The statistical structures of reinforcement learning with asymmetric value updates

Kentaro Katahira

Department of Cognitive and Psychological Sciences, Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8601, Japan

ABSTRACT

Reinforcement learning (RL) models have been broadly used in modeling the choice behavior of humans and other animals. In standard RL models, the action values are assumed to be updated according to the reward prediction error (RPE), i.e., the difference between the obtained reward and the expected reward. Numerous studies have noted that the magnitude of the update is biased depending on the sign of the RPE. The bias is represented in RL models by differential learning rates for positive and negative RPEs. However, which aspect of behavioral data that the estimated differential learning rates reflect is not well understood. In this study, we investigate how the differential learning rates influence the statistical properties of choice behavior (i.e., the relation between past experiences and the current choice) based on theoretical considerations and numerical simulations. We clarify that when the learning rates differ, the impact of a past outcome depends on the subsequent outcomes, in contrast to standard RL models with symmetric value updates. Based on the results, we propose a model-neutral statistical test to validate the hypothesis that value updates are asymmetric. The asymmetry in the value updates induces the autocorrelation of choice (i.e., the tendency to repeat the same choice or to switch the choice irrespective of past rewards). Conversely, if an RL model without an intrinsic autocorrelation factor is fitted to data that possess an intrinsic autocorrelation, a statistical bias to overestimate the difference in learning rates arises. We demonstrate that this bias can cause a statistical artifact in RL-model fitting leading to a “pseudo-positivity bias” and a “pseudo-confirmation bias.”

© 2018 The Author. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Reinforcement learning (RL) models have been broadly used in modeling action selection in various fields such as psychology, neuroscience, and psychiatry (Corrado & Doya, 2007; Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Maia & Frank, 2011; Yechiam, Busemeyer, Stout, & Bechara, 2005). In standard RL models, the action values are assumed to be updated according to the reward prediction error (RPE), i.e., the difference between the actual reward and the expected reward. Numerous studies have noted that the magnitude of an update is biased depending on the sign of the RPE. The bias is represented in RL models by differential learning rates for positive and negative RPEs (Frank, Moustafa, Haughey, Curran, & Hutchison, 2007; Gershman, 2015, 2016). Such asymmetric RL models have been used in various contexts. Differential learning rates have been introduced to incorporate the neuroscientific knowledge that positive feedback and negative feedback are processed by separate neural substrates (Frank et al., 2007). The asymmetric RL models have also been used to explain risk-seeking/aversion behaviors (Niv, Edlund, Dayan, & O’Doherty, 2012). The larger learning rate for the negative RPE than that for the positive RPE leads to risk aversion, and the opposite leads to risk-seeking behaviors. Differential learning rates have also been related to cognitive bias, such as positivity bias and confirmation bias (Kuzmanovic & Rigoux, 2017; Lefebvre, Lebreton, Meyniel, Bourgeois-Gironde, & Palminteri, 2017), in studies on humans; the positivity bias refers to the tendency to weigh “positive news” compared to “negative news”, whereas confirmation bias refers to the tendency to put weight on the outcomes that are consistent with their hypothesis (updates are larger when the chosen option produces good results, while updates are smaller when the unchosen option produces better results) (Palminteri, Lefebvre, Kilford, & Blakemore, 2017). The adaptive properties of asymmetric value updates has also been discussed in Cazé and van der Meer (2013). Cazé and van der Meer (2013) suggested the existence of a situation in which the steady-state behavior of asymmetric RL models yields better separation of the action values compared to symmetric RL models.

However, it is not yet known how the asymmetric value updates manifest in the behavior and what properties of behavioral data

E-mail address: kawahira@lit.nagoya-u.ac.jp.

https://doi.org/10.1016/j.jmp.2018.09.002
0022-2496/© 2018 The Author. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
the differential learning rates reflect. Can a researcher claim that asymmetric value updates underlie subjects’ behaviors by merely observing the choice sequence without fitting RL models to the data? Most previous studies examining asymmetric value updates have relied solely on the parameter estimates of RL-model fits. The availability of statistical testing without computational model fitting is desirable because computational model-based analysis can suffer from the influence of a confounding variable, as we show in this study.

The steady-state, asymptotic properties of asymmetric RL models have been clarified in Cazé and van der Meer (2013). However, parameter estimates obtained from model fitting reflect not only the steady-state but also transient trial-by-trial dynamics of behaviors (cf. Katahira, Yuki, & Okanoya, 2017). We thus especially focus on the transient properties of the behavioral consequences of asymmetric value updates. To achieve this, we perform theoretical analysis and numerical simulations that examine how the differential learning rates influence the impact of past experiences on current choices, following Katahira (2015). Specifically, we compare the RL models and logistic regression models that have been used for analyzing the impact of past events on future choices. Because regression models can explicitly represent how the reward and choice histories influence future choices, investigating the relation between RL models and regression models would be helpful to clarify the transient, statistical structure of asymmetric RL models (Katahira, 2015).

The results indicate the following. In asymmetric RL models, the extent to which past outcomes influence the present choice also depends on the outcomes of intermediate choices, unlike standard (symmetric) RL models. Based on this effect, we propose a statistical method for testing the existence of an asymmetry in value updates without relying on the RL model fit. The asymmetry in the value updates induces the autocorrelation of choice (i.e., the tendency to repeat the same choice or to switch to another choice irrespective of past outcomes). Consequently, if the RL model without intrinsic autocorrelation is fitted to data that possess intrinsic autocorrelation, a statistical bias that overestimates the difference in learning rates will arise. By simulation, we demonstrate that this can cause a statistical artifact leading to “pseudo-positivity bias” and “pseudo-confirmation bias” as a conclusion of RL-model fitting.

The remainder of this article is organized as follows. First, we introduce asymmetric RL models and a logistic regression model (Section 2). Second, we discuss the property of asymmetric RL models by investigating the relation between the RL models and the logistic regression model (Section 3). Third, we examine the statistical artifact that can arise from the basic properties of asymmetric RL models (Section 4). Fourth, we describe the proposed statistical method for testing the asymmetry in value updates without RL-model fitting (Section 5). Finally, we discuss several implications of our findings (Section 6).

