) + \lambda(r-\boldsymbol{Q}(\boldsymbol{s}1,\ \boldsymbol{a}1))], \tag{15}$$

$$Q_{\mathrm{NET}}(s1,\ a2) \leftarrow w[(1-T)\cdot\max(\boldsymbol{Q}(\boldsymbol{s}2,\ \boldsymbol{a}1),\ Q(s2,\ a2)) + T\cdot\max(Q(s2,\ a3),\ Q(s2,\ a4))] + (1-w)Q(s1,\ a2), \tag{16}$$

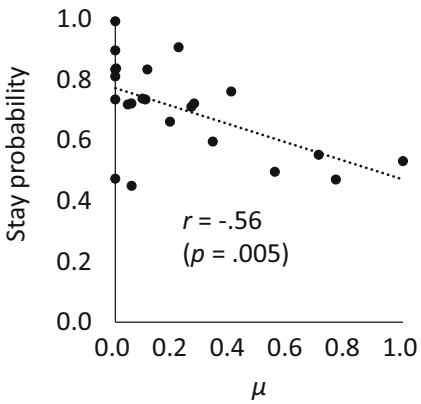

Fig. 7 Relation between the average stay probabilities of all trials at the first stage and the default-value parameter μ of the EA–FD model. As predicted, these variables show a negative correlation, suggesting that the participants who had a high default value tended to adopt exploratory behavior.

where Q is equal to the model-free value estimated in the previous trial, and Q_{NET} is the ultimate updated value after the experience of a trial. T and r represent the immediately experienced transition probability and the outcome, respectively. In the EA model, the net values are updated using Eq. 9 for the chosen action and Eq. 10 for the unchosen action as follows:

$$Q_{NET}(s1,\ a1) \leftarrow w[\boldsymbol{Q}(\boldsymbol{s}1,\ \boldsymbol{a}1) + \lambda T(r-\boldsymbol{Q}(\boldsymbol{s}1,\ \boldsymbol{a}1))] + (1-w)[\boldsymbol{Q}(\boldsymbol{s}1,\ \boldsymbol{a}1) + \lambda(r-\boldsymbol{Q}(\boldsymbol{s}1,\ \boldsymbol{a}1))], \tag{17}$$

$$Q_{NET}(s1,\ a2) \leftarrow w\,[Q(s1,\ a2) + \lambda(1-T)(r-Q(s1,\ a2))] + (1-w)Q(s1,\ a2). \tag{18}$$

By comparing the equations between the two models, it becomes clear that the difference between the two models appeared in the model-based part (the term multiplied by w) and not in the model-free part [the term multiplied by $(1-w)$].

Regarding the model-based parts, two candidate factors might have caused the fit to be better in the EA than in the parallel-learning model. First, these two models have differences in the weight of the immediate experience. The EA model has a structure that directly reflects the interaction of the transition probability (T) and reward outcome (r) to the value update for the next choice. In contrast, in the parallel-learning model this effect is diminished. Although the immediate reward outcome is reflected in $\boldsymbol{Q}(\boldsymbol{s}2, \boldsymbol{a}1)$, the max function

