Model-based estimation of subjective values using choice tasks with probabilistic feedback

Kentaro Katahira,[1,2,3*] Shoko Yuki,[4] Kazuo Okanoya[2,4]

Institutional affiliations:

1 Center for Evolutionary Cognitive Sciences, The University of Tokyo, Meguro, Tokyo, 153-0041, Japan.

2 Behavior and Cognition Joint Research Laboratory RIKEN Brain Science Institute, Wako, 351-0198, Saitama, Japan.

3 Department of Psychology, Graduate School of Informatics, Nagoya University, Nagoya, Aichi, 464-8601, Japan.

4 Graduate School of Arts and Sciences, The University of Tokyo, Tokyo, 153-0041, Japan.

* Corresponding author:

Kentaro Katahira

Department of Psychology, Graduate School of Informatics, Nagoya University, Nagoya, Aichi, 464-8601, Japan

E-mail: katahira@lit.nagoya-u.ac.jp

Abstract

Evaluating the subjective value of events is a crucial task in the investigation of how the brain implements the value-based computations by which living systems make decisions. This task is often not straightforward, especially for animal subjects. In the present paper, we propose a novel model-based method for estimating subjective value from choice behavior. The proposed method is based on reinforcement learning (RL) theory. It draws upon the premise that a subject tends to choose the option that leads to an outcome with a high subjective value. The proposed method consists of two components: (1) a novel behavioral task in which the choice outcome is presented randomly within the same valence category and (2) the model parameter fit of RL models to the behavioral data. We investigated the validity and limitations of the proposed method by conducting several computer simulations. We also applied the proposed method to actual behavioral data from two rats that performed two tasks: one manipulating the reward amount and another manipulating the delay of reward signals. These results demonstrate that reasonable estimates can be obtained using the proposed method.

Keywords: Subjective value, Model-based estimation, Reinforcement learning, Choice behavior, Random feedback

Introduction

Subjective values (or utilities) assigned to positive or negative events by living systems in general differ from their objective value (e.g., amount of money). Rewards with larger amounts and less delay are basically preferable, but the subjective values are not linearly related to objective, measurable values such as amount and delay (e.g., Kahneman & Tversky, 1979). Investigations into the valuation systems of living systems have gained significant attention in various fields such as psychology, neuroscience, and psychiatry (e.g., Rangel et al., 2008; O'Doherty, 2014). For example, some psychiatric disorders (e.g., depression) can be characterized by altered subjective values (for a review, see Chen et al., 2015). Thus, the validity of animal models of a psychiatric disorder may be evaluated based on the subjective values of the subjects.

Traditional econometric methods of estimating subjective value cannot be applied to animals because they rely on verbal instruction (e.g., Kahneman & Tversky, 1979; Kable & Glimcher, 2007). Several methods have been used to estimate subjective values or preferences in animal studies. A typical procedure is to have the subjects learn the relationship between a specific response (e.g., pressing a lever or remaining in a specific location) and the resulting outcome, from which the subjective value is measured (e.g., Green and Estle, 2003). This approach requires sufficient training so that the animals learn the relationships among all of the items and the choice behavior reaches the steady state. Another common method utilizes the law of how animals distribute their responses depending on the reinforcement, i.e., the matching law (Miller, 1976). Both approaches rely on the pairwise comparison of preferences for two items. Thus, to measure the subjective values of several items, the researcher must examine preferences for multiple combinations of items. This method requires much time and sophisticated experimental considerations.

In the present study, we propose a novel method for estimating subjective values especially from animal behaviors using novel behavioral tasks and reinforcement learning (RL) model-based analysis. RL is usually formulated as an algorithm that attempts to maximize the total reward that a decision-maker can obtain. Recent studies, however, have begun to use the RL framework to model human behavior that does not necessarily lead to reward maximization (Neiman and Loewenstein, 2011; Shteingart and Loewenstein, 2014). For example, basketball players tend to choose to make a 3-point shot immediately after an experience of success; however, this dependence decreases the success rate. This choice behavior is modeled using an RL model (Neiman and Loewenstein, 2011, 2014). Additionally, RL models have been important data analysis tools for experiments involving value-based, decision-making tasks (Corrado and Doya, 2007; O'Doherty et al., 2007; Daw, 2011).

Standard RL theory assumes that there is an increased probability of choosing an option that has been reinforced in the immediate past. The magnitude of dependence decays exponentially with the passage of time (trials) (Katahira, 2015). The main idea of the proposed method is to utilize this property. RL theory also assumes that the larger the subjective value of an outcome, the more frequently the decision-maker repeats the same choice in the immediate future. Using the model parameter fit of RL models to trial-by-trial data, one can effectively estimate the subjective values of various decision-outcomes. The proposed method takes advantage of transient, trial-level dynamics of behavior, whereas other conventional methods examine only steady-state behavior. By using the transient effect of outcome on subsequent choices, it can estimate the value of multiple types of outcomes in a single experiment consisting of only two options.

The remainder of this paper is organized as follows. First, we describe the proposed

method, which consists of the novel experimental design and RL model-based analysis. Next, we examine the validity and several properties of the proposed method based on synthetic data. We then apply the proposed method to actual behavioral data from rats. In the demonstration, we examined the rats' subjective values regarding amounts of rewards and delays of the reward (and no-reward) signal. Finally, we discuss the advantages and limitations of the proposed method.

Proposed method

The proposed method consists of novel experimental tasks and RL model-based trial-by-trial analysis of behavioral data. In the following, we describe the basic task structure, the RL models, and the statistical analysis procedure.

Basic task properties

The proposed choice task has the following structure. First, the outcome of choice (decision-outcome) should contain at least one appetitive outcome. This point is particularly crucial in animal studies to ensure that there is an incentive that will motivate animals to engage in the task. Second, the task must have contingency between the valence of outcome (appetitive, neutral, or aversive) and the animal's choice, as in conventional decision-making tasks. Contingency is required because it provides the animal with an incentive to learn the value of its actions. Within the outcome valence, however, the outcomes may be randomly chosen, irrespective of the animal's choice. Third, the contingency between choice and outcome valence must change during the task so that the animals' choice does not converge with the same option. Although a class of RL models, as employed in the present study, does not converge to

deterministic choice behavior, actual animals' choice behavior often becomes deterministic if they are exposed to the constant contingency condition. This also occurs with other RL models, such as actor-critic learning (Sakai & Fukai, 2008). Using dynamic changing contingency prevents subjects from converging to a deterministic choice behavior, which is less useful for estimating subjective values. Specific task examples are presented in the following simulation and in experiments using rats.

Reinforcement learning model

In this section, we introduce RL models (Sutton & Barto, 1998). Specifically, we consider several variants of Q-learning with a single state (Watkins & Dayan, 1992), which is the model most commonly used in the model-based analysis of choice behavior. The model assigns each action an action value denoted as $Q_i(t)$, where *i* is the index of the action and *t* is the index of the trial. In the common setting, the initial action values are set to zero, i.e., $Q_i(1) = 0$ for all *i*. Based on the outcome, the action values for the action *i* are updated as follows:

$$Q_i(t+1) = Q_i(t) + \alpha_L\big(r(t) - Q_i(t)\big), \quad \textbf{(1)}$$

where α_L is the learning rate that determines the degree to which the model updates the action value depending on the reward prediction error, $r(t) - Q_i(t)$. The range of the learning rate is restricted between 0 and 1. For the unchosen action option *j* (*i*≠*j*), the action value is updated as follows:

$$Q_j(t+1) = (1 - \alpha_F)Q_j(t), \quad \textbf{(2)}$$

where α_F is the forgetting rate (Erev & Roth, 1998; Ito & Doya, 2009). In a typical RL model, the action value of the unchosen option is not updated. This convention can be represented by setting $\alpha_F = 0$. We call this the standard Q-learning model. The model with an identical learning

rate and forgetting rate ($\alpha_L = \alpha_F$) is called Q-learning with forgetting (F-Q-learning). We also consider the model in which the learning rate and forgetting rate are allowed to differ ($\alpha_L \neq \alpha_F$) and the learning rate can take a non-zero value. This is called Q-learning with differential forgetting (DF-Q-learning).

We suppose that there are at least two different types of decision-outcomes. To estimate the subjective value of the outcomes, we propose two methods to represent the subjective value. One method is non-parametric and assigns a single parameter for each outcome type. The other method uses a parametric function, which represents subjective values as a function of the objective quantity of outcomes (e.g., amount, or delay).

In the non-parametric method, we set the value of outcome *m* (*m* is the index of decision-outcome) as κ_m, and $r(t) = \kappa_m$ if outcome *m* appears at trial *t* (Katahira et al., 2011, 2014). For example, the index of outcome indicates the reward amount (in the following simulations and Task 1 of the rat experiments) and the delay in reward (no-reward) signals (Task 2 of the rat experiments). We denote the value of the reference outcome (such as absence of reward) as κ_0. This is often fixed at zero, but we also examine the case in which κ_0 is estimated as a free parameter. We call $r(t)$ and thus κ_ms the reward value. We assume that the reward value reflects the subjective value of the corresponding outcome.

When the outcome types are quantifiable, one can parameterize the value function. In the parametric method, for example, the reward value may be parameterized by the prospect utility function, which is a power function of the reward amount (e.g., Ahn et al., 2008), as

$$r(t) = x(t)^{\alpha_U}, \tag{3}$$

where $x(t)$ is the amount of reward given in trial *t*. When $\alpha_U = 0$, the reward values are

homogeneous over different amounts. If $\alpha_U = 1$, the reward value function is a linear function of the actual reward amount. Intermediate values of α_U indicate a non-linear relationship between the actual reward amount and subjective reward values. Other examples of parametric value functions (a hyperbolic function and a polynomial function) are given in the application in the rat experiment (Task 2). Based on the set of action values, the model computes the probability of choosing option *i* at trial *t* using the softmax function:

$$P(a(t) = i) = \frac{\exp(\beta \cdot Q_i(t) + \varphi \cdot c_i(t))}{\sum_{j=1}^{K} \exp(\beta \cdot Q_j(t) + \varphi \cdot c_j(t))}, \quad (4)$$

where β is the inverse temperature parameter that determines the sensitivity of the choice probabilities to differences in values. *K* represents the number of possible actions. The term $\varphi \cdot c_j(t)$ is a choice autocorrelation factor. The parameter φ controls the tendency to repeat (when the parameter is positive) or avoid (when the parameter is negative) recently chosen options. The choice trace $c_j(t)$ is set to 1 if the subject chooses option *j* in the previous trial (at trial *t*-1) and zero otherwise. Choice perseverance (repetition of a choice) is also caused by decay of an unchosen option value (forgetting) because the decreases in the value of the unchosen value increase the relative value of the chosen option (Worthy et al., 2013). However, as the choice autocorrelation factor causes the perseveration (or switching) tendency irrespective of outcomes, the effect is not identical to forgetting (Katahira, 2015). Thus, we include both components. If all initial values of Q_i are zero, the magnitude of inverse temperature β has the equivalent effect of the scale of the reward value. Formally, the transformation $r(t) \leftarrow a \cdot r(t)$ with a constant *a* has the same choice probability as the transformation $\beta \leftarrow a \cdot \beta$ (Katahira, 2015). Thus, the scale of κ_m is confounded with the magnitude of the inverse temperature. Thus, only the relative magnitude among decision-outcomes can be estimated. In the present paper, we set β = 1 as a fixed parameter in all models except for those with the prospect utility function; because the

prospect utility function given in Eq. (3) has no scaling factor (when x(t) = 1, the value is 1, irrespective of the parameter α_U), we used β as a scaling factor.