2. Models

2.1. Reinforcement learning models with differential learning rates

Here, we introduce an RL model with differential learning rates for positive RPE and negative RPE. Hereafter, we refer to this model as the asymmetric RL model. Throughout this paper, we basically consider cases with only two options \( i = 1 \) or \( 2 \).\(^1\)

The model assigns each option \( i \) an action value \( Q_i(t) \), where \( t \) is the index of the trial. In common settings, the initial action values are set to zero (i.e., \( Q_1(0) = Q_2(0) = 0 \)). The model updates the action values depending on the outcome of the action. The outcome (reward) at trial \( t \) is denoted by \( R(t) \). We typically consider a binary outcome case whereby we set \( R(t) = 1 \) if a reward is given and \( R(t) = 0 \) if no reward is given, unless stated otherwise. In the asymmetric RL model, the action value for the chosen option \( i \) is updated as follows:

\[
\delta(t) = R(t) - Q_i(t),
\]

\[
Q_i(t + 1) = Q_i(t) + \alpha_i^+ \delta(t) \quad \text{if } \delta(t) \geq 0,
\]

\[
Q_i(t + 1) = Q_i(t) - \alpha_i^- \delta(t) \quad \text{if } \delta(t) < 0,
\]

where \( \alpha_i^+ \) and \( \alpha_i^- \) are the learning rates that determine how much the model updates the action value depending on the RPE, \( \delta(t) \). Although it is generally assumed that the sign of the learning rates depends on the RPE, in the binary outcome case, the sign of the RPE coincides with whether the reward is given \( (R(t) = 1) \) or omitted \( (R(t) = 0) \). For the unchosen option \( j \) \((i \neq j)\), the action value is updated as follows:

\[
Q_j(t + 1) = Q_j(t) + \alpha_f (1 - \alpha_i^-) Q_j(t) - \alpha_f \delta(t),
\]

where \( \alpha_f \) is the forgetting rate. The forgetting rate is typically set to zero \( \alpha_f = 0 \) (i.e., the action value of the unchosen option is not updated). However, many studies have reported that zero-forgetting rates improve the goodness of fit to behavioral data (e.g., Ito & Doya, 2009; Katahira et al., 2017; Toyama, Katahira, & Ohira, 2017).

Here, we denote the option that was chosen at trial \( t \) by \( a(t) \) \((= 1 \text{ or } 2)\). Based on the set of action values, the model assigns the probability of choosing option \( i \) using the softmax function:

\[
P(a(t) = i) = \frac{\exp(\beta Q_i(t))}{1 + \exp(-\beta (Q_i(t) - Q_j(t)))},
\]

where \( \beta \) is called the inverse temperature parameter, which determines the sensitivity of the choice probabilities to differences in action values. It has been shown that humans tend to repeat recent choices, a tendency called “choice perseveration”, “choice stickiness”, or “decision inertia”. Such a tendency is often explicitly modeled by a choice-autocorrelation factor that is added to the action value as follows (Akaishi, Umeda, Nagase, & Sakai, 2014; Gershman, Pesaran, & Daw, 2009; Huys et al., 2011):

\[
P(a(t) = i) = \frac{\exp(\beta Q_i(t) + \psi C_i(t))}{\sum_{i=1}^2 \exp(\beta Q_i(t) + \psi C_i(t))},
\]

where \( C_i(t) \) represents the choice trace, which quantifies how frequently option \( i \) was chosen recently. The choice trace weight \( \psi \) is a parameter that controls the tendency to repeat (when positive) or avoid (when negative) recently chosen options. The choice trace is computed using the following update rule (Akaishi et al., 2014):

\[
C_i(t + 1) = (1 - \tau) C_i(t) + \tau I(a(t) = i),
\]

where the indicator function \( I(.) \) takes on a value of 1 if the statement is true and 0 if the statement is false. The parameter \( \tau \) is the decay rate of the choice trace. The initial values are set to zero, i.e., \( C_1(1) = C_2(1) = 0 \). Many RL models that represent the choice perseverence or choice stickiness only incorporate the influence of the immediate prior trial (e.g., Gillan, Kosinski, Whelan, Phelps, & Daw, 2016; Huys, Moutoussis, & Williams, 2011). Such a convention can be represented by setting \( \tau = 1 \).

When the choice is binary, the above choice-autocorrelation factor is equivalently implemented by

\[
C(t + 1) = \begin{cases} (1 - \tau) C(t) + \tau & \text{if } a(t) = 1, \\ (1 - \tau) C(t) - \tau & \text{if } a(t) = 2, \end{cases}
\]
with the soft max function:
\[
P(a(t) = 1) = \frac{1}{1 + \exp(-\beta [Q_i(t) - Q_j(t)])
\]
We implemented the model with this expression in the following simulations.

Examples of the model behavior are shown in Fig. 1. In Fig. 1A, the update was pessimistic \((\alpha_i^+ < \alpha_i^-)\): the magnitude of the decrease in action values when no reward is obtained is larger than that of the increase in action values when a reward is obtained (bottom panel). In this case, the model often changes the choice probability and switches the choice frequently because the larger decrease in an action value when the model does not obtain a reward decreases the probability of the same choice. In contrast, in an optimistic case (Fig. 1B, \(\alpha_i^+ > \alpha_i^-\)), the omission of the reward causes a smaller effect on the action value. This makes the model likely to repeat the same choice.

2.2. Logistic regression

Next, we introduce a regression model that predicts current choice from the histories of rewards and choices in previous trials (e.g., Corrado, Sugrue, Seung, & Newsome, 2005; Kovach et al., 2012; Lau & Glimcher, 2005; Seymour, Daw, Roiser, Dayan, & Dolan, 2012). Following the convention of Corrado et al. (2005) and Lau and Glimcher (2005), we represent the reward history \(h(t)\) as follows:

\[
r(t) = \begin{cases} 
1 & \text{if option } 1 \text{ is chosen and a reward is given at trial } t \\
-1 & \text{if option } 2 \text{ is chosen and a reward is given at trial } t \\
0 & \text{if no reward is given at trial } t.
\end{cases}
\]

We represent the choice history \(c(t)\) as follows:

\[
c(t) = \begin{cases} 
1 & \text{if option } 1 \text{ is chosen at trial } t \\
-1 & \text{if option } 2 \text{ is chosen at trial } t.
\end{cases}
\]

With these history variables, the logistic regression model is constructed as
\[
p(a(t) = 1) = \frac{1}{1 + \exp(-h(t))}.
\]

\[
h(t) = \sum_{m=1}^{M_r} b_r^{(m)}(t - m) + \sum_{m=1}^{M_c} b_c^{(m)}c(t - m),
\]

where \(b_r^{(m)}\) and \(b_c^{(m)}\) are the regression coefficients for the trial \(m\) trials ago. The constants \(M_r\) and \(M_c\) are the history length for the reward history and the choice history, respectively. We estimate the regression coefficients using the maximum likelihood method.