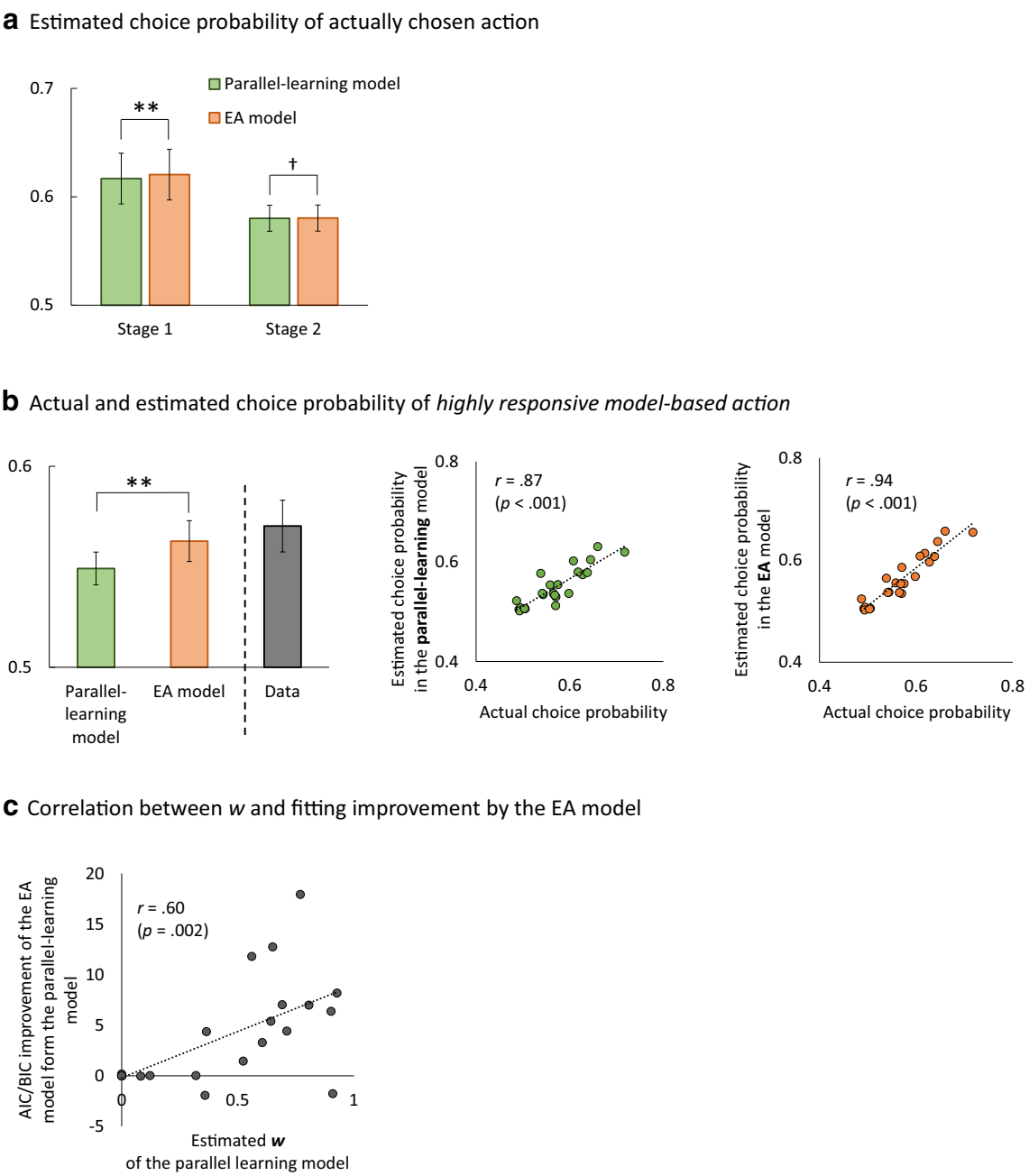

Fig. 8 Characteristics of the EA model compared with the parallel-learning model. (A) Average estimated choice probabilities of the actually chosen option in each trial in the parallel-learning model and the EA model. The EA model had greater predictability, especially in the first stage. (B) The left panel shows the average estimated and actual first-stage choice probabilities of *highly responsive model-based actions*, which are defined in each trial on the basis of the transition and outcome of the previous trial. The EA model predicted more model-based actions. The middle and right panels show the correlations between the actual and estimated choice probabilities of highly responsive model-based actions (middle panel, for the parallel-learning model; right panel, for the EA model). (C) Correlation between the estimated w in the parallel-learning model and AIC/BIC improvement of the EA model over the parallel-learning model. The greater the value of w, the greater the improvement in fit of the EA model. $^{\dagger}p < .10$, $^{**}p < .005$.

sometimes prevents it from affecting the Q_{NET}, although the prevention does not always lead to a nonoptimal choice. The influence of the immediate outcome on the model-based part depends on the comparison with the estimated value of the unchosen option in the same state. To confirm that the EA model indeed made choice predictions that reflected the interaction between the immediately experienced transition and outcome more sensitively than the parallel-learning model, we calculated, for each model, the average estimated first-stage choice probability of the model-based action defined for each trial, based only on the transition and outcome of the previous trial. Here, we call these the *highly responsive model-based*