Model fit, model selection, and hypothesis testing

Parameter estimation

The model parameters are estimated based on the maximum likelihood method. Specifically, the negative log likelihood for each subject is minimized using the Matlab function "fmincon". To prevent a poor local minimum solution, the algorithm is run several times for each subject; each run is initiated from a random initial value (from the uniform distribution with the range between 0.0 and 1.0 for each parameter), and the parameter set that provides the lowest negative log likelihood is selected. In the following simulations, we empirically confirmed that it only in very rare cases was the solver "fmincon" trapped by a poor local minimum. Thus, we ran the optimization only three times, which is considered sufficient to obtain the maximum likelihood estimate. To evaluate the reliability of parameter estimates, the standard error of the parameter estimate can be estimated by computing the Hessian of the negative log likelihood at the maximum likelihood point. The Hessian is computed by using the function "hessian" in DERIVESTsuite (https://jp.mathworks.com/matlabcentral/fileexchange/13490-adaptive-robust-numerical-differentiation). The square roots of the diagonal terms of its matrix inverse correspond to one standard error for the parameter (Daw, 2011).

Model selection

To determine which model among the candidate models best describes the given

behavioral data and whether different outcomes indeed yield different subjective values, we consider two model selection criteria: Akaike's Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarz, 1978). They are defined as

$$\mathrm{AIC} = -2L + 2k, \qquad (5)$$

$$\mathrm{BIC} = -2L + k\ln(n), \qquad (6)$$

respectively, where L is the maximized value of the log likelihood function for the model, k is the number of estimated parameters in the model, and n is the number of samples (here, the total number of trials). Both criteria introduce a penalty for the number of parameters in the model. AIC is a measure of the predictive ability of a model: a smaller AIC score indicates better prediction ability. BIC approximates the model evidence, which quantifies how likely the model is for the given data set. A smaller BIC score indicates a more likely model.

Hypothesis testing

In addition to model selection based on AIC or BIC, the proposed method performs hypothesis testing (specifically, the likelihood-ratio test) when the best model contains heterogeneous reward values ($\kappa_1 \neq \kappa_2 \neq \cdots \neq \kappa_M$ for non-parametric value models, and $\alpha_U \neq 0$ for models with the prospect utility function). We compare the log likelihood of the null model (L_{null}) to that of the alternative model (L_{alt}). Here, the alternative model is selected based on AIC or BIC. The null model has homogeneous reward value parameters ($\kappa_1 = \kappa_2 = \cdots = \kappa_M$, or $\alpha_U = 0$), whereas the other model structure is identical to that of the alternative model. When the null model is the true model, the statistic $D = -2\,(L_{\mathrm{null}} - L_{\mathrm{alt}})$ obeys the chi-square distribution with the degree of freedom $M - 1$. If the null model is rejected, we consider that the subjective values

are not homogeneous among the outcomes to be tested.

Simulations

To examine the validity of the proposed method, we perform computer simulations based on the synthesized data set. One of the advantages of simulations using synthetic data compared with real behavioral data is that we can know the ground truth about the underlying subjective reward values. Thus, we can evaluate how the method works in a straightforward manner. Specifically, we address the following 7 points.

1. The validity of the estimates of subjective values (addressed in Case 1).

   – Can the proposed method recover true subjective values? How accurate are the estimates?

2. The validity of the model selection criterion (addressed in Case 1).

   – Can AIC and BIC lead to correct conclusions about the difference or equivalence of subjective values? Which criterion, AIC or BIC, is more appropriate for our purpose?

3. The robustness of the proposed method (addressed in Case 1).

   – What happens if the true model (for generating synthesized data) differs from the fitted model?

4. The validity of hypothesis testing (addressed in Case 2).

   – Is the frequency of type I error (the error in which the method detects a difference in the reward values although the actual reward values are homogeneous) suppressed to a specific range (a critical value) as intended?

5. The effect of (true) parameter values (addressed in Case 3).

   – How does the true parameter value, especially for learning rates, affect the quality of

estimates?

6. The effect of reference outcome (addressed in Case 4).

   – What happens if the reward value of reference κ_0 is also estimated as a free parameter rather than fixed to zero?

7. Effectiveness of value parameterization (addressed in Case 5).

   – When does the parameterization of reward value lead to better estimation, and when is the non-parametric value function better?

Simulation procedure

Throughout the simulations, we adopt the following procedure. First, we generate the choice data from the Q-learning models (hypothetical subjects) that perform hypothetical decision-making tasks. Each hypothetical subject performs 10 sessions; each session consists of 200 choice trials. We assume that there are two options. An option may correspond to a lever press or nose poke in a typical choice experiment using rats. One option is an advantageous (good) option, which has an 80% probability of a reward outcome and a 20% probability of the absence of a reward. Another option is a disadvantageous (bad) option, which has an 80% probability of the absence of a reward and a 20% probability of a reward. The advantageous and disadvantageous options are switched when the choice frequency of the advantageous option during the previous 10 choice trials reaches 80%. In a rewarded trial, the reward amount (a number of pellets for a rat experiment) is randomly drawn from 1, 2, or 3 pellets for Cases 1 – 4. For Case 5, the maximum number of the reward amount is varied as 3, 5, or 7.

Model for generating synthetic data

A "true model" for generating synthetic data is described for each simulation setting. In Cases 1 – 4, the reward values are set as

$$r(t) = \begin{cases} \kappa_1 \text{ if the reward amount was one at trial } t \\ \kappa_2 \text{ if the reward amount was two at trial } t \\ \kappa_3 \text{ if the reward amount was three at trial } t \\ \kappa_0 \text{ if no reward was given at trial } t \end{cases}, \quad (7)$$

where the values of the parameter κ_m and the other parameters are also described for each simulation. For Case 5, the true reward value is the prospect utility function model described in Eq. (3).

Parameter fit and model selection

For all candidate models, the model parameters are fit to each hypothetical subject's data using the maximum likelihood method as described in the Proposed Method section. For Cases 1 – 4, models with only non-parametric values are fit (listed in Table 1). For Case 5, models with both parametric and non-parametric value functions are used.

Evaluation of subject value estimates

The estimation accuracy is evaluated by computing the root mean square error (RMSE), defined as

$$\text{RMSE} = \sqrt{\frac{1}{N\,K}\sum_i \sum_m (\hat{\kappa}_{i,m} - \kappa^*_{i,m})^2} \quad (8)$$

where $\kappa^*_{i,m}$ is the true reward value of outcome *m* for subject *i*, $\hat{\kappa}_{i,m}$ is its estimate and *N* is the number of simulated subjects. For parametric value cases, $\kappa^*_{i,m}$ and $\hat{\kappa}_{i,m}$ are defined using the

true parametric function (with the true parameter) and with the estimated parametric function, respectively.

Case 1 – Estimating heterogeneous subjective values

First, we consider an ideal case in which the true subjective values are indeed heterogeneous (different from each other). The model for generating the synthesized data is as follows. The true value of the absence of reward is zero ($\kappa_0 = 0$), and the reward values of the three reward amounts (1, 2 and 3) are $\kappa_1 = 1.0, \kappa_2 = 1.5, \kappa_3 = 2.0$. The true model is the standard Q-learning model with a choice-autocorrelation factor (Model 4). The other true parameters are α_L= 0.7, $\alpha_F = 0$, $\varphi = 0.3$. We generated data from 200 hypothetical subjects and performed the parameter fit, the model selection, and the likelihood-ratio test for each subject's data.

Results

Figure 1 shows the typical model results for the first 100 trials of one hypothetical subject. Each model assigns the probability of choosing each option via the softmax function. The probabilities of choosing option 1 of four different models (the combinations of standard-Q or F-Q and the common reward value variant or different reward value variant) and the true model are depicted in Figure 1A. As expected, Model 4 (red, solid line), which has the same structure as the true model, shows the closest probability to the true model (gray line). The standard Q-learning retains the action value of the unchosen option. Thus, the action value of the bad option tends to persist even after the contingency changes (Figure 1B). In contrast, in the F-Q-learning, the value

of the bad option decays to nearly zero when the contingency changes (Figure 1C). The effect of this difference on the choice probability is evident around the 16th-19th trials (Figure 1A). In addition, differences between common value models (broken lines) and differential value models (solid lines) are also evident at approximately the 16th-19th trials (Figure 1A). This is because in the 15th, 16th, and 17th trials, the largest reward amount (=3) appears consecutively, and the action value of differential value models (including the true model) increases compared with the models with common reward values (Figure 1B, C).

The parameter estimates are shown in Table 2. Model 4, which has the same structure as the true model, yields the closest mean value to the true reward values ($\kappa_1 = 1.0, \kappa_2 = 1.5, \kappa_3 = 2.0$). Model 12 is a redundant model that includes the forgetting rate α_F as a free parameter but also includes the true model (with $\alpha_F = 0$). Thus, its parameter estimates are close to the true values. Other models (Models 3, 7, 8, 11) do not include the true model. Accordingly, the reward value estimates differ substantially from the true value; however, the relative order among the three outcomes (i.e., $\kappa_1 < \kappa_2 < \kappa_3$) is retained. We also evaluate the estimation accuracy of the reward value by computing the RMSE (Table 2). As expected, Model 4, which has the true model structure, gives the lowest RMSE. Although Model 12 contains the true model, it gives a slightly larger RMSE than Model 4 due to the estimation error of the redundant parameter (α_F).

For each hypothetical subject, the best model is selected based on AIC and BIC (Table 2). AIC selected Model 4 a total of 185 times (of 200 simulations), Model 11 once, and Model 12 a total of 14 times. All of these models have different reward values. For every hypothetical subject, a log likelihood test shows that the best model has a significantly better fit than its variant with the common reward value ($p < 0.05$). In contrast, BIC selects Model 4 a total of 172 times,

Model 2 a total of 24 times, and Model 11 four times. Model 2 has the common reward value. Thus, BIC cannot detect the true difference in reward value for 12% of the simulations, i.e., BIC has low detection power for the differences in reward values due to the over-penalization of free parameters.

Case 2 – Type I error

We next consider the case in which the true reward values are identical among decision-outcomes. In particular, we examine how often type I error occurs, that is, how often the method erroneously determines that the reward values are different. In Case 2, the true model parameter for the no-reward outcome is zero, and the true reward values are identical irrespective of the reward amount: $\kappa_0 = 0,\ \kappa_1 = \kappa_2 = \kappa_3 = 1$. The other true model parameters are $\alpha_L = 0.7$, $\alpha_F = 0,\ \varphi = 0.3$. Thus, the true model structure is Model 2 (Standard Q-learning with a choice-autocorrelation factor). We generate the synthesized data for 2000 subjects, and each model is fit to each subject's data. Then, we examine the distribution of the log likelihood difference statistic *D* (see the Proposed Method section).