3. Relation between RL models and regression models

3.1. Theoretical consideration

Following Katahira (2015), we discuss the relation between the parameters of the asymmetric RL model and the regression coefficients of the logistic regression model. Let \(R_i(t)\) denote the reward value obtained from choosing option \(i\) at trial \(t\). If option \(i\) is not chosen at trial \(t\), \(R_i(t) = 0\). Then, the update rules (Eqs. (2) and (4)) can be summarized as

\[
Q_i(t+1) = Q_i(t) + \alpha_i^+(R_i(t) - Q_i(t))
\]

\[= (1 - \alpha_i^+)Q_i(t) + \alpha_i^+R_i(t),\]

where

\[
\alpha_i^+ = \begin{cases} 
\alpha_i^+ & \text{if } a(t) = i \text{ and } \delta(t) \geq 0 \\
\alpha_i^- & \text{if } a(t) = i \text{ and } \delta(t) < 0 \\
\alpha_F & \text{if } a(t) = j (\neq i).
\end{cases}
\]

Let us consider the influence of the outcome at trial \(t - k\) after choosing option \(i\) (i.e., \(R_i(t - k)\)) on the trial \(t + 1\). Expanding Eq. (13) back into the past (see Appendix A; Katahira, 2015), we can confirm that the term \(R_i(t - k)\) is multiplied by \(\alpha_i^+\) at trial \(t - k\), by \((1 - \alpha_i^-)\) at the next trial \(t - k + 1\), by \((1 - \alpha_i^-\delta)\) at trial \(t - k + 2\), etc. A specific example is shown in Fig. 2A. In this case, the impact of the outcome given at trial \(t - 3\) is \(\alpha_i^+\) at trial \(t - 2\) due to the update rule for \(\delta(t - 3) > 0\). In the next trial, \(t - 2\), the reward was absent \((R_i(t - 2) = 0)\), and thus, \(\delta(t - 2) < 0\). Therefore, the impact of \(R_i(t - 3)\) is multiplied by \((1 - \alpha_i^-)\) at trial \(t - 1\). Because the option is unchosen at trial \(t\), \((1 - \alpha_F)\) is multiplied by the impact for the next trial \(t + 1\).

Let \(n^+, n^-, n_F\) be the numbers of trials for each trial type after trial \(t - k\) and before trial \(t + 1\) \((n^+\): for positive RPE trials with option \(i\) being chosen; \(n^-\): for negative RPE trials with option \(i\) being chosen; \(n_F\): for trials with option \(i\) being not chosen). By construction, we have the relationship \(n_F = k - 1 - n^+ - n^-\).

The impact of the reward provided at \(t - k\) on the choice at trial \(t + 1\) is given as a function of \(n^+, n^-, n_F\):

\[
\text{Impact of a reward on trial } t + 1 = \alpha_i^+ (1 - \alpha_i^-)^{n^+} (1 - \alpha_i^-)^{n^-} (1 - \alpha_F)^{n_F}
\]

\[
\times (1 - \alpha_F)^{n_F}.
\]

In the special case wherein \(\alpha_i^+ = \alpha_i^- = \alpha_F (\equiv \alpha)\), this reduces to

\[
\alpha_i(1 - \alpha)^{n^+ - 1},
\]

which is a constant that is independent of subsequent outcomes. From this fact, one can find that if the history length of the regression model is sufficiently large, this RL model is equivalent to a logistic regression model whose coefficients are written with the parameters of the RL model as \(b_r^{(k)} = \beta \alpha(1 - \alpha)^{k-1}\), where \(\beta\) is the inverse temperature of the RL model (Katahira, 2015).

Eq. (15) indicates that if the learning rates differ depending on the outcome \((\alpha_i^+ \neq \alpha_i^-)\), the influence of the reward (at trial \(t - k\)) on the future choice (at trial \(t + 1\)) depends on the outcomes of subsequent choices at trials \(t - k + 1, t - k + 2, ..., t\). An example of the impact of past rewards is shown in Fig. 2B. This property violates the assumption of the logistic regression model that assumes that the influence of a past event on the current choice is independent of the events at other trials. We can utilize this property to test the asymmetry in value updates, as we will discuss later.

3.2. Simulation results

To understand the properties of asymmetric value updates, we conducted simulations that examine the relation between the asymmetric RL models and the logistic regression model. In the simulations, we first generated the choice data from RL models used to perform a hypothetical decision-making task and then fitted the logistic regression model to the data (the detailed procedure is given in Appendix B). Here, the RL models were set to have no intrinsic choice autocorrelation (\(\varphi = 0\)).

The first case varied the balance of the learning rates \((\alpha_i^+ + \alpha_i^-)\) while the sum of the learning rates was fixed as \(\alpha_i^+ + \alpha_i^- = 0.8\) and the forgetting rate was fixed as \(\alpha_F = 0.4\) (Fig. 3). Fig. 3A compares the predictions of the RL models and regression models. Each dot represents the prediction for one trial. If the two models' predictions perfectly agree, the samples should lie on the diagonal line. As the theory predicted, when the learning rates and the forgetting rate are identical \((\alpha_i^+ = \alpha_i^- = \alpha_F = 0.4)\), the two models exhibited perfect agreement. In contrast, when there is an asymmetry \((\alpha_i^+ \neq \alpha_i^-)\), mismatches between the predictions of the two models (discrepancies from the diagonal lines) were observed. This is because of the change in the impact of past outcomes on future choices, depending on the outcomes of
Fig. 1. Examples of RL model behavior for the simulated task of a pessimistic update case (A; \(\alpha^+ = 0.2, \alpha^- = 0.6\)) and an optimistic update case (B; \(\alpha^+ = 0.7, \alpha^- = 0.1\)). For both cases, the first 100 trials of one hypothetical subject are depicted. The probability of choosing option 1 was calculated from the “true” RL model (bold, gray lines) and the logistic regression model (thin, black lines). The corresponding action values (Q-values) of the RL model are shown at each bottom panel. The other common parameter settings were \(\alpha_F = 0.4\), \(\phi = 0.0\) (no choice-autocorrelation) and \(\beta = 3.0\). In the first 50 trials, option 1 was optimal (the probability of reward associated with option 1 was 0.7 and that with option 2 was 0.3). At the 51-st trial, the reward probability was reversed.