actions, which concretely refer to an action that is identical to the previously chosen action after a "common-transition and reward" trial or a "rare-transition and no-reward" trial and an action that is identical to the previously unchosen action after "common-transition and no-reward" or "rare-transition and reward." A paired *t* test was conducted on the estimated choice probabilities of highly responsive model-based actions by the two models (the parallel-learning model: $M = .549$, $SE = .008$; the EA model: $M = .563$, $SE = .010$). This comparison revealed that the EA model predicted this type of model-based action with higher probabilities than the parallel-learning model [$t(22) = 3.77$, $p = .001$, $d = 0.79$; see the left panel of Fig. 8B]. To confirm that this effect was not simply caused by the difference in estimated *w* in the two models (19 of 23 participants showed a higher *w* in the EA model than in the parallel-learning model; see also Table 5), two more *t* tests were conducted on the estimated choice probabilities of highly responsive model-based actions from the two models using the same *w* values and optimized values for all other parameters. Concretely, one test compared the estimated choice probabilities of the two models using the best-fitting *w* values for the parallel-learning model (the parallel-learning model: $M = .549$, $SE = .008$; the EA model: $M = .561$, $SE = .010$); the other test compared the estimated probabilities of the two models using the best-fitting *w* values for the EA model (the parallel-learning model: $M = .552$, $SE = .009$; the EA model: $M = .563$, $SE = .010$). Both results again revealed significantly higher probabilities of highly responsive model-based actions in the EA model [$t(22) = 3.61$, $p = .002$, $d = 0.75$; $t(22) = 3.01$, $p = .006$, $d = 0.63$] and supported the claim that the EA model has a structure that tends to produce highly responsive model-based actions, as compared with the parallel-learning model.

It is still unclear which of these two mechanisms—the model-based part of the parallel-learning model or of the EA model—is better suited to real choice behavior. The left panel of Fig. 8B includes the average actual choice probability of highly responsive model-based actions ($M = .570$, $SE = .013$). To see the correspondence between the estimated probabilities by the two models and the actual probability, we examined the correlations between the actual choice probability and those estimated by the parallel-learning model and the EA model. In both results, a strong positive correlation was confirmed between the actual and estimated choice probabilities, and a stronger correspondence was observed in the EA model (the parallel-learning model: $r = .87$, $p < .001$; the EA model: $r = .94$, $p < .001$; see the middle and right panels of Fig. 8B). In addition, as is shown in Fig. 8C, individual AIC/BIC improvements in the fit of overall choice behavior in the EA model from the parallel-learning model were positively correlated with their estimated *w*, showing that those who had relatively higher weight in the model-based part showed greater improvement in the EA model ($r = .60$, $p = .002$). Taken together, we may conclude that the model-based part of the EA model is better suited to actual choice behavior than is that part of the parallel-learning model, and that this difference made the EA model a better fit for the overall choice behavior according to the AIC/BIC criterion.

Another factor can diminish the model fit of the parallel-learning model. This model uses all second-stage estimated values in the model-based part to update the first-stage values. They are treated equally except for the weighting of the transition probability. However, considering the existence of the forgetting mechanism, shown in the Examination of the Forgetting Mechanism section, the second-stage values computed without this mechanism caused mismatches between the estimated second-stage values and the actual values. These mismatches, in turn, can induce noise in the model-based part and cause worse prediction of the next choice at the first stage. In this context, it is noteworthy that the fitting differences between the two models are diminished when they are combined with the FD model, which can improve the estimation of model-free values; however, the EA–FD model is still favored over the parallel–FD model (see Table 4).

In summary, the EA model has a computational structure that directly reflects the interaction of the transition and outcome in the model-based part of the Q_{NET} estimation, and this causes sensitive adjustments of the choice probabilities reflecting them. Because such adjustments may well capture the strategy adopted by the actual participants who use transition information, this leads to better fits than under the parallel-learning model. In addition, in the parallel-learning model, mismatches between the estimated second-stage values and the real subjective values can cause somewhat harmful effect on the estimation regarding the first-stage choices.

Action tendency and model fits at the individual level

The results of logistic regression analyses for each participant are shown in Table 6. These results confirmed that those who reported model use showed behavior corresponding to the model-based choice. That is, all of them showed significant interactions of previous outcome and previous transition. Table 6 also shows the best-fitting models for each participant according to AIC and BIC. This result revealed that models including the eligibility adjustment rule provided the best fit among the participants who reported model use, and the EA–FD model was supported by them relatively often. On the other hand, the F model most often provided the best fit among those who did not report use of the transition model. This might be because the parameter corresponding to the model-based component was redundant for these participants.