Results

First, we consider the hypothesis testing in which we assume that the null model is the true model (Model 2) and that the alternative model is the one whose reward values are permitted to differ (Model 4). For this case, the distribution of statistic *D* agrees with the theoretically predicted chi-square distribution (Figure 2A), thus validating the use of the likelihood-ratio test. Even when the true model is not included among the models (null model: Model 5 vs. the

alternative model: Model 7; both are F-Q-learning) and when both models are redundant (null model: Model 9 vs. the alternative model: Model 11; both are DF-Q-learning), the distribution of *D* agrees with the chi-square distribution (Figure 2B, C). Accordingly, if the significance level is set to 0.05, the likelihood-ratio test rejects the null model 0.053 (Model 2 vs. Model 4), 0.051 (Model 5 vs. Model 7), and 0.034 (Model 9 vs. Model 11) times; all are close to the type I error (false positive) rate as intended ($\alpha = 0.05$).

AIC selected Model 4 (with different reward values) 273 times in 2000 simulations, although the likelihood-ratio test suppressed the type I error to approximately 0.05. This result suggests that demonstrating that AIC selects the model with different reward values is not sufficient to assert that different outcomes have different reward values; likelihood-ratio tests should also be performed. In contrast, BIC selected Model 1 a total of 4 times, Model 2 a total of 1960 times, Model 5 a total of 5 times, and Model 9 a total of 29 times; all selected models had a common reward value. For this case, BIC leads to the correct conclusion more frequently than AIC, albeit at the cost of detection power (as shown in Case 1).

Case 3 – Effect of learning rate

The next case (Case 3) considers the effect of the true model parameters on the accuracy of the reward value estimates. Our random feedback paradigm utilizes the property that the impact of a single outcome decays rapidly; thus, a lower learning rate is expected to degrade the estimation accuracy by slowing the decay of the impact and thus conflating the influence of several past outcomes (Katahira, 2015). Because of this conflation, the impacts of outcomes are averaged out.

The models for generating the synthesized data are as follows. The true value of the absence of reward is zero ($\kappa_0 = 0$), and the reward values of three reward amounts (1, 2, and 3) are $\kappa_1 = 1.0, \kappa_2 = 1.5, \kappa_3 = 2.0$. For Case 3-1, the true model is the standard Q-learning model (Model 4), and for Case 3-2, the true model is the F-Q-learning model (Model 8). The true parameter of choice autocorrelation is $\varphi = 0.3$. For Case 3-1, the forgetting rate is fixed at $\alpha_F = 0$, and the learning rate, α_L, varies from 0.01 to 0.99. For Case 3-2, forgetting is equated to the learning rate, $\alpha_L = \alpha_F$, and varies from 0.01 to 0.99. We performed 200 simulations for each condition. Both Model 4 and Model 8 are fit to both data sets.

Results

The results of Case 3 are shown in Table 3. As we expected, when the learning rate is very small (0.01), the RMSE is much larger. In addition, when the fitted model includes the true model, the RMSE is minimized. The error rate of model selection based on AIC and BIC also increases as the learning rate decreases. For this case, AIC and BIC provide identical results because the number of free parameters is identical between the two models [Footnote 1: Although AIC and BIC give different penalizations, the two models are subject to the same penalization if they have the same number of parameters and are compared by either AIC or BIC. Thus, only the log likelihood matters in the model comparison here.].

Figure 3 shows examples of reward value estimates. We can confirm that the variances of the estimates decrease as the true learning rate increases. When the model with the same structure as the true model (here, F-Q-learning, Model 8; Figure 3A, B, C) is used, the means of the estimates are close to the true values (1.0, 1.5 and 2.0). In contrast, when the fitted model is different from the true model, a bias arises. When the true model is F-Q-learning but the standard

Q-learning model is fit to the data, the estimates for the reward value are overestimated; however, the order among outcomes is basically retained (Figure 3D, E, F).

Case 4 – Effect of reference outcome

Thus far, we have set the value of the no-reward outcome to zero in both the true model and fitted models. Here, we examine what happens if this reference outcome is absent, i.e., when the reward value of the no-reward outcome is also estimated as a free parameter. The reward values of the true model are shifted by -0.5 from their values in Case 1; that is, $\kappa_0 = -0.5$ for the value of the absence of reward, and $\kappa_1 = 0.5$, $\kappa_2 = 1.0$, and $\kappa_3 = 1.5$ for the reward values of three reward amounts (1, 2, and 3, respectively). For Case 4-1, the true model is the standard Q-learning model (a variant of Model 4, referred to as Model 4'). The other true parameters are $\alpha_L = 0.6$, $\alpha_F = 0$, and $\varphi = 0.3$. For Case 4-2, the true model is the F-Q-learning model (a variant of Model 8, referred to as Model 8'). The other true parameters are $\alpha_L = 0.6$, $\alpha_F = 0.6$, and $\varphi = 0.3$. We performed 100 simulations for each condition.

Results

The results are shown in Table 4. Even when the true value of no-reward is non-zero (κ_0 = -0.5), both AIC and BIC often select the model with κ_0 = 0. The reason is that the absolute values (rather than the relative values) of the reward values are meaningful only in the difference from the initial Q-values, which we set to zero (Katahira, 2015; Katahira et al., 2015). Thus, when the impact of the initial value decays rapidly, such as when the learning rate is sufficiently high, the absolute values do not have a significant impact on the prediction (and thus the likelihood) of

the model. Rather, what affects the model prediction is the relative reward values among outcomes. When the models with free κ_0 are fit to the data, the variance of the estimates is larger (Figure 4B, D, F, H) than that of the models with a fixed κ_0 (Figure 4A, C, E, G). Although the relative reward values among the three rewards are relatively stable and the order is recovered, the absolute level of the reward values is unstable. This result suggests that the absolute value of the reference outcome may be obtained, but it may be unreliable. In contrast, the relative values of the outcomes are more reliable. The root mean square error of the relative values is defined as

$$\mathrm{RMSE}_{(\mathrm{relative})} = \sqrt{\frac{1}{3N}\sum_i \sum_{k=1}^{3}[\,(\hat{\kappa}_{i,k} - \hat{\kappa}_{i,0}) - (\kappa^*_{i,k} - \kappa^*_{i,0})]^2} \quad . \qquad (9)$$

When the fitted model is the standard Q-learning model, the relative RMSE is smaller than the absolute RMSE even if the true model is also the standard Q-learning model (Table 4).

Case 5 – Parameterization of reward values

Here, we examine the effect of reward value parameterizations. As the number of reward types increases, the number of parameters increases for non-parametric value models but not parametric value models. Thus, value parameterization is more effective for suppressing the estimation error for larger reward-type cases. To confirm this, we manipulate the number of reward types (as 3, 5, and 7) and confirm the absolute RMSE. For the data generation model (true model) and fitted model, the prospect utility function given in Eq. (3) is used. The true model is the standard Q-learning model with a choice-autocorrelation factor. The reward values of the true model were determined by the prospect utility function described in Eq. (3), with $\alpha_U = 0.4$. The parameter values other than those for reward values are $\alpha_L = 0.4$, $\alpha_F = 0$, and $\varphi = 0.3$. We perform the parameter fitting using both the non-parametric value model and the parametric value

model with the prospect utility function. For the later models, the inverse temperature parameter β was included as a free parameter. The other parts of the two models have the same structure as the true model. We performed 100 simulations for each condition.

Results

Figure 5 shows the reward value estimates for both the non-parametric value model (left panels) and the parametric value model (right panels). For parametric value models, the reward values are obtained by multiplying the prospect utility function by the inverse temperature parameter β. When the number of reward amount types is three (top panels, A and B), the two models give estimates with similar reliability. Overall, as the number of reward types increases (middle panels and bottom panels), the estimates become unstable even though both models recover the true reward value on average (red dotted lines). The increase in the estimation error is larger for the non-parametric value model than for the parametric value model, as confirmed by the RMSE (Table 5). The number of parameters is smaller for parametric models, and the difference become marked as the number of reward types increases. Correspondingly, AIC and BIC tend to select parametric models, especially when the number of reward-type cases is large.

Summary of the simulations

The simulation results suggest the following points. First, the estimation of the reward values in this framework relies on an assumption about the model structure. Generally, the correct model provides a better estimation. It is desirable to consider several candidate model structures and to perform model selection based on AIC or BIC. As we showed, BIC tends to favor an overly

parsimonious model in terms of the reward values. It is known that BIC generally tends to over-penalize more complex models (Kuha, 2004), consistent with our results.

Second, an extremely low learning rate leads to poor estimates of reward values. The learning rate in humans changes depending on the volatility of the reward schedule (Behrens et al., 2007). To increase the learning rate of a subject, introducing volatility into the task design may be effective. A simple way to introduce volatility is to switch the contingency, as in the proposed method. In addition, whether the estimated learning rate has an extremely small value should be confirmed.

Third, the estimates of the reward values are meaningful only in a relative sense (between different outcomes). The absolute value may be estimated, but it tends to be unstable and meaningful only compared with the initial value. Thus, if the assumption about the initial value (which is usually assumed to be zero) is incorrect, the absolute value is less meaningful. Accordingly, one should not ask about the absolute value of the outcome, e.g., “Is the subjective value of the absence of reward zero or negative?” In many cases, it is better to set the value of the reference outcome as a fixed value because it stabilizes the parameter estimates.

Fourth, when the number of reward types increases, the reliability of estimates becomes worse. The use of parametric value functions can alleviate the deterioration of the estimates. However, if the parametric function cannot represent the underlying true relation between the outcome and subjective value, a misleading conclusion is obtained. Thus, it is recommended that both non-parametric value models and parametric value models be used and that one should confirm that the results of non-parametric estimates do not differ substantially from those of the parametric estimates.

Application to rat experiments

We next demonstrate how the proposed method works for an actual animal experiment using rats. The goal here is not to extract general conclusions about rat behavior. Rather, we intend to confirm that the proposed method can extract valid and interpretable estimates from individual rats. Two rats both performed two tasks. In Task 1, we randomly manipulated the reward amount. It is reasonable to expect that the reward value should be a non-decreasing function of the reward amount because it is unlikely that a small reward is preferred to a large reward. The non-parametric value models do not include such a non-decreasing assumption. We checked whether such a non-decreasing property is obtained as a result of the parameter fit. Task 2 considered the effect of delays in feedback on outcomes. Task 2 presents a more complicated situation than Task 1; a shorter delay would be preferred (by better reinforcing the choice) for the reward signal, whereas for the non-reward signal, the impact of a delay would not be straightforward. For the data from Task 2, we employed parametric value models as well as non-parametric value models.

Materials and methods

Subjects

Two adult male Long–Evans rats (labeled as Rat 1 and Rat 2) were used in both tasks. They were housed individually under a normal light/dark cycle (lights on at 8:00 A.M. and off at 8:00 P.M.). The experiments were performed during the light phase. Water was supplied ad libitum. Food was provided after the task so that body weights were maintained at no less than 85% of the initial level. All experiments were performed in accordance with the guidelines of the Animal Experiment Committee at the University of Tokyo.