Fig. 2. Schematic of how the impact of past rewards on the current choice changes. (A) An example illustrating how the weight for a past reward (here, \(R\) at trial \(t - 3\)) decays as the agent experiences new trials. (B) The impact of the reward on the current choice as a function of trial lag and subsequent experience. This figure is created by setting the learning and forgetting rates as \(\alpha^+ = 0.6, \alpha^- = 0.2\), and \(\alpha_F = 0.4\).
intermediate choices, as described in Eq. (15). This dynamic change cannot be represented by a regression model that assumes that the influence of a reward is independent of subsequent events (without incorporating an interaction term, as we will discuss in Section 5).

Although the logistic regression models are not statistically equivalent with asymmetric RL models, the corresponding regression coefficients would be helpful in understanding the statistical properties of the RL models (Katahira, 2015). Fig. 3B shows the regression coefficients of the logistic regression models obtained by fitting to the simulated data. When the learning rates for the positive RPE and the negative RPE of the RL model were different, a dependence on the choice history arose (i.e., $b_c \neq 0$) even though the RL model had no intrinsic autocorrelation factor ($\psi = 0$). More specifically, when $\alpha_1^+ > \alpha_1^-$ (optimistic updates), a positive dependence on choice history is observed ($b_c > 0$). In contrast, when $\alpha_1^+ > \alpha_1^-$, a negative dependence is observed ($b_c < 0$). The dependence on the choice history arose through the following mechanism: Consider the effect of the outcome of trial $t-1$. If $\alpha_1^+ > \alpha_1^-$, the value of chosen action at trial $t-1$ can gain more from a reward than it can lose by a non-reward, which makes the repetition of the same choice at trial $t$ more likely than when the learning rates are identical ($\alpha_1^+ = \alpha_1^-$). This tendency can be captured by a regression model with a positive regression coefficient for the choice at trial $t-1$. On the other hand, if $\alpha_1^+ < \alpha_1^-$, the value of chosen action can lose more than it can gain, which makes switching the choice at trial $t$ more likely and leads to opposite consequences.

Next, we examined the effects of the forgetting rate when the balance of the learning rates is optimistic ($\alpha_1^+ = 0.8$, $\alpha_1^- = 0.2$; Fig. 4A). In this case, a lower forgetting rate results in a smaller impact of the choice history (Fig. 4A, right panel). When the forgetting rate was very small ($\alpha_F < 0.05$), the coefficients for the choice history were negative. The following reason for this is considered. When the forgetting rate is close to zero, the action value remains almost the same value if the option is unchosen (see Fig. 4B). Thus, in the case of optimistic updates, the omission of the reward causes a decay with $(1 - \alpha_1^-)$, the action value easily becomes lower than that of the opposite options. Thus, the model easily switches the choice compared to the regression model, which has no mechanism for retaining the action value of the unchosen option. In contrast, when the forgetting rate is large (e.g., Fig. 4C), the action value of the unchosen options quickly decays to zero, and thus, the action value of the chosen option tends to be superior to that of the other option. Because of this property, the model presents less alternation.

Fig. 3. The effects of the asymmetry in the value updates. (A) The scatter plot of the prediction regarding the current choice ($P(a|t-1) = 1$) derived from the RL models with varied learning rates and a fixed forgetting rate ($\alpha_F = 0.4$) and the logistic regression models. (B) The regression coefficients for the reward history (left) and the choice history (right). The common parameters of the RL models were set as $\psi = 0.0$, $\tau = 0.0$, and $\beta = 3.0$.
4. Biases in parameter estimates of asymmetric updating models

The results in the previous section suggest that the differential learning rates for positive RPE and negative RPE cause an apparent choice autocorrelation (i.e., a tendency to continue with the same option or switch choices), even when the true model (the data generation process) has no explicit mechanism for generating such an autocorrelation. This property raises important questions. If the true model has an intrinsic choice autocorrelation and the model
without the autocorrelation factor but with differential learning rates is fitted to the data, is there a possibility that the estimated learning rates are biased such that they can represent the choice autocorrelation? In addition, are the true choice autocorrelation and the effects of the differential learning rates statistically distinguishable? To address these issues, we performed simulations of RL-model fitting and statistical model selection. Specifically, we simulated the two experiments reported in Palminteri et al. (2017). In Experiment 1 of Palminteri et al. (2017), they reported a “positivity bias”, where the estimate of the learning rate for positive RPE was larger than that for the negative RPE ($\alpha_+ > \alpha_-$), while in Experiment 2, they reported a “confirmation bias”, which we explain later. We examine whether these estimates can be reproduced by assuming that the true model (data generation process) has an intrinsic choice-autocorrelation factor but no asymmetry characterizing the value updates.

### 4.1. Pseudo-positivity bias

First, we examined whether the “pseudo-positivity bias” ($\alpha_+ > \alpha_-$) can arise even when the true process has no such bias but has a positive choice autocorrelation (i.e., choice perseverance). In Cases 1 and 2, we generated hypothetical data from true models with identical learning rates ($\alpha_+ = \alpha_- = 0.5$) with intrinsic choice-autocorrelation factors (Fig. 6). The models performed the task with the same structure of Experiment 1 in Palminteri et al. (2017) (see Appendix B for details). In Case 1, the true model has an impulsive autocorrelation (the model has a tendency to repeat the same choice as in the previous trial; $\tau = 1.0$, $\phi = 2.5$). In Case 2, the true model has a gradual autocorrelation ($\tau = 0.4$, $\phi = 2.5$). The other true parameters were $\beta = 3.0$ and $\alpha_0 = 0.0$ (no forgetting). We choose these parameter values so that the model can approximately reproduce the learning curve of Experiment 1 in Palminteri et al. (2017). We fitted six models (Table 1). Model 1–Model 3 were allowed to have differential learning rates depending on the valence of RPE, whereas Models 4–6 were not. Model 1 and Model 4 had no choice-autocorrelation factor, Models 2 and 5 could have an impulsive autocorrelation ($\tau$ was fixed to 1), and Models 3 and 6 could have a gradual choice autocorrelation ($\tau$ was a free parameter; Table 1). For all candidate models, the model parameters were fitted to each hypothetical subject’s data using the maximum likelihood method as described in Appendix B.