Discussion

In this study, we examined hypotheses regarding a mechanism that combines model-free and model-based

Table 6 Results of logistic regression analyses predicting stay probability and the best-fitting model for each participant

Participants	Model user	Results of logistic regression analyses											Best-fit Model	
		Intercept			Outcome			Transition			Outcome × Transition			AIC-based	BIC-based
		β(*SE*)	*z*	*p*	*β*(*SE*)	*z*	*p*	*β*(*SE*)	*z*	*p*	*β*(*SE*)	*z*	*p*		
1	✓	1.05 (0.14)	7.41	<.0001***	-0.07 (0.28)	-0.26	.798	0.14 (0.28)	0.51	.613	1.22 (0.56)	2.17	.030*	EA-F	EA-F
2		-0.07 (0.13)	-0.57	.567	-0.19 (0.26)	-0.75	.455	-0.30 (0.26)	-1.15	.251	-0.09 (0.52)	-0.17	.864	F	F
3	✓	0.25 (0.13)	1.99	.047*	0.35 (0.25)	1.36	.174	0.19 (0.25)	0.75	.450	1.51 (0.51)	2.96	.003**	EA-FD	EA-FD
4	✓	1.00 (0.17)	5.96	<.0001***	0.21 (0.34)	0.63	.530	1.36 (0.34)	4.06	<.0001***	3.54(0.67)	5.26	<.0001***	EA-FD	EA-F
5	✓	1.01 (0.16)	6.20	<.0001***	0.48 (0.33)	1.48	.138	1.12 (0.33)	3.43	<.001***	3.46	(0.65)5.31	<.0001***	EA-F	EA-F
6	✓	1.55 (0.22)	7.06	<.0001***	1.23 (0.44)	2.82	.005**	1.17 (0.44)	2.68	.007**	4.27 (0.88)	4.87	<.0001***	EA-FD	EA-FD
7	✓	1.05 (0.17)	6.36	<.0001***	0.69 (0.33)	2.10	.036*	0.81 (0.33)	2.44	.015*	3.48 (0.66)	5.26	<.0001***	Parallel-FD	Parallel-FD
8		2.14 (0.26)	8.18	<.0001***	0.31 (0.52)	0.59	.552	0.17 (0.52)	0.33.	742	4.08 (1.05)	3.90	<.0001***	EA-F	EA-F
9	✓	0.44 (0.13)	3.24	.001**	0.42 (0.27)	1.56	.119	0.03 (0.27)	0.10	.920	1.92 (0.54)	3.56	<.0005***	EA-FD	EA-FD
10		1.74 (0.18)	9.44	<.0001***	0.93 (0.37)	2.53	.011*	0.27 (0.37)	0.73	.466	1.87 (0.74)	2.53	.011*	EA-FD	EA-F
11		0.16 (0.12)	1.25	.213	0.45 (0.25)	1.79	.073	-0.06 (0.25)	-0.24	.809	0.82 (0.50)	1.64	.100	Parallel-FD	SARSA (λ) TD
12		5.79 (268.85)	0.02	.983	10.60 (537.70)	0.02	.984	7.60 (537.70)	0.01	.989	18.72 (1075.40)	0.02	.986	EA-FD	EA-FD
13		-0.11 (0.12)	-0.94	.349	-0.17 (0.24)	-0.71	.476	-0.07 (0.24)	-0.30	.762	-0.18 (0.48)	-0.37	.710	F	F
14		0.72 (0.14)	5.25	<.0001***	1.12 (0.27)	4.10	<.0001***	0.46 (0.27)	1.67	.094	0.72 (0.55)	1.32	.188	EA	SARSA (λ) TD
15		17.90 (2728.90)	0.01	.995	9.34 (5457.80)	0.00	.999	-9.34 (5457.80)	0.00	.999	18.67 (10915.59)	0.00	.999	F	F
16		0.02 (0.13)	0.16	.877	0.64 (0.25)	2.54	.011*	0.08 (0.25)	0.32	.746	0.21 (0.50)	0.42	.672	FD	FD
17		1.00 (0.14)	7.03	<.0001***	0.18 (0.28)	0.65	.518	0.06 (0.28)	0.20	.841	0.10 (0.57)	0.18	.860	F	F
18		2.14 (0.21)	10.27	<.0001***	0.43 (0.42)	1.04	.300	0.48 (0.42)	1.15	.250	0.83 (0.83)	0.99	.321	F	F
19		1.92 (0.23)	8.23	<.0001***	1.49 (0.47)	3.19	.001**	1.12 (0.47)	2.40	.016*	2.74 (0.93)	2.93	.003**	SARSA (λ) TD	SARSA (λ) TD
20		1.36 (0.21)	6.43	<.0001***	1.07 (0.42)	2.52	.012*	-0.83 (0.42)	-1.95	.051	-1.31 (0.85)	-1.54	.123	F	F
21		6.72 (308.17)	0.02	.983	9.81 (616.33)	0.02	.987	-7.30 (616.33)	-0.01	.991	-13.16 (1232.66)	-0.01	.991	FD	FD
22		-0.25 (0.14)	-1.81	.070	0.40 (0.28)	1.41	.159	0.34 (0.28)	1.20	.230	-0.29 (0.56)	-0.51	.610	F	F
23		1.42 (0.16)	9.07	<.0001***	-0.31 (0.31)	-1.00	.315	0.33 (0.31)	1.06	.290	0.26 (0.62)	0.42	.677	F	F