Apparatus

All experiments were conducted in an experimental chamber (ENV-009L, Med Associates, Vermont, USA) placed in a soundproof box (Muromachi Kikai, Tokyo, Japan). The chamber had two levers (ENV-112CM), a pellet dispenser (ENV-203-45), and a speaker (ENV-224AM) on one side and 9 poke holes (2.5 cm × 2.5 cm × 2.2 cm; ENV-115-9NP) on the other side. An audio generator (ANL-926) was used to generate pure tones that were delivered via the speaker.

Procedure

Before performing the main tasks, the rats were trained to learn the association between feedback tones and outcomes (existence or absence of rewards). For this procedure, the rats performed a simplified version of Task 1, in which the reward amount was fixed to one pellet. This pre-training was conducted until the rats had learned to press the advantageous lever. Both rats performed both Task 1 and Task 2, which are detailed below. Each task was conducted on consecutive days (excluding weekends) until a sufficient number of trials was obtained. The order of the tasks differed between the two rats.

Task 1 – random amount task

Task design

Each trial started with a tone presentation (800 Hz; 500 ms) and the lighting of the center poke hole. When the rat performed a nose poke in the center hole, two levers were inserted into

the chamber. After the rat pressed one of the levers, a feedback tone that informed the rat whether the reward would appear was presented. A 500-ms feedback tone at 2000 Hz indicated that a reward was present, and a 500-ms tone at 1200 Hz indicated that a reward was absent. The onset of the feedback tone was 200 ms after the detection of the lever press. Whether the trial was rewarded or non-rewarded was determined by the lever pressed by the rat. One lever was the advantageous lever, which was associated with a reward at a probability of 80% and with no reward at a probability of 20%. The second lever was the disadvantageous lever, which had a no-reward probability of 80% and a reward probability of 20%. The advantageous and disadvantageous levers switched when the choice frequency of the advantageous lever over the previous 10 choice trials reached 80%. The initial advantageous lever was randomly selected. In the rewarded trial, sucrose pellet(s) were delivered 0.5 s from the onset of the reward tone. The number of pellets was randomly drawn from 1, 2, or 3 pellets. A new trial started 3 s after the final pellet was delivered in the reward trials and 3 s after the onset of the no-reward tone in the no-reward trials.

A session was terminated when the rats obtained 300 pellets or when 60 minutes passed from the first trial. The rats performed one session per day. Rat 1 underwent Task 1 for 10 sessions, resulting in a total of 2,515 trials. Rat 2 experienced this task for 12 sessions, for a total of 3,012 trials.

Models

To evaluate whether the reward value is a non-decreasing function of the reward amount, we primarily focus only on the non-parametric reward value modes, although it is natural to use parametric models because the reward types are quantifiable. This analysis is also intended to be

a model case for general cases with non-quantifiable reward types. Thereafter, we present the results of the parametric value models. For the non-parametric value models, we used the same model set with simulations as shown in Table 1. For the parametric value models, we used the prospect utility function described in Eq. (3), with the combination of standard Q-learning / F-Q-learning / DF-Q-learning and the absence / presence of the auto-correlation factor, resulting in a total of 6 models. For these models, the inverse temperature parameter β was included as a free parameter to scale the magnitude of reward value.

Results

We first report the results of the non-parametric value models. Within these models, for Rat 1, AIC selected Model 12 (DF-Q-learning with differential reward values and choice-autocorrelation; AIC = 3045), and BIC selected Model 6 (F-Q-learning with common reward values and choice-autocorrelation; BIC = 3077). For Rat 2, both AIC and BIC selected Model 8 (F-Q-learning with differential reward values and choice-autocorrelation; AIC = 2888, BIC = 2918). With the exception of the BIC results for Rat 1, the best model had different values for different reward amounts. The likelihood-ratio test showed that the differences in the reward value were significant (for Rat 1; Model 10 (null model) vs. Model 12, $\chi^2(2) = 11.50$, $p = 0.00318$; for Rat 2; Model 10 (null model) vs. Model 15, $\chi^2(2) = 100.86$, $p < 10^{-10}$). Figure 6 plots the reward value estimates of the best non-parametric value model for each subject (gray bars). We obtained reasonable results: the estimated reward value increased (for Rat 1, from one pellet to three pellets; for Rat 2, from one pellet to two pellets) or remained constant (for Rat 2, 2 pellets and 3 pellets). To confirm the robustness of the estimates, we divided the data into odd-number sessions and even-number sessions. We then fit the best non-parametric value model selected

using the data from all sessions to divide the data set independently [Footnote 2: A standard method to confirm the robustness of model fit is cross-validation. However, cross-validation is not practical for our purpose: to estimate a reward value parameter rather than to explain or to predict the behavioral data themselves.]. The estimates are shown in Figure 6 with symbols (triangles for the odd-number sessions and crosses for the even-number sessions). The estimates for the divided data set showed similar tendencies, suggesting the robustness of the parameter estimates.

Next, we report the results of the parametric value models. For Rat 1, the parametric value models yielded smaller values of AIC and BIC (indicating better fit) than the non-parametric value models did. AIC selected DF-Q-learning with choice-autocorrelation (AIC = 3043; $\alpha_U = 0.30.\ \alpha_L = 0.80,\ \alpha_F = 99,\ \varphi = -0.44$, and $\beta = 1.64$), and BIC selected F-Q-learning with choice-autocorrelation (BIC = 3070; $\alpha_U = 0.30.\ \alpha_L = \alpha_F = 0.89,\ \varphi = -0.42$, and $\beta = 1.47$). This is because the parametric reward value function (scaled by the inverse temperature parameter, β) agreed with the non-parametric reward value (Figure 6A, broken line) and the parametric value model has one fewer parameter than the non-parametric value model. In contrast, for Rat 2, the parametric value models did not provide a better fit than non-parametric value models (the minimum value of AIC was 2899, and that of BIC is 2923; both were given by F-Q-learning with choice-autocorrelation; $\alpha_U = 0.40.\ \alpha_L = \alpha_F = 0.99,\ \varphi = -0.74$, and $\beta = 2.32$). This is because the parametric value function cannot capture the saturated pattern of reward value in Rat 2 (Figure 6B, broken line). This result suggests that when the parametric function does not match the true reward values, non-parametric models would be the better models.

Task 2 – random-delay task

Task 2 manipulated the feedback delays regarding outcomes while fixing the reward amount in the reward trials. In this task, we examined the utility of the parametric models. As there are five outcome types for both rewarded trials and non-rewarded trials, parameterization may facilitate the estimation of the underlying patterns.

Task design

The basic task design of Task 2 was similar to Task 1, with some exceptions. First, the latency of the feedback tones (reward tone: 2000 Hz, 500 ms; no-reward tone: 1200 Hz, 500 ms) was drawn randomly from 0 sec, 1 sec, 2 sec, 4 sec, or 6 sec for both feedback types. A single pellet was always provided for the reward trials. The reward probability was 85% for the advantageous lever and 15% for the disadvantageous lever. These probabilities made the discrimination slightly easier on average than in Task 1. The goal of this modification was to alleviate the difficulty due to long feedback delays. Independent of the feedback delay, the next trial started 10 sec after the lever was pressed; for example, if the feedback delay was 6 sec, the next trial began 4 sec after the feedback tone.

A session was terminated when the rats obtained 300 pellets or when 90 minutes passed from the first lever push. Each rat performed one session per day. Rat 1 experienced Task 2 for 10 sessions, resulting in 4,069 trials, whereas Rat 2 experienced this task for 15 sessions, resulting in 5,455 trials.

Models

The fitted models with non-parametric values consist of all combinations of (1) standard-Q learning ($\alpha_F = 0$) / F-Q-learning ($\alpha_L = \alpha_F$) / DF-Q-learning ($\alpha_L \neq \alpha_F$); (2) reward values of reward signals: homogeneous / heterogeneous, (3) reward values of no-reward signals: homogeneous / heterogeneous (no-fixed values) / heterogeneous with the value of the no-delay signal (κ_{N0}) being fixed at zero, resulting in a total of 24 models. We included models in which the choice-autocorrelation factor was a free parameter only because we confirmed that the no-autocorrelation factor model was clearly a worse fit to the data (data not shown), as in Task 1.

As the models with reward-value parameterization, we included the two parameterization methods for reward feedback. One variation assumed the polynomial function for the reward trials:

$$r\,(t) = w_{R,0} + w_{R,1} \cdot D(t) + w_{R,2} \cdot D(t)^2, \tag{11}$$

where $D\,(t)$ denotes the delay of reward signal at trial *t*. The w coefficients are free parameters. We also examined the first-order function by setting $w_{R,2} = 0$ and the zero-th order function (only in the constant term version) by setting $w_{R,1} = 0$ and $w_{R,2} = 0$. The polynomial function for the no-reward trials are similarly defined as $r\,(t) = w_{NR,0} + w_{NR,1} \cdot D(t) + w_{NR,2} \cdot D(t)^2$. The different parameterization function for the reward signal is the hyperbolic discounting function:

$$r\,(t) = \frac{A}{1+k \cdot D(t)}, \tag{12}$$

where A and k are the free parameters. The hyperbolic discounting function is often used to represent the effect of delay on a subjective value (e.g., Mazur & Biondi, 2011) but less frequently to represent the reward value in reinforcement learning models. For the models with a hyperbolic

discounting function, polynomial functions (first-order, second-order, or zero-th- order) are used for non-reward trials. Null models regarding the hyperbolic discount function with k = 0 (indicating homogeneous reward values) are included. For each combination of the value parameterizations, there are standard Q-learning, F-Q-learning and DF-Q-learning versions, constituting a total of 45 models.

Results

For both subjects, a parametric model was selected by AIC and BIC. For the data from Rat 1, AIC selected standard the Q-learning model with hyperbolic discounting (for the reward signal) and the second-order polynomial model (for the no-reward signal). The estimated parameter values were α_L = 0.155, $\varphi = -0.167$, $A = 2.50$, $k = 0.163$, $w_{NR,0} = 0.285, w_{NR,1} = -0.26, w_{NR,2} = 0.062$. Regarding the hyperbolic discounting function, the likelihood was significantly improved compared with the null model (with k = 0), which assigns common values for reward signals ($\chi^2(1) = 8.08$, $p = 0.0045$). Regarding the polynomial function for non-reward trials, the selected model (including the second-order term) exhibited a marginally significant difference in likelihood compared with the zero-th order model, with $w_{NR,1} = 0$ and $w_{NR,2} = 0$ ($\chi^2(2) = 5.98$, $p = 0.05$), but not when compared with the first-order model, with $w_{NR,2} = 0$ ($\chi^2(1) = 2.53$, $p = 0.112$). Thus, we deemed the first-order model the best model, with estimates of α_L= 0.115, $\varphi = -0.166$, $A = 2.55$, $k = 0.162$, $w_{NR,0} = 0.021,$ and $w_{NR,1} = 0.13$. Regarding Rat 2, AIC selected the standard Q-learning model with a first-order polynomial function (linear function) for both the reward signal and the no-reward signal, with estimated parameters of α_L = 0.091, $\varphi = -0.477$, $w_{R,0} = 3.309, w_{R,1} = -0.382,$ $w_{NR,0} = 0.394,$ and $w_{NR,1} = 0.29$. Both polynomial functions exhibited significant differences in

likelihood compared with the first-order models (vs. model with $w_{R,2} = 0$, $\chi^2(1) = 27.44$, $p < 10^{-10}$; $w_{NR,2} = 0$, $\chi^2(1) = 16.57$, $p = 0.000047$), indicating that the linear trends are significant.