The candidate models were evaluated based on Akaike’s Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Both criteria introduce a penalty for the number of parameters in the model (see B.2). A lower value indicates a better model. More specifically, AIC is a measure of the predictive ability of a model for future data: a smaller AIC value indicates better prediction ability. BIC approximates the model evidence which quantifies how likely the model is. A smaller BIC value indicates that the model is more likely. Thus, to examine which model is likely to be the true model from the data, BIC would be more appropriate. However, given that both criteria are commonly used in RL model fitting, and BIC generally tends to more strongly penalize models with more parameters than AIC (cf., Katahira et al., 2017), we report the values of both criteria.

Table 1 shows AIC and BIC values (mean across subjects) for each model. For both cases, both criteria selected the models with identical learning rates. For Case 1, Model 5, which has an impulsive choice-autocorrelation factor, as does the true model,

yielded the smallest mean value of AIC and that of BIC, indicating that this was the best model for the simulated data. However, if the candidate models are restricted to the models with no choice-autocorrelation factor (Model 1 and Model 4), Model 1, which possesses the differential learning rates, was selected as the best model in both cases despite the fact that the true learning rates did not differ. This is because the asymmetric RL models can represent a positive choice autocorrelation, as we discussed in the previous section. Fig. 6 shows the resulting estimates of the learning rates in Models 1, 2, and 3 for both cases. For Case 1, the learning rate for the positive outcome $\alpha_+$ was significantly higher than that for the negative outcome $\alpha_-$ for Model 1, which has no choice autocorrelation ($t(19) = 10.43, p < .001$, paired t-test), and also for Model 2, which has an impulsive choice autocorrelation ($t(19) = 2.10, p = .049$, paired t-test).

For Case 2, in which the true model had a gradual choice autocorrelation, Model 6, which has the same structure as the true model, was selected by both AIC and BIC (Table 1). If the candidate models were restricted to the models that cannot have a gradual choice-autocorrelation factor (Models 1, 2, 4, and 5), both criteria selected Model 2, which has differential learning rates and an impulsive autocorrelation. This is because the asymmetric RL models can represent the impact of the choices made more than one trial ago, which cannot be represented by the impulsive choice-autocorrelation factor. These results suggest that including a model with only an impulsive autocorrelation factor is not sufficient when there is a gradual perseverance in the true model; models with an asymmetric RL model achieve a better fit to the actual gradual autocorrelation than the symmetric RL model with the impulsive autocorrelation. The learning rate for the positive outcome $\alpha_+$ was significantly higher than that for the negative outcome $\alpha_-$ for Model 1 and Model 2, which do not have a gradual choice-autocorrelation factor ($t(19) = 7.55, p < .001$ and $t(19) = 2.63, p = .017$, respectively; Fig. 6). The difference in the estimates of the learning rates was suppressed in the parameter estimates for Model 3, which includes the gradual autocorrelation factor ($t(19) = 0.38, p = .705$).

As a summary, an overestimation of the differences in the learning rates can be observed when a fitted model did not have a choice-autocorrelation factor that can represent the intrinsic choice autocorrelation in the true model. This effect can be regarded as a “pseudo-positivity bias”.

We conducted an additional simulation (Case 3), in which we generated hypothetical data from the true model, where the differential learning rates were indeed different (optimistic) ($\alpha_+ = 0.5$, $\alpha_- = 0.2$) but with no choice-autocorrelation factor ($\phi = 0$). In this case, both criteria correctly selected Model 1, which has the same structure as the true model (Table 1). This result indicates that the effects of asymmetric value updates and that of intrinsic choice perseverance are similar but statistically distinguishable based on the goodness of fit to data.

### 4.2. Pseudo-confirmation bias

Next, we consider a further example of statistical artifacts as a consequence of the similarity between the asymmetric value updates and choice perseverance. Palminteri et al. (2017) reported that when the forgone outcomes (i.e., the outcomes of the unchosen option) are given, people tend to greatly update values when a negative prediction error is given compared to when a
positive prediction error is given. A negative outcome from the unchosen option conforms to the “hypothesis” of a decision maker (i.e., “This option will yield a worse outcome compared to another option.”), and a positive outcome does not. Thus, Palminteri et al. (2017) deemed this tendency, together with the positivity bias regarding the obtained outcomes, as “confirmation bias”. Here, we examined whether the parameter estimates corresponding to the confirmation bias also arise as a mere consequence of the effects of the intrinsic choice perseverence of a decision maker.

The asymmetric RL model can incorporate the counterfactual learning as follows (Palminteri et al., 2017). The action value of chosen option $i$ is updated depending on the outcome from chosen option $R_i$ as

$$
\delta_i(t) = R_i(t) - Q_i(t),
$$

(17)

$$
Q_i(t+1) = \begin{cases} 
Q_i(t) + a_c^+ \delta_i(t) & \text{if } \delta_i(t) \geq 0 \\
Q_i(t) + a_c^- \delta_i(t) & \text{if } \delta_i(t) < 0
\end{cases}
$$

(18)

while that of unchosen option $j$ is updated depending on the forgone outcome $R_j$ as

$$
\delta_j(t) = R_j(t) - Q_j(t),
$$

(19)

$$
Q_j(t+1) = \begin{cases} 
Q_j(t) + a_u^+ \delta_j(t) & \text{if } \delta_j(t) \geq 0 \\
Q_j(t) + a_u^- \delta_j(t) & \text{if } \delta_j(t) < 0
\end{cases}
$$

(20)

The confirmation bias is represented by differential learning rates: $a_c^+ > a_c^-$ and $a_u^+ < a_u^-$. Our question here is whether such a

---

**Table 1**

Model selection results for the simulation of positivity bias.