Participants are ordered by their estimated values of *w* in the EA–FD model. The participants who reported model use are marked with ✓ as Model user. Logistic regression coefficients indicate the influences of the outcome of the previous trial and the transition of the previous trial, along with their interaction. At the right of the table, the best-fitting model is shown for each participant, based on the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

components, and we also tested whether the introduction of the forgetting mechanism improved the fit to the data. The EA model assumes a sequential value-updating process in which the model-free learning system is a core mechanism; in this mechanism, the environmental model is used to adjust the degree of updating. The EA model is a natural extension of traditionally used SARSA (λ) TD learning, and it showed better predictability of choice behavior than the original model, which assumes that the parallel calculations of two types of independent values are based on two distinct learning systems (Daw et al., 2011; Daw et al., 2005). In addition, the application of the F model and its variation, the FD model, improved the data fit by updating the unchosen action values. Thus, we provided a hybrid EA–FD model that assumed model-based adjustments in eligibility trace updating and introduced the forgetting mechanism, which resulted in the best data fit.

Some theoretical and experimental models are in agreement with the ideas of the EA model. First, regarding the idea of the EA model that both model-free and model-based values can be integrated in single update, the same idea is seen in the experience-weighted attraction (EWA) learning model proposed by Camerer and Ho (1999). This model has widely succeeded in capturing the features of human choice dynamics in multiperson noncooperative games. The EWA and EA models have different computational structures because they treat completely different *models*. However, the core mechanism in the model-based component is the same between the two models, in that the unchosen action values are updated according to the expected rewards if the actions are chosen in the current trial. Second, regarding the assumption of the EA model that transition probabilities attenuate the degree of updating, in the field of classical conditioning the agent's attention to the conditioned stimulus (CS) is considered to adjust the learning rate (MacKintosh, 1975; Pearce & Hall, 1980; Rescorla & Wagner, 1972). Particularly in Mackintosh's model, the degree of value updating is proportional to the degree of the CS's relative ability to predict an outcome. This model is close to the EA model's assumption that a well-established transition leads to more updating. In the field of neuroscience, a similar framework has been proposed in which the learning rate depends on either the reliability of the environment (Daw, Courville, & Touretzky, 2006) or the accuracy of reward prediction (Bertin, Schweighofer, & Doya, 2007). Thus, the EA model is a natural expansion of these theories to SARSA (λ) TD learning.

Biologically, the eligibility trace can be implemented in the brain as sustained firing in reverberating circuits at the synaptic level (Florian, 2007; Houk, Adams, & Barto, 1995). Some studies have reported the validity of the eligibility accumulation beyond the last episode or trial (Bogacz, McClure, Li, Cohen, & Montague, 2007), which is different from the original assumption of the eligibility trace but has led to the idea for this article. In addition, the neural circuit of model-based decision making proposed by Friedrich and Lengyel (2016) might be related to both the EA model and the parallel-learning model, because it provided the neural basis for value updating using transition probabilities. It is also noteworthy that Daw et al. (2011) reported that the striatum RPE signal reflected both model-free and model-based valuations. This observation is consistent with our framework of the EA model, in which the RPEs include both model-free and model-based components. Although it still remains unclear what type of mechanism exists in the neural RPEs to reflect both systems, the EA model provides a likely and testable structure to capture their activation. We hope that the proposed computational models will be examined at the implementation level in future studies.