Figure 7 shows the estimated reward values obtained from the selected parametric models (lines) with the results of the non-parametric value estimates (whose model structure is the same as that of the selected parametric models). For reward signals, the reward value decreased as the delay increased (Figure 7A, B). For no-reward feedback, the parametric model suggested that the value increased in a positive direction as the delay increased. The non-parametric model suggested that the values are non-monotonic, reaching a minimum when the feedback latency was 1 s and then increasing, but our analysis indicated that these non-monotonic properties were not supported by the data. The estimates for the divided data set showed that the estimated tendency was robust (triangles for the odd-number sessions and crosses for the even-number sessions), even though the linear increasing trend for the no-reward signal in Rat 1 was weak for the even-number sessions.

Discussion

In the present paper, we have proposed a novel framework for estimating subjective values. The framework consists of novel behavioral tasks and model-based analysis of the behavioral data. Our methods utilize the history dependence of choice behavior (i.e., the larger the subjective value of the outcome, the more likely the action is to be repeated; this influence decays as subjects experience additional trials). This tendency is represented by the RL framework, from which we can estimate the subjective values. The proposed method has at least two advantages over previous methods of estimating subjective values. First, it can estimate the subjective values of multiple outcomes within a single experimental design. Second, the number

of outcomes types from which the subjective value is estimated is flexible and can be changed.

Several previous studies have attempted to estimate the value of various outcomes using similar RL model-based methods (Katahira et al., 2011, 2014; Lindström et al., 2014). In contrast to our experimental design, in previous studies, the outcome stimulus type (e.g., the valence of pictures) was stochastically contingent on the subjects' choices. This approach limits the number of outcome categories. With this constraint, increasing the number of outcome types complicates the design of the experiment schedule. Indeed, these studies have estimated the subjective values of only a few stimulus outcome categories. In contrast, our novel experimental design randomly chooses the outcome (within the same valence) independent of the subject's choice; thus, researchers can easily include an arbitrary number of outcome types without special consideration of the task design, albeit with the number of outcome types being limited according to the number of trials.

As we have discussed, the proposed method relies in part on the assumption that the RL model structure is valid. If the underlying decision-making processes in the actual animal behavior differ from the fitted model, the resulting estimates can differ from the true subjective value. Thus, it is desirable to consider several candidate model structures and perform statistical model selection. For example, the subjects may learn the higher-order structure of the task, such as the timing of reward schedule reversal (Myers, 1976). For our task design, if the subjects were able to take advantage of such a structure, the estimated subjective value would be flat over different reward types. This was not the case, suggesting that our subjects could not follow the strategy and that at least a part of the RL model was valid. Nevertheless, the researcher should consider such a possibility and design the task so that such an unintended factor does not contaminate the estimates. In addition, estimated reward values are meaningful in the relative

value between different outcome types; however, one advantage of the proposed method is that relative values among outcomes are often retained even if the fitted model is incorrect. Nevertheless, this property may depend on the discrepancy between the fitted model structure and the true model structure.

We assumed that the subjective value of the decision-outcome is represented as the "reward value" in the RL framework, i.e., the outcome evaluation (OE) hypothesis (Lindström et al., 2014). Others have argued that the stimulus property of outcome in the RL framework influences the learning rate, a view that is termed the outcome learning (OL) hypothesis (Lindström et al., 2014). In general, the higher the learning rate, the more recent outcomes influence future choices. However, the total impact of past outcomes essentially does not change even if the learning rate changes because a higher learning rate also induces greater decay (Katahira, 2015). Therefore, it is not reasonable to assume that the subjective value is reflected in the magnitude of the learning rate. Indeed, we examined data from both the OL model and a hybrid of the OL and OE models from the random amount experiment (Task 1), but neither provided a better fit than the (OE) models used in the proposed framework (data not shown).

We demonstrated how the proposed method works for actual animal behavior using data from rat experiments. As the goal of the demonstration was to confirm that the proposed method could obtain valid estimates from individual rats, we used only two rats. We showed that the reward values were not identical among three reward amounts for both rats and that the value monotonically increased or remained the same as the reward amount (the number of pellets) increased, a reasonable result. We also showed that the reward values roughly decreased as the delay in the reward signal increased, in accordance with previous studies on delay discounts (Green et al., 2004; Hirsh et al., 2010; Paglieri, 2013). Regarding the delay of the no-reward signal,

the value increased as the delay increased. This result may indicate that a longer delay causes the next trial to begin sooner based on how we fixed the inter-trial interval in the present experimental design. However, these results are preliminary and far from conclusive. The proposed framework provides a promising way to understand these complex psychological processes. Although we intended to propose a framework for extracting the subjective value of animal subjects, future work is needed to examine how the proposed framework also works for human subjects. For human subjects, econometric methods are available, but using the same task with animal studies enables a direct comparison between animal models and humans.

In the present paper, we have focused most of our attention on the estimation of subjective value. Perhaps more importantly, however, the advantage of the model-based approach is that it can estimate and represent animals' internal computational processes. Constructing a better model is beneficial for estimating internal variables such as the action value, reward prediction error, and attention. By incorporating the estimation of subjective values, we can derive better estimates of the variables involved in the subjective experiences of animals. In addition, the computational approaches carry implications for psychiatry (Maia and Frank, 2011; Montague et al., 2012; Huys et al., 2013). The method presented here will contribute to this promising field.

References

Ahn, W.-K., Busemeyer, J. R., Wagenmakers, E.-J., & Stout, J. C. (2008). Comparison of decision learning models using the generalization criterion method. *Cognitive Science*, 32, 1376–1402.

Akaike H. (1974). A new look at the statistical model identification. *IEEE Trans. Autom. Control. 19*, 716–723.

Behrens T.E.J, Woolrich M.W., Walton M.E., Rushworth M.F.S. (2007). Learning the value of

information in an uncertain world. *Nat. Neurosci. 10*, 1214–1221.

Chen, C., Takahashi, T., Nakagawa, S., Inoue, T., & Kusumi, I. (2015). Reinforcement learning in depression: A review of computational research. *Neuroscience and Biobehavioral Reviews, 55*, 247–267.

Corrado G., Doya K. (2007). Understanding neural coding through the model-based analysis of decision making. *J. Neurosci. 27*, 8178–8180.

Daw N.D. (2011). Trial-by-trial data analysis using computational models. *Decis. Making, Affect. Learn. Atten. Perform. XXIII 23*, 1–26.

Green L., Estle S.J. (2003). Preference reversals with food and water reinforcers in rats. *J. Exp. Anal. Behav., 79*, 233–242.

Green L., Myerson J, Holt D.D., Slevin J.R., Estle S.J. (2004). Discounting of delayed food rewards in pigeons and rats: is there a magnitude effect? *J. Exp. Anal. Behav. 81*, 39–50.

Hirsh J.B., Guindon A., Morisano D., Peterson J.B. (2010). Positive mood effects on delay discounting. *Emotion, 10*, 717–721.

Huys Q.J., Pizzagalli D.A., Bogdan R., Dayan P. (2013). Mapping anhedonia onto reinforcement learning: a behavioural meta-analysis. *Biol. Mood Anxiety Disord., 3*, 12.

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nat. Neurosci. 10*, 1625–1633.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: : Journal of the Econometric Society*, 263–291.

Katahira K., Fujimura T., Matsuda Y-T., Okanoya K., Okada M. (2014). Individual differences in heart rate variability are associated with the avoidance of negative emotional events. *Biol. Psychol. 103*, 322–331.

Katahira K., Fujimura T., Okanoya K., Okada M. (2011). Decision-Making Based on Emotional Images. *Front. Psychol., 2*, 311.

Katahira K., Matsuda Y-T., Fujimura T., Ueno K., Asamizuya T., Suzuki C., Cheng K., Okanoya K., Okada M. (2015). Neural basis of decision-making guided by emotional outcomes. *J. Neurophysiol. 113*, 3056-3068.

Katahira K. (2015). The relation between reinforcement learning parameters and the influence of reinforcement history on choice behavior. *J. Math. Psychol. 66*, 59–69.

Kuha, J. (2004). AIC and BIC comparisons of assumptions and performance. *Sociological Methods & Research, 33*, 188-229.

Lindström B., Selbing I., Molapour T., Olsson A. (2014). Racial bias shapes social reinforcement learning. *Psychol. Sci. 25*, 711–719.

Maia T.V., Frank M.J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nat. Neurosci. 14*, 154–162.

Myers, J. L. (1976). Probability learning. In W. K. Estes (Ed.), Handbook of Learning and Cognitive Processes: Vol. 3. Approaches to Human Learning and Motivation (pp. 171-205). Hillsdale, NJ: Erlbaum.

Mazur, J. E., & Biondi, D. R. (2011). Effects of time between trials on rats' and pigeons' choices with probabilistic delayed reinforcers. *J. Exp. Anal. Behav. 95*, 41–56.

Miller H.L. (1976). Matching-based hednic scaling in the pigeion. *J. Exp. Anal. Behav. 26*, 335–347.

Montague P.R., Dolan R.J., Friston K.J., Dayan P. (2012). Computational psychiatry. *Trends Cogn. Sci. 16*, 72–80.

Neiman T., Loewenstein Y. (2011). Reinforcement learning in professional basketball players. *Nat. Commun. 2*, 569.

Neiman T., Loewenstein Y. (2014). Spatial generalization in operant learning: lessons from professional basketball. *PLoS Comput. Biol. 10*, e1003623.

O'Doherty J.P., Hampton A., Kim H. (2007). Model-based fMRI and its application to reward learning and decision making. *Ann. N. Y. Acad. Sci. 1104*, 35–53.

O'Doherty J.P. (2014). The problem with value. *Neurosci. Biobehav. Rev. 43*, 259–268.

Paglieri F. (2013). The costs of delay: waiting versus postponing in intertemporal choice. *J. Exp. Anal. Behav. 99*, 362–377.

Rangel A., Camerer C., Montague P.R. (2008). A framework for studying the neurobiology of value-based decision making. *Nat. Rev. Neurosci. 9*, 545–556.

Schwarz G. (1978). Estimating the dimension of a model. *Ann. Stat. 6*, 461–464.

Sakai, Y., & Fukai, T. (2008). The actor-critic learning is behind the matching law: Matching versus optimal behaviors. *Neural Computation*, *20*, 227–251.

Shteingart H., Loewenstein Y. (2014). Reinforcement learning and human behavior. *Curr. Opin. Neurobiol. 25*, 93–98.

Acknowledgments

This work was partially supported by Grants-in-Aid for Scientific Research (KAKENHI) Nos. 24700238, 26118506, 15K12140, and 23118003.