| Simulation condition | Fitted model | Learning rates | choice-autocorrelation | # Free parameter | AIC | BIC |
|----------------------|--------------|----------------|-------------------------|----------------|-----|-----|
| Case 1               | Model 1      | Differential  | None                    | 3              | 125.04 (32.64) | 134.81 (32.64) |
|                      | Model 2      | Differential  | Impulsive               | 4              | 82.71 (14.78)  | 95.74 (14.78)  |
|                      | Model 3      | Differential  | Gradual                 | 5              | 83.96 (14.95)  | 100.25 (14.95) |
|                      | Model 4      | Identical     | None                    | 2              | 153.15 (32.70) | 159.67 (32.70) |
|                      | Model 5      | Identical     | Impulsive               | 3              | **81.89 (15.04)** | 91.66 (15.04)  |
|                      | Model 6      | Identical     | Gradual                 | 4              | 83.16 (15.14)  | 96.19 (15.14)  |
| Case 2               | Model 1      | Differential  | None                    | 3              | 94.33 (26.75)  | 104.10 (26.75) |
|                      | Model 2      | Differential  | Impulsive               | 4              | 83.93 (18.76)  | **96.96 (18.76)** |
|                      | Model 3      | Differential  | Gradual                 | 5              | 82.41 (18.24)  | 98.70 (18.24)  |
|                      | Model 4      | Identical     | None                    | 2              | 111.38 (29.36) | 117.89 (29.36) |
|                      | Model 5      | Identical     | Impulsive               | 3              | 86.55 (17.96)  | 93.32 (17.96)  |
|                      | Model 6      | Identical     | Gradual                 | 4              | **81.91 (18.31)** | 94.94 (18.31)  |
| Case 3               | Model 1      | Differential  | None                    | 3              | 137.41 (18.70) | 147.18 (18.70) |
|                      | Model 2      | Differential  | Impulsive               | 4              | 138.83 (18.78) | 151.86 (18.78) |
|                      | Model 3      | Differential  | Gradual                 | 5              | 139.77 (18.92) | 156.06 (18.92) |
|                      | Model 4      | Identical     | None                    | 2              | 142.94 (19.05) | 154.46 (19.05) |
|                      | Model 5      | Identical     | Impulsive               | 3              | 141.86 (18.86) | 151.63 (18.86) |
|                      | Model 6      | Identical     | Gradual                 | 4              | 141.18 (18.70) | 154.21 (18.70) |

Standard deviations are in parentheses.

*Best model among models with no choice-autocorrelation factor.

*Best model when gradual auto-correlation models are excluded.
relation can appear in the parameter estimates even when the true learning rates are identical \((\alpha^C = \alpha^C = \alpha^D = 0.5)\). Similar to the simulation of the pseudo-positivity bias, we considered two cases regarding the choice-autocorrelation factor: with an impulsive autocorrelation (Case 1; \(\tau = 1.0, \psi = 3.5\)) and with a gradual autocorrelation (Case 2; \(\tau = 0.4, \psi = 3.5\)). The other true parameters were \(\beta = 1.5\) and \(\alpha_1 = 0.0\) (no forgetting). We again chose these parameter values so that the true model can approximately reproduce the learning curve for Experiment 2 in Palminteri et al. (2017). We fitted 10 models, which consisted of 6 models used in Palminteri et al. (2017) and an additional 4 models with choice-autocorrelation factors (Table 2). Model 1–Model 3 were allowed to have differential learning rates (all four learning rates can differ from each other), while in Model 8–Model 10, all four learning rates share the same value. Model 4–Model 7 are intermediate models with reduced parameters, all of which Palminteri et al. (2017) used. In the ‘Information’ model (Model 4), the learning rates can differ depending on whether the option is chosen or unchosen. This corresponds to the following restrictions: \(\alpha^C = \alpha^C = \alpha^D = \alpha^D\). In the ‘Valence’ model (Model 5), the learning rates can vary according to the valence (positive or negative) of the RPE, resulting in the restrictions \(\alpha^C = \alpha^D = 0\) and \(\alpha^C = 0\). In the ‘Confirmation’ model (Model 6 and Model 7), the learning rates can differ depending on whether the outcomes are “confirmatory” (positive obtained or negative forgone) or “disconfirmatory” (negative obtained or positive forgone). This corresponds to the following restrictions: \(\alpha^C = \alpha^D = \alpha^C = \alpha^D\). For this variant, Palminteri et al. (2017) also included the version with an impulsive choice-autocorrelation factor (Model 7).

The estimated learning rates in the models whose four learning rates were allowed to have different values (Model 1–Model 3) are shown in Fig. 7. As expected, when the model with no choice-autocorrelation factor (Model 1) was fitted to the generated data, the strong bias in the estimates of the learning rates appeared. The resulting estimates of the learning rates correspond to “confirmation bias”. Even when the fitted model has impulsive choice perseverance (Model 2), the confirmation bias appeared if the true model possessed gradual choice perseverance (Case 2, Fig. 7). When the true model has only impulsive choice perseverance, the confirmation bias was suppressed if the fitted model can also have impulsive perseverance (Case 1, Fig. 7).

In Case 1, in which the true model had impulsive perseverance, both model selection criteria selected Model 9, which has the same structure as the true model (Table 2). However, if the candidate models are restricted to the models with no choice-autocorrelation factor (Models 1, 4, 5, 6, and 8), the ‘Confirmation’ model with no autocorrelation factor (Model 6) was selected by both AIC and BIC. The parameter estimates of the ‘Confirmation’ models are also shown in Fig. 7. The learning rate for confirmatory outcomes, \(\alpha^C\), in Model 6 was significantly higher than that for disconfirmatory outcomes, \(\alpha^D\) \((t(19) = 26.77, p < 10^{-15})\). In contrast, there was no significant difference in the learning rates for Model 7, which has impulsive perseverance \((t(19) = 1.05, p = 0.306)\).

In Case 2, in which the true model had gradual perseverance, Model 10, which has the same structure as the true model, showed the smallest AIC value, whereas the ‘Confirmation’ model with no autocorrelation factor (Model 6) presented the smallest BIC value. This is because Model 6 has one less parameter than Model 10; as we mentioned above, BIC tends to heavily penalize models with more parameters. When the models with gradual choice-autocorrelation factor were not included in the candidate models (i.e., Models 3 and 10 are excluded), AIC also selected Model 6. In this case, for both Model 6 (with no perseverance) and Model 7 (with impulsive perseverance), the two learning rates of the ‘Confirmation’ models \((\alpha^C\) and \(\alpha^D\)) showed a significant difference \((t(19) = 11.59, p < 10^{-9}\) and \(t(19) = 6.46, p < 10^{-3}\), respectively).