To clarify the characteristics of the EA model, it might be useful to note its differences from the other models. First, the EA model realizes model-based value updates by adding two features to SARSA (λ) TD learning. One is a weight parameter that controls the effects of the model-free and model-based systems in eligibility traces, and the other is a similar eligibility trace rule for the unchosen action. We found that both of them contributed to a better fit to the data (Supplementary Text 3). Second, although both the EA model and the parallel-learning model can treat model-based decision making, there are critical differences between them at the algorithm level. The EA model uses only one learning mechanism, whereas the parallel-learning model uses two learning mechanisms and an additional mechanism for the integration of values. With respect to the model-based component, the EA model relies on the eligibility trace and uses the reward outcome to update the values in a model-based manner, whereas the parallel-learning model relies on the Bellman optimality equation, and the maximum values from the second stage are used to update the values in a model-based manner. The parallel-learning model may be useful if the full task structure is obvious and available, although it is also more costly, because it uses the full information about transition and related state–action values for value updating (Fig. 9, left). In contrast, the EA model uses only partial information—the reward outcome and transition model related to the most recently experienced state. The EA model receives the influence of selection bias but incurs less of a cost (Fig. 9, right). Thus, the two models show the trade-off between the degree of model use and ease of calculation.

In the Results section, we have noted the possible factors that may explain why the EA model showed a better fit than the parallel-learning model. It is notable that the stronger influence of the interaction of the immediately experienced transition and the outcome in the EA model suited the choice behavior in the present task. In addition, we noted that in the model-based part of the parallel-learning model, the mismatches between the estimated second-stage values and the real subjective values became a factor that worsened the fit of the model. Although the present data supported the EA model due to this mix of factors, it is possible that human and animal

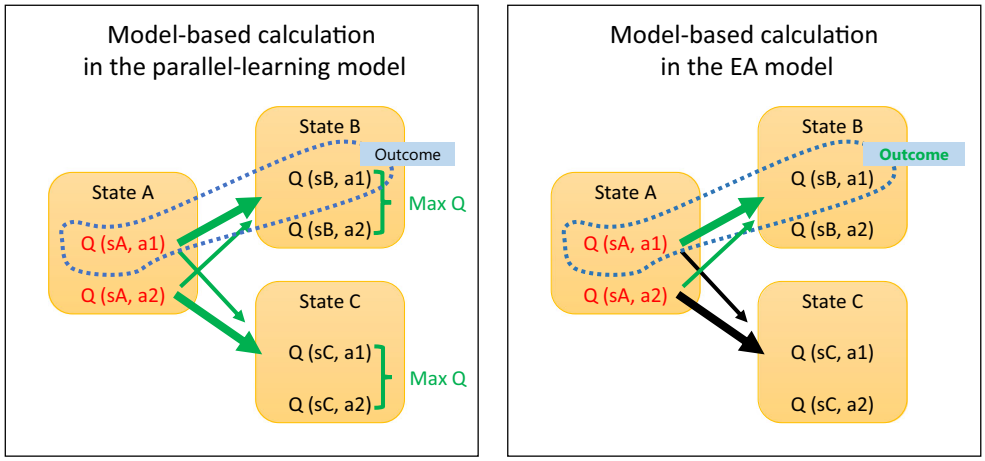

Fig. 9 Schematics of model-based calculations in the parallel-learning model (left panel) and the EA model (right panel). The dotted lines indicate examples of an experienced transition. The first-stage model-based calculations are performed for the color Q values. The parallel-learning model uses all transitions as well as the two maximum values in each state of the second stage, marked in cooler color. The EA model uses the transition relating to the visited second-stage state and the reward outcome, also marked in cooler color.

strategies of model use would change under different levels of complexity or certainty of the available internal models. Further research will be needed to clarify this point.

One of the limitations of the present models is that they treat the relative contributions of the model-free and model-based systems as constant for the duration of the task. This approach is suitable to the present purpose because it provides the total choice tendency of each person. However, as Table 1 revealed, the relative weights of the systems seem to change over time. Therefore, developing a model that can capture these dynamic changes will be valuable for describing a metasystem that determines which system tends to be used in certain situations.

We also showed that an assumption regarding the decay of unchosen option values explained the data better than not hypothesizing it, which is another important mechanism. It is reasonable to believe that the action–reward association decays over time if it is not experienced. To the best of our knowledge, this report has been the first to show the usefulness of the forgetting mechanism to predict human choice behavior. However, it might still be unreasonable to assume that all values always decay to zero. Considering that many studies have demonstrated that uncertainty prompts exploration behavior (Badre, Doll, Long, & Frank, 2012), it is supposed that relatively unchosen options might induce some vague expectation related to their increased uncertainty. Thus, we additionally introduced the concept of a default value to the F model. The FD model can express the phenomena that unchosen actions obtain either an increment of value or a decrement of value by regressing to the default-value parameter μ. In this study, the FD model showed a slightly improved fit relative to the F model. Interestingly, the degree of improvement was relatively high in the parallel-learning model, which, compared with the other models, gives greater weight to the accuracy of the second-stage model-free value estimations in the model fit.