Figures

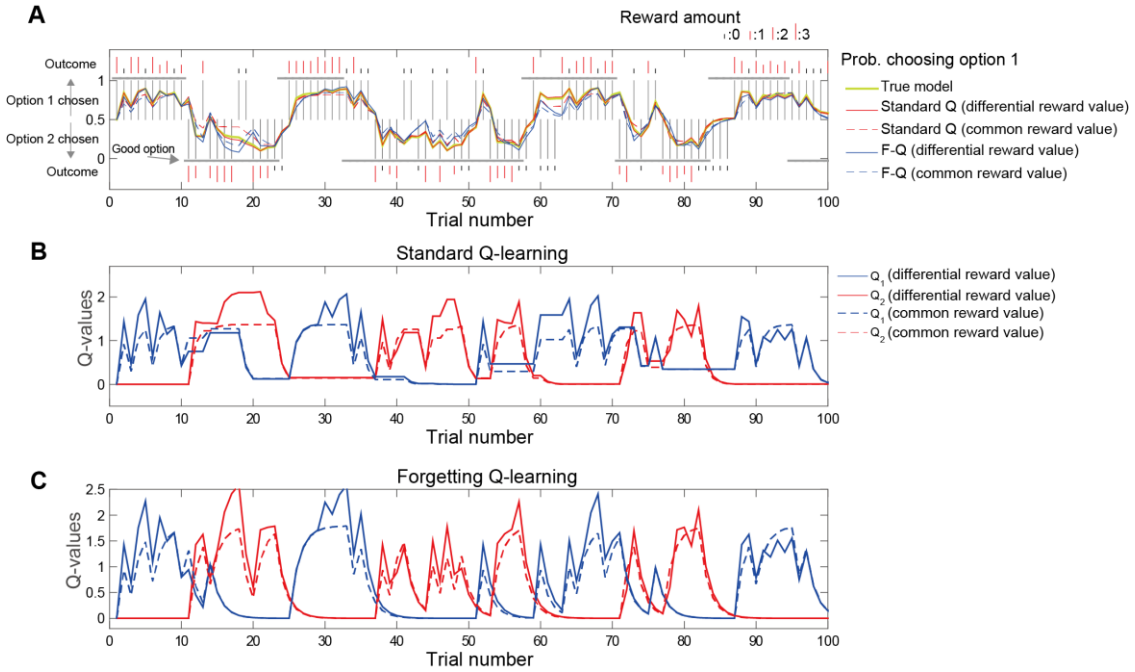

Figure 1.

An example of RL model behavior for the simulated task (Case 1). **A**: An example of choice data from a single hypothetical subject (true model; Model 4) and the fit of four RL models. The first half of the session is depicted. The vertical bars in the top panel indicate the chosen option. Their color and length represent the reward amount. The probability of choosing option 1 was calculated from the fitted Q-learning models (see the legend). Corresponding action values (Q-values) are shown for the standard Q-learning (**B,** solid line: Model 4, broken line: Model 2) and forgetting Q-learning models (**C**, solid line: Model 8, broken line: Model 6).

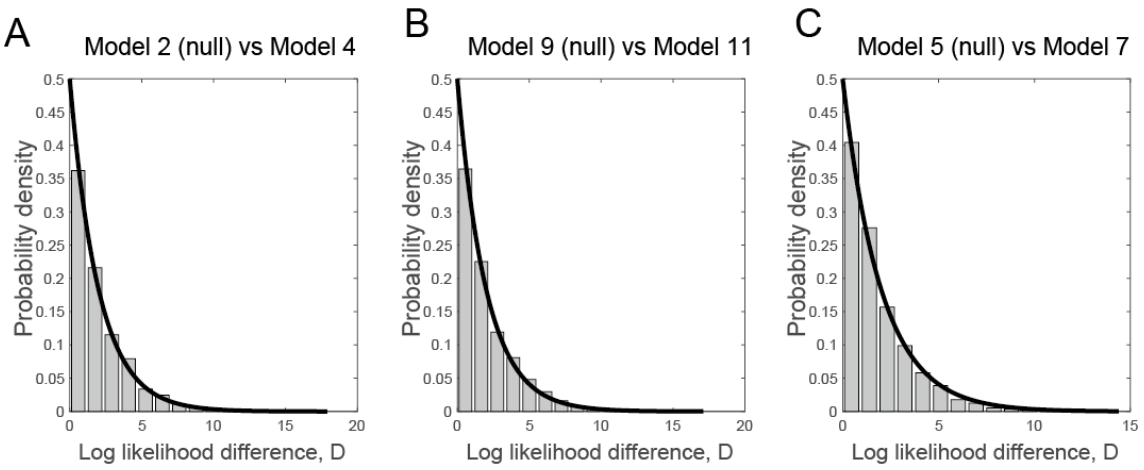

Figure 2.

The histogram of the likelihood-ratio statistic *D* when the null hypothesis is correct (true reward values were identical for different reward outcomes; Case 2). The results of various model comparisons are shown in each panel. The solid line represents the chi-square distribution with 2 degrees of freedom. The agreements between the histograms and the chi-square distributions validate the use of the likelihood-ratio test.

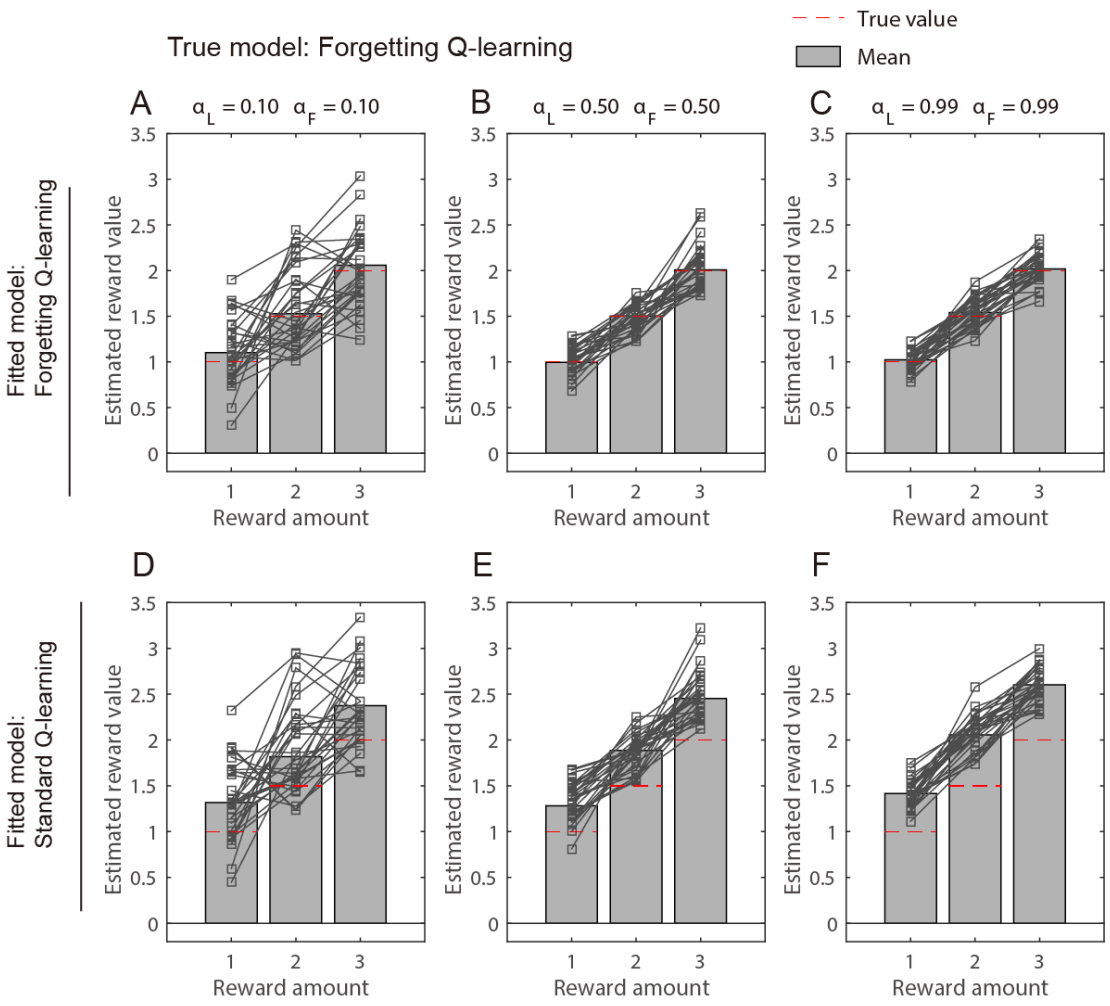

Figure 3.

Individual reward value estimates for the simulation examining the effect of the true learning rate (Case 3). The true model was F-Q-learning, and the fitted model was either the F-Q-learning (**A**-**C**) model or the standard Q-learning (**D**-**F**) model. The gray bars indicate the mean over the estimates for 100 subjects. The estimates for the same subject are connected by lines. Estimates for 30 subjects (of 100) are shown for visibility. Broken red lines represent the true reward value.

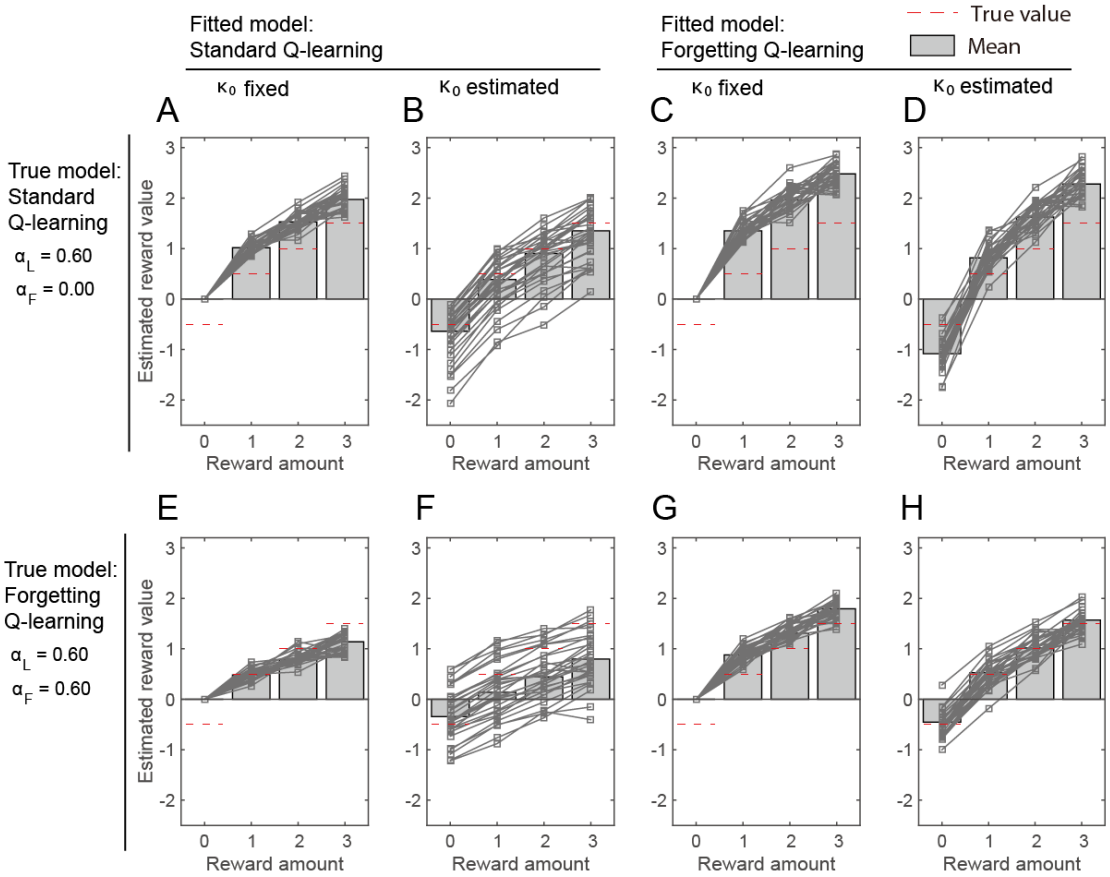

Figure 4.