Taken together, the statistical artifact causes an apparent confirmation bias to arise when the true model has an intrinsic choice-autocorrelation factor while the fitted model does not have this factor. If the model that can represent the true structure of choice perseverance is included, the model is selected based on a model selection criterion. However, if all the candidate models cannot represent the underlying choice perseverance, the models with differential learning rates can be selected.

5. Model-neutral analysis of asymmetry in learning

We have demonstrated that confounding factors can systematically bias parameter estimation. As a confounding factor, we have considered the exponentially decaying choice-autocorrelation factors. There might be other types of confounding factor characterized by a different functional form. In reality, it is generally difficult for a researcher to include all of the possible confounding factors into their candidate models. Thus, using non-parametric, model-neutral analysis is helpful to make the conclusion strong. Here, based on the theoretical consideration given in Section 3.1 (as illustrated in Fig. 2), we propose a model-neutral method to test the asymmetry in value updates.

5.1. Proposed analysis

Here, we again focus on the factual learning, in which only the outcome from the chosen option is presented. As we have discussed, one behavioral consequence of differential learning is that the impact of a past outcome depends on subsequent outcomes. This indicates that there is an interaction between the outcome at trial \(t\) and that at trial \(t - 1\) \((R(t)\) and \(R(t - 1)\)) as a factor that influences the current choice, \(a(t + 1)\). Thus, to claim an asymmetry in the value updates, we need to determine whether there is such an interaction. Here, we denote the three successive trials by \(t + 1\), \(t\), and \(t - 1\). To simplify the analysis, we select the trials in which the subjects choose the same option at trial \(t\) and at trial \(t - 1\) (trials \(t\) and \(t - 1\) such that \(a(t) = a(t - 1)\)). Next, we construct the logistic regression model to predict the probability that the subject chooses the same option at trial \(t + 1\) \((i.e., a(t + 1) = a(t - 1))\), denoted by \(p(stay(t + 1))\). The regressors are \(R(t)\) and \(R(t - 1)\), and their interaction term is \(R(t) \times R(t - 1)\). This constitutes the logistic regression model:

\[
\begin{align*}
\log \left( \frac{p(stay(t + 1))}{p(switch(t + 1))} \right) &= b_0 + b_1 R(t) + b_2 R(t - 1) \\
&\quad + b_3 R(t) R(t - 1),
\end{align*}
\]

where \(p(stay(t + 1))\) is the probability that the subject switches the choice at trial \(t + 1\) \((i.e., a(t + 1) \neq a(t - 1))\). The intercept \(b_0\) represents the overall tendency to repeat the same choice, which may absorb the effects of the choice autocorrelation and the influence of the rewards provided before trial \(t - 1\).

The proposed method might apply to the counter-factual learning task if one includes the outcomes from the unchosen option as regressors in a similar way as done for the chosen option. A future study would examine this extension.
5.2. Theoretical consideration

If we ignore the events before trial \( t - 1 \), the asymmetric-RL model yields the stay probability as

\[
p(\text{stay}(t+1)) = \frac{1}{1 + \exp(-\beta h(t+1))},
\]

\[
h(t + 1) \approx \alpha_t^+ R(t) + \alpha_t^- (1 - \alpha_t^+) R(t - 1) \quad \text{if} \quad \delta(t) \geq 0
\]

\[
= \alpha_t^+ R(t) - \alpha_t^+ R(t - 1) \quad \text{if} \quad \delta(t) < 0.
\]

Here, we consider the condition whereby \( R(t) \) takes on a binary value \( <1\) if a reward is given and \( <0\) if no reward is given), and thus, \( h(t) > 0 \) if \( R(t) = 1 \) and \( h(t) < 0 \) if \( R(t) = 0 \)(given action values are not exactly 0 or 1). Under this condition, Eq. (23) can be rewritten as

\[
h(t + 1) \approx \alpha_t^+ R(t) + \alpha_t^+ (1 - \alpha_t^-) R(t - 1) + \alpha_t^- (\alpha_t^- - \alpha_t^+) R(t - 1).
\]

A comparison of Eqs. (21) and (24) indicates that the following relation holds if we can neglect the effects of rewards given more than two trials ago:

\[
b_1 = \beta \alpha_t^+,
\]

\[
b_2 = \beta \alpha_t^+ (1 - \alpha_t^-),
\]

\[
b_{12} = \beta \alpha_t^+ (\alpha_t^- - \alpha_t^+).
\]

This indicates that the interaction term \( b_{12} \) represents the degree of the difference in the learning rates. If \( b_{12} \) significantly differs from zero, one can claim an asymmetry in the value updates. The negative \( b_{12} \) indicates optimistic updates (the learning rate for a positive outcome, \( \alpha_t^+ \), is larger than that for a negative outcome, \( \alpha_t^- \)), whereas a positive \( b_{12} \) indicates pessimistic updates (a learning rate for a positive outcome, \( \alpha_t^+ \), is smaller than that for a negative outcome, \( \alpha_t^- \)).

5.3. Simulation

The proposed analysis relies on the assumption that one can neglect the events from more than two trials ago. Next, we examine whether this assumption causes undesirable statistical bias. To achieve this and to demonstrate how the proposed method works, we conducted simulations based on data generated from various RL models that perform the hypothetical learning task (see Appendix B.3 for details). For each simulation, the choice datasets of the 20 hypothetical subjects, each consisting of 200 trials, were generated from an RL model.

Fig. 8 shows the estimated regression coefficients. Broken horizontal lines represent theoretical values given by Eqs. (25)–(27), which ignore the impact of events at more than two trials ago. When the true learning rates were identical \( \alpha_t^+ = \alpha_t^- = 0.4 \), the theory predicts that the coefficient of the interaction term, \( \beta R(t - 1) \), is zero (i.e., \( b_{12} = 0 \), see Eq. (27)). For each simulation, the hypothesis test for the interaction term was conducted with a significance level \( \alpha = 0.05 \). The mean values over 1,000 simulations (indicated by bars) almost agreed with this prediction (panel A: case without choice-autocorrelation factor; panel B: case with gradual choice-autocorrelation factor). Accordingly, the false positive rates, the fraction of the simulations where the null hypothesis \( b_{12} = 0 \) was rejected, were 0.045 and 0.060. These false positive rates were close to the intended type I error rate \( \alpha = 0.05 \), validating the proposed analysis.