As far as we know, there are two classifications of methods that promote exploration in reinforcement learning: increasing randomness of choices (e.g., e-greedy, Boltzmann exploration by Sutton & Barto, 1998) or increasing value of the unchosen options (e.g., the exploration bonus of Dyna-Q by Sutton, 1990, prioritized sweeping by Moore & Atkeson, 1993). In the former method, exploration occurs in undirected manner, whereas the latter method assumes that the exploration behavior is prompted because options that have not been selected recently obtain some value because of their uncertainty. Neurological evidence has been provided for both types of exploration (Badre et al., 2012; Humphries, Khamassi, & Gurney, 2012; Krebs, Schott, Schütze, & Düzel, 2009). Parameter μ in our models is in line with the latter method, because the high value for parameter μ tends to raise unchosen option values followed by a temporal choice shift to them, and this parameter has an advantage that can adapt relatively easily to traditional TD learning.

Here we proposed a hybrid EA–FD model and showed its relative goodness of fit among all the other models compared in this study. The parameters that are used in this hybrid model work as useful indicators that capture important characteristics of choice behavior. First, parameter w from the EA model worked better as a predictor of model-based actions than the parameter w from the parallel-learning model. The results of the yes–no question regarding the intentional use of the transition model confirmed that w in the eligibility trace adjustment models was a better predictor of intentional model use. It is also important to note that if λ is very low, the value of w is meaningless; otherwise, it would become a credible index of model use. Second, parameter μ from the FD model enables the capture of individual differences in exploration behavior, as shown by a negative correlation with the stay probability.

In conclusion, the hybrid model and its subcomponent models that were proposed here provide a renewed framework to understand the learning process with model use. The proposed hybrid model is a simple and reasonable extension of the traditional SARSA (λ) TD framework and has a reduced computational cost, relative to a competing model that assumes a dual value-updating process. Simultaneously, this model produced both better empirical predictions and useful indexes of individual differences in model use and exploration behavior. The proposed model will promote better understanding of the learning process at the behavioral and neural levels, serving research that connects learning strategy with other cognitive functions or mental disorders.

Author note This work was supported by a Grant-in-Aid for Japan's Society for the Promotion of Science Fellows.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716–723. doi:10.1109/TAC.1974.1100705

Badre, D., Doll, B. B., Long, N. M., & Frank, M. J. (2012). Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. *Neuron, 73,* 595–607. doi:10.1016/j.neuron.2011.12.025

Barraclough, D. J., Conroy, M. L., & Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience, 7,* 404–410. doi:10.1038/nn1209

Bertin, M., Schweighofer, N., & Doya, K. (2007). Multiple model-based reinforcement learning explains dopamine neuronal activity. *Neural Networks, 20,* 668–675. doi:10.1016/j.neunet.2007.04.028

Bogacz, R., McClure, S. M., Li, J., Cohen, J. D., & Montague, P. R. (2007). Short-term memory traces for action bias in human reinforcement learning. *Brain Research, 1153,* 111–121. doi:10.1016/j.brainres.2007.03.057

Camerer, C., & Ho, T. (1999). Experience-weighted attraction learning in normal form games. *Econometrica, 67,* 827–874. doi:10.1111/1468-0262.00054

Curtis, C. E., & Lee, D. (2010). Beyond working memory: The role of persistent activity in decision making. *Trends in Cognitive Sciences, 14,* 216–222. doi:10.1016/j.tics.2010.03.006

Daw, N. D., Courville, A. C., & Touretzky, D. S. (2006). Representation and timing in theories of the dopamine system. *Neural Computation, 18,* 1637–1677. doi:10.1162/neco.2006.18.7.1637

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,* 1204–1215. doi: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,* 1704–1711. doi:10.1038/nn1560

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society B, 308,* 67–78. doi:10.1098/rstb.1985.0010

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron, 80,* 312–325. doi:10.1016/j.neuron.2013.09.007

Florian, R. V. (2007). Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity. *Neural Computation, 19,* 1468–1502. doi:10.1162/neco.2007.19.6.1468