Individual reward value estimates for the simulation examining the effect of fixed reference outcomes (Case 4). **A-D**: The true model for generating data is the standard Q-learning model. **E-H**: The true model for generating data is the F-Q-learning model. **A, C, E, G**: The reward value of the reference outcome (no-reward) is fixed. **B, D, E, H**: The reward value of the reference outcome (no-reward) is a free parameter. The conventions are the same as in Figure 3.

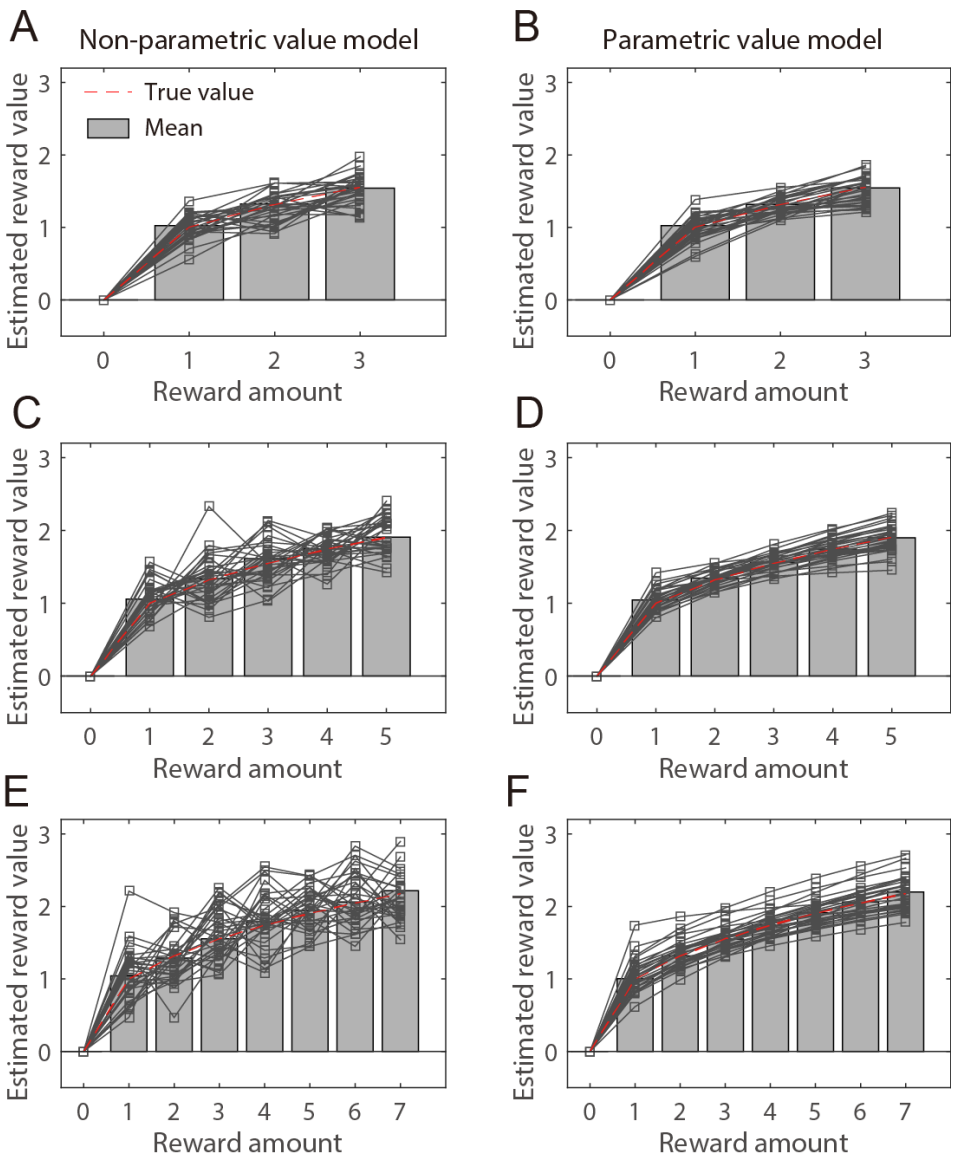

Figure 5.

Individual reward value estimates for the simulation examining the effect of reward value parameterization (Case 5). **A, C, E**: The fitted models have non-parametric values. **B, D, F**: The fitted models are parameterized with the prospect utility function (multiplied by the parameter β). The estimates for the same subject are connected by lines. The conventions are the same as in Figure 3.

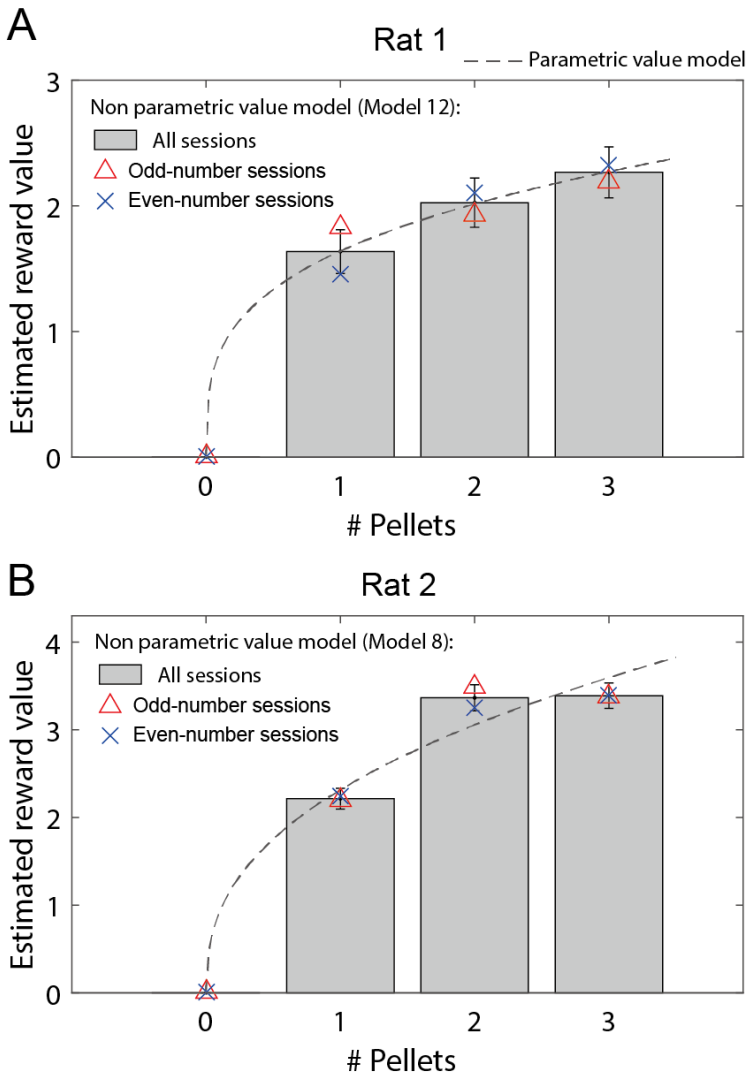

Figure 6.

The reward value estimates for the random amount task (Task 1). **A**: Rat 1. **B**: Rat 2. The best non-parametric value model selected by AIC was used for each subject. The reward value of the no-reward outcome was fixed at zero. Gray bars represent the estimates from the whole data set. Error bars represent the standard error. Symbols indicate the estimates from the divided independent data set (triangles for the odd-number sessions and crosses for the even-number sessions). The reward value functions obtained from the parametric value model selected by AIC are plotted as broken lines.

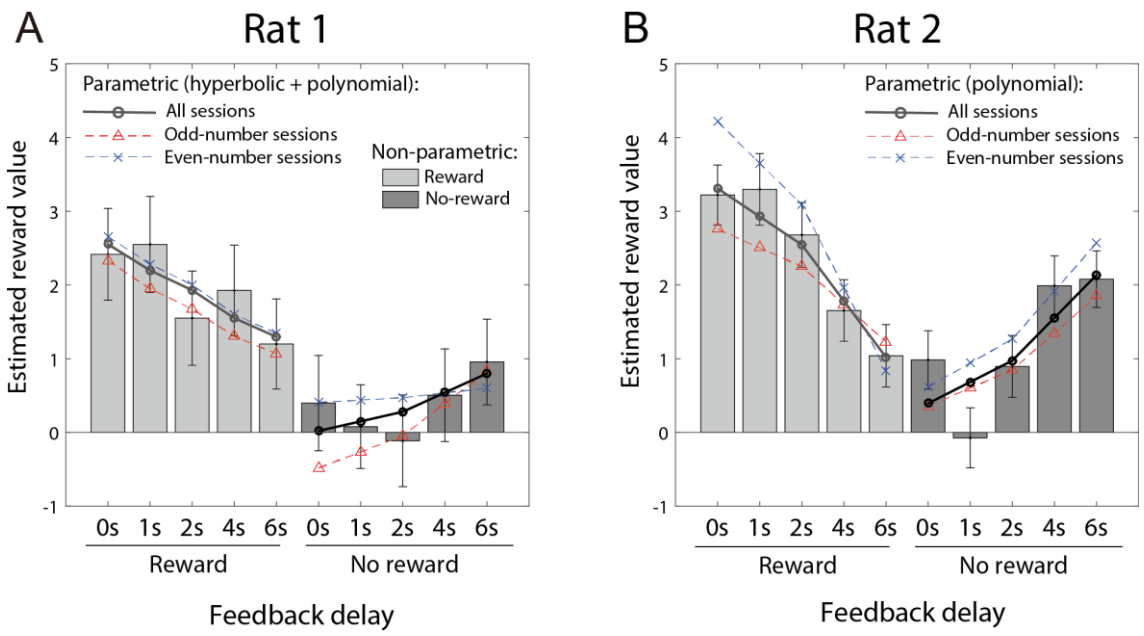

Figure 7.

The reward value estimates for the random-delay task (Task 2). The reward value estimates of the selected parametric models are shown as lines (solid lines for the whole data set and broken lines for the divided data set). For Rat 1, the standard Q-learning models with a hyperbolic function and polynomial function were selected (**A**). For Rat 2, the model with first-order polynomial functions for both the reward signal and no-reward signal was selected (**B**). The estimates from the non-parametric models are shown with bars for reference. The error bars represent the standard error of the estimates obtained from the non-parametric models.

Tables

Table 1.

The list of non-parametric value models for the simulations (Cases 1 – 4) and the rat random-amount task.