Friedrich, J., & Lengyel, M. (2016). Goal-directed decision making with spiking neurons. *Journal of Neuroscience, 36,* 1529–1546. doi:10.1523/JNEUROSCI.2854-15.2016

Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General, 143,* 182–194. doi:10.1037/a0030844

Ghalanos, A., & Theussl, S. (2015). Package Rsolnp: General non-linear optimization using augmented Lagrange multiplier method (R package version 1.16). Retrieved from https: /cran.r-project.org/web/package=Rsolnp

Gillan, C. M., Otto, A. R., Phelps, E. A., & Daw, N. D. (2015). Model-based learning protects against forming habits. *Cognitive, Affective & Behavioral Neuroscience, 15,* 523–536. doi:10.3758/s13415-015-0347-6

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,* 585–595. doi:10.1016/j.neuron.2010.04.016

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 249–270). Cambridge: MIT Press.

Humphries, M. D., Khamassi, M., & Gurney, K. (2012). Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Frontiers in Neuroscience, 6,* 9. doi:10.3389/fnins.2012.00009

Ito, M., & Doya, K. (2009). Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *Journal of Neuroscience, 29,* 9861–9874. doi:10.1523/JNEUROSCI.6157-08.2009

Kahneman, D. (2010). *Thinking, fast and slow.* New York: Farrar, Straus & Giroux.

Katahira, K. (2015). The relation between reinforcement learning parameters and the influence of reinforcement history on choice behavior. *Journal of Mathematical Psychology, 66,* 59–69. doi:10.1016/j.jmp.2015.03.006

Krebs, R. M., Schott, B. H., Schütze, H., & Düzel, E. (2009). The novelty exploration bonus and its attentional modulation. *Neuropsychologia, 47,* 2272–2281. doi:10.1016/j.neuropsychologia.2009.01.015

MacKintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82,* 276–298. doi:10.1037/h0076778

Moore, A. W., & Atkeson, C. G. (1993). Prioritized sweeping: Reinforcement learning with less data and less real time. *Machine Learning, 13,* 103–130. doi:10.1023/a:1022635613229

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, 24,* 751–761. doi:10.1177/0956797612463080

Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A., & Daw, N. D. (2013). Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences, 110,* 20941–20946. doi:10.1073/pnas.1312011110

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 87,* 532–552. doi:10.1037/0033-295X.87.6.532

Redish, A. D., Jensen, S., & Johnson, A. (2008). A unified framework for addiction: Vulnerabilities in the decision process. *Behavioral and Brain Sciences, 31,* 415–437. doi:10.1017/S0140525X0800472X

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.),

Classical conditioning II: Current Research and Theory (pp. 64–99). New York: Appleton-Century-Crofts.

Rummery, G., & Niranjan, M. (1994). *On-line Q-learning using connectionist systems (Technical Report CUED/F-INFENG/TR 166)*. Cambridge: Cambridge University.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464. doi:10.1214/aos/1176344136

Sebold, M., Deserno, L., Nebe, S., Nebe, S., Schad, D. J., Garbusow, M., … & Huys, Q. J. (2014). Model-based and model-free decisions in alcohol dependence. *Neuropsychobiology*, *70*, 122–131. doi:10.1159/000362840

Skatova, A., Chan, P. A., & Daw, N. D. (2013). Extraversion differentiates between model-based and model-free strategies in a reinforcement learning task. *Frontiers in Human Neuroscience, 7,* 525. doi:10.3389/fnhum.2013.00525

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,* 914–919. doi:10.1016/j.neuron.2013.08.009

Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting on the basis of approximating dynamic programming. In B. W. Porter & R. J. Mooney (Eds.), *Proceedings of the Seventh International Conference on Machine Learning* (pp. 216–224). San Francisco: Morgan Kaufmann.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge: MIT Press.

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. New York: Macmillan.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review, 55,* 189–208. doi:10.1037/h0061626

Voon, V., Derbyshire, K., Rück, C., Irvine, M. A., Worbe, Y., Enander, J., … & Bullmore, E. T. (2015). Disorders of compulsivity: A common bias toward learning habits. *Molecular Psychiatry*, *20*, 345–352. doi:10.1038/mp.2014.44

Wunderlich, K., Smittenaar, P., & Dolan, R. J. (2012). Dopamine enhances model-based over model-free choice behavior. *Neuron, 75,* 418–424. doi:10.1016/j.neuron.2012.03.042