Model ID	**Learning rate α_L**	**Forgetting rate α_F**	**Reward value $\kappa_1, \kappa_2, \kappa_3$**	**Choice-autocorrelation φ**	**# Free parameters**
1	free	0 (fixed)	homogeneous	0 (fixed)	2
2	free	0 (fixed)	homogeneous	free	3
3	free	0 (fixed)	heterogeneous	0 (fixed)	4
4	free	0 (fixed)	heterogeneous	free	5
5	free	$= \alpha_L$	homogeneous	0 (fixed)	2
6	free	$= \alpha_L$	homogeneous	free	3
7	free	$= \alpha_L$	heterogeneous	0 (fixed)	4
8	free	$= \alpha_L$	heterogeneous	free	5
9	free	free	homogeneous	0 (fixed)	3
10	free	free	homogeneous	free	4
11	free	free	heterogeneous	0 (fixed)	5
12	free	free	heterogeneous	free	6

Table 2.

The results of the simulation - Case 1.

Model ID	Estimated parameter values							AIC	BIC	# Selected (BIC)	# Selected (AIC)	RMSE
	Learning rate α_L	Forgetting rate α_F	κ_0	κ_1	κ_2	κ_3	φ					
1	0.70 (0.06)	0.0 (fixed)	0.0 (fixed)	1.64 (0.11)	1.64 (0.11)	1.64 (0.11)	0.0 (fixed)	2314.3 (39.5)	2325.5 (39.5)	0	0	–
2	0.70 (0.06)	0.0 (fixed)	0.0 (fixed)	1.42 (0.10)	1.42 (0.10)	1.42 (0.10)	0.33 (0.05)	2279.5 (39.9)	2296.3 (39.9)	24	0	–
3	0.70 (0.05)	0.0 (fixed)	0.0 (fixed)	1.14 (0.13)	1.69 (0.17)	2.31 (0.21)	0.0 (fixed)	2282.8 (40.5)	2305.2 (40.5)	0	0	0.264 (0.102)
4	0.70 (0.05)	0.0 (fixed)	0.0 (fixed)	1.00 (0.12)	1.49 (0.15)	2.04 (0.19)	0.29 (0.05)	2255.7 (40.4)	2283.7 (40.4)	172	185	0.146 (0.064)
5	0.55 (0.04)	0.55 (0.04)	0.0 (fixed)	1.82 (0.10)	1.82 (0.10)	1.82 (0.10)	0.0 (fixed)	2323.3 (40.5)	2334.5 (40.5)	0	0	–
6	0.56 (0.05)	0.56 (0.05)	0.0 (fixed)	1.89 (0.12)	1.89 (0.12)	1.89 (0.12)	−0.07 (0.07)	2323.5 (40.6)	2340.3 (40.6)	0	0	–
7	0.56 (0.04)	0.56 (0.04)	0.0 (fixed)	1.23 (0.17)	1.84 (0.18)	2.45 (0.20)	0.0 (fixed)	2307.9 (41.0)	2330.3 (41.0)	0	0	0.385 (0.101)
8	0.57 (0.04)	0.57 (0.04)	0.0 (fixed)	1.33 (0.18)	1.93 (0.19)	2.55 (0.22)	−0.08 (0.07)	2307.6 (41.1)	2335.6 (41.1)	0	0	0.472 (0.117)
9	0.68 (0.05)	0.19 (0.05)	0.0 (fixed)	1.74 (0.10)	1.74 (0.10)	1.74 (0.10)	0.0 (fixed)	2294.0 (40.1)	2310.8 (40.1)	0	0	–
10	0.70 (0.06)	0.03 (0.04)	0.0 (fixed)	1.46 (0.11)	1.46 (0.11)	1.46 (0.11)	0.29 (0.07)	2280.4 (40.0)	2302.8 (40.0)	0	0	–
11	0.69 (0.05)	0.14 (0.05)	0.0 (fixed)	1.22 (0.14)	1.77 (0.16)	2.35 (0.20)	0.0 (fixed)	2270.9 (40.8)	2298.9 (40.8)	4	1	0.314 (0.102)
12	0.70 (0.05)	0.02 (0.03)	0.0 (fixed)	1.01 (0.12)	1.51 (0.16)	2.06 (0.20)	0.28 (0.06)	2257.2 (40.5)	2290.8 (40.5)	0	14	0.151 (0.068)

The average parameter estimates across subjects are shown with the S.D. in parentheses.

Table 3.

The results of the simulation - Case 3.

		Learning rate (forgetting rate)								
		0.01	0.1	0.2	0.4	0.5	0.6	0.7	0.8	0.99
(i) True model: Standard-Q (Model 4)										
Model 4	RMSE	5.24 (10.42)	0.36 (0.16)	0.27 (0.13)	0.19 (0.08)	0.15 (0.07)	0.15 (0.07)	0.14 (0.07)	0.14 (0.06)	0.13 (0.06)
	Frac. value order correct	0.31	0.68	0.74	0.94	0.98	1.00	0.98	1.00	1.00
Model 8	RMSE	7.24 (12.34)	0.50 (0.23)	0.40 (0.15)	0.40 (0.13)	0.41 (0.12)	0.47 (0.12)	0.51 (0.11)	0.53 (0.12)	0.55 (0.12)
	Frac. value order correct	0.23	0.62	0.73	0.89	0.99	1.00	0.99	0.97	0.99
Frac. model selection correct		0.83	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00
(ii) True model: F-Q-learning (Model 8)										
Model 4	RMSE	7.13 (11.76)	0.63 (0.18)	0.59 (0.16)	0.63 (0.12)	0.63 (0.11)	0.65 (0.10)	0.65 (0.09)	0.69 (0.10)	0.66 (0.08)
	Order correct	0.15	0.46	0.64	0.89	0.81	0.91	0.91	0.9	0.96
Model 8	RMSE	6.80 (11.14)	0.44 (0.20)	0.33 (0.14)	0.21 (0.08)	0.20 (0.08)	0.19 (0.08)	0.18 (0.08)	0.17 (0.06)	0.14 (0.06)
	Frac. value order correct	0.15	0.53	0.74	0.92	0.84	0.94	0.96	0.99	0.99
Frac. model selection correct		0.33	0.9	0.99	0.99	1.00	1.00	1.00	1.00	1.00

The average RMSEs across subjects are shown in addition to the S.D. in parentheses.

Table 4.

The results of the simulation – Case 4.

	Fitted model			
	Standard Q-learning		Forgetting Q-learning	
True model	κ_0 fixed	κ_0 free	κ_0 fixed	κ_0 free
Selected by AIC				
Standard Q-learning (Model 4')	69	31	0	0
Forgetting Q-learning (Model 8')	0	0	34	66
Selected by BIC				
Standard Q-learning (Model 4')	96	4	0	0
Forgetting Q-learning (Model 8')	1	0	80	19
RMSE (absolute)				
Standard Q-learning (Model 4')	0.52 (0.08)	0.43 (0.35)	0.86 (0.11)	0.64 (0.12)
Forgetting Q-learning (Model 8')	0.34 (0.05)	0.63 (0.36)	0.40 (0.07)	0.21 (0.11)
RMSE (relative)				
Standard Q-learning (Model 4')	0.15 (0.06)	0.16 (0.07)	0.46 (0.13)	1.19 (0.25)
Forgetting Q-learning (Model 8')	0.72 (0.09)	0.72 (0.09)	0.22 (0.09)	0.19 (0.08)

The average RMSEs across subjects are shown in addition to the S.D. in parentheses.

Table 5.

The results of the simulation – Case 5.

		Fitted model		
	Max. reward size	Homogeneous	Non-parametric	Parametric
RMSE				
	3	0.23 (0.03)	0.16 (0.07)	0.15 (0.11)
	5	0.31 (0.03)	0.23 (0.08)	0.13 (0.07)
	7	0.39 (0.04)	0.29 (0.08)	0.15 (0.07)
Selected by AIC				
	3	25	13	62
	5	9	18	73
	7	0	7	93
Selected by BIC				
	3	75	0	25
	5	46	0	54
	7	18	0	82

Table 6.

The results of the rat random-amount experiment (Task 1, non-parametric value models). The minimum values of AIC and BIC are shown in bold.

	Learning rate α_L	Forgetting rate α_F	κ_0	κ_1	κ_2	κ_3	φ	AIC	BIC
Rat 1									
Model 1	0.999	0.0 (fixed)	0.0 (fixed)	0.930	0.930	0.930	0.0 (fixed)	3179	3190
Model 2	0.999	0.0 (fixed)	0.0 (fixed)	0.905	0.905	0.905	0.059	3179	3196
Model 3	0.996	0.0 (fixed)	0.0 (fixed)	0.700	0.929	1.241	0.0 (fixed)	3159	3183
Model 4	0.996	0.0 (fixed)	0.0 (fixed)	0.687	0.912	1.218	0.039	3161	3190
Model 5	0.887	0.887	0.0 (fixed)	1.112	1.112	1.112	0.0 (fixed)	3151	3163
Model 6	0.898	0.898	0.0 (fixed)	1.763	1.763	1.763	-0.616	3056	3074
Model 7	0.882	0.882	0.0 (fixed)	0.819	1.159	1.395	0.0 (fixed)	3143	3166
Model 8	0.893	0.893	0.0 (fixed)	1.472	1.814	2.050	-0.617	3048	**3077**
Model 9	0.999	0.434	0.0 (fixed)	1.103	1.103	1.103	0.0 (fixed)	3127	3145
Model 10	0.8	0.999	0.0 (fixed)	1.972	1.972	1.972	-0.716	3052	3076
Model 11	0.999	0.406	0.0 (fixed)	0.856	1.112	1.370	0.0 (fixed)	3117	3146
Model 12	0.803	0.999	0.0 (fixed)	1.637	2.025	2.267	-0.716	**3045**	3080
Rat 2									
Model 1	0.999	0.0 (fixed)	0.0 (fixed)	1.293	1.293	1.293	0.0 (fixed)	3480	3492
Model 2	0.999	0.0 (fixed)	0.0 (fixed)	1.725	1.725	1.725	-0.63	3325	3343
Model 3	0.999	0.0 (fixed)	0.0 (fixed)	0.820	1.684	1.645	0.0 (fixed)	3402	3426
Model 4	0.999	0.0 (fixed)	0.0 (fixed)	1.180	2.304	2.268	-0.739	3199	3229
Model 5	0.999	0.999	0.0 (fixed)	1.205	1.205	1.205	0.0 (fixed)	3620	3632
Model 6	0.999	0.999	0.0 (fixed)	2.910	2.910	2.910	-1.704	2985	3003
Model 7	0.999	0.999	0.0 (fixed)	0.509	1.66	1.684	0.0 (fixed)	3523	3547
Model 8	0.999	0.999	0.0 (fixed)	2.214	3.365	3.39	-1.704	**2888**	**2918**
Model 9	0.999	0.16	0.0 (fixed)	1.367	1.367	1.367	0.0 (fixed)	3464	3482
Model 10	0.999	0.956	0.0 (fixed)	2.889	2.889	2.889	-1.674	2986	3010
Model 11	0.999	0.155	0.0 (fixed)	0.848	1.764	1.722	0.0 (fixed)	3389	3419
Model 12	0.999	0.951	0.0 (fixed)	2.193	3.346	3.364	-1.674	2890	2926