In contrast, when the true learning rates differed between the positive and negative outcomes, the interaction coefficients were
non-zero (negative in the case of optimistic updates, $a_{i}^{-} > a_{i}^{+}$, panel C, D; positive in the case of pessimistic updates, $a_{i}^{+} < a_{i}^{-}$, panel E, F). For these cases, the mean estimates of the regression coefficients (bars) deviated from the theoretical values (broken horizontal lines) because of the ignored impacts of previous events. However, the relative values of the coefficients agreed with the theoretical values. Because the impact of the choice autocorrelation in the true model was absorbed into the intercept, $b_{0}$, it did not inflate the other coefficients (right-hand-side panels). The significant interactions were observed in 33%–53% of the simulations (“p(sig.)” in Fig. 8).

In these simulations, we simply assumed that all the subjects had the same parameters in both the true models and the fitted regression models (i.e., all the regression coefficients were treated as fixed effects). In a real-world case, the parameters should vary...
across subjects. To address the individual differences in the parameters, it would be better to use mixed-effect models (e.g., implemented with the lme4 package, Bates, Mächler, Bolker, & Walker, 2015), which can treat the main effects of outcomes and their interactions as random effects, instead of the logistic regression models with fixed effects (cf, Gillan et al., 2016).

6. Discussion

In computational model fitting based on likelihoods, the computational model (e.g., RL model) is regarded as a statistical (probabilistic) model. Understanding the statistical structure of a model (e.g., the dependence of choice probabilities on past events) is crucial for understanding the type of properties of the data that the results reflect. In this study, we have investigated the statistical structure of the RL models with asymmetric (biased) value updates. Although the steady-state behavior of the asymmetric RL models have been discussed in a prior study (Cazé & van der Meer, 2013), the transient property of the choice behavior (i.e., how the current choice depends on the recent rewards and choices), which is critical for model fitting, has not yet been addressed.

We have demonstrated that the influence of the choice outcomes on future choices also depends on the outcomes of intermediate choices in the asymmetric RL models. This property is not directly represented in conventional regression models that assume the independence of the impact of past events. If the logistic regression model is fitted to the behavioral data generated based on asymmetric value-updates, an apparent choice history dependence is observed. This suggests that asymmetric value updates cause choice autocorrelation. Our simulation results suggest that this effect leads to a statistical artifact when asymmetric-RL models are fitted to data with an intrinsic choice autocorrelation (perseverance). The choice autocorrelation can be a critical confounding factor in estimating the differential learning rates. The results of the present study demonstrate the problem of model misspecification, which leads to an incorrect interpretation. Although such a type of pitfall of computational-model-based analysis has been discussed based on other types of model misspecification (Katahira, Bai, & Nakao, 2017; Nassar & Gold, 2013), the effect presented in this paper is perhaps the most striking example of the pitfall.

Of course, the existence of such an artifact in parameter fitting does not necessarily suggest that the previous results that have
reported asymmetries in learning are due to this artifact. However, we should be careful when interpreting the results. In addition, re-analysis of data with RL models with choice-autocorrelation factor may be desirable. The simulation in this study clarified that including only the impulsive choice-autocorrelation factor (a tendency to repeat the choice made in the previous trial), as many studies performing RL-model fitting have done, might not be sufficient; including models with a choice-autocorrelation factor that can account for the autocorrelation between distant trials is desired.

The proposed model-neutral statistical test, which employs the logistic regression with interaction term of outcomes in the past two trials, would be useful for claiming that asymmetric value update indeed occur, as we discussed in Section 5. However, the statistical power of the test is relatively weak, as the simulation results showed. Thus, a large data set, which may be obtained through online experiments, is required (e.g., Gershman, 2015, 2016; Gillan et al., 2016). Combining the model-neutral analysis for such large data sets and computational models fit to a small number of but carefully collected data sets would be effective. Also, the proposed analysis only addresses the impact of the influence of the two most recent trials. The use of (parametric) computational models is useful to check the long-term history dependence of action selection. Thus, model-neutral analysis and model-based analysis have their own advantages and can be used in a complementary fashion.

In the simulation of RL model fitting presented in Section 4, we considered the maximum likelihood method for parameter estimation. One might wonder whether the results presented there are applicable to Bayesian approaches, including hierarchical Bayesian methods, which have been also used for fitting RL models (Ahn, Haines & Zhang, 2017; Ahn, Krawitz, Kim, Busemeyer, & Brown, 2011). While hierarchical Bayesian parameter estimation tends to produce stable, better results by utilizing both group-level information and individual-level information (Katahira, 2016), there is no explicit mechanism to reduce statistical biases due to the confounding factor, which is not included in the fitted models. Therefore, we expect that the systematic bias reported in this paper can occur in such a hierarchical Bayesian approaches. We leave this issue to a future study.

The temporal modulation of learning rates is a core component of various learning models. For example, the notable property of Bayesian update models (e.g., Kalman filters) is the modulation of the learning rates depending on uncertainty about task structure (e.g., reward probability) (Behrens, Woolrich, Walton, & Rushworth, 2007; Daw, O’Doherty, Dayan, & Seymour, 2006; Mathys, Daunizeau, Friston, & Stephan, 2011). In addition, modulating value updates using the “model” (knowledge) of the task structure can account for the “model-based” control of choice (Toyama et al., 2017). The asymmetric RL models discussed in this study may be the simplest case of a wide range of learning models with learning rate modulation. However, the results presented in this study have several implications for various models. For example, the results tell us how modulating the learning rates changes the history dependence of choice. This study is an important starting point for exploring the various phenomena in learning and decision making.

Acknowledgments

This work was supported in part by JSPS KAKENHI, Japan Grant Numbers JP17H05946 and 18K03173. The author thanks Asako Toyama for stimulating discussions and helpful comments on this study. The author is also grateful to the associate editor Matt Jones and two anonymous reviewers for their insightful suggestions during the review process.

Appendix A. Analytical calculation

The general expression of the update rule (Eq. (13)) can be expanded back into the past as

\[ Q(t + 1) = \left(1 - \alpha_t^*\right)Q(t) + \alpha_t^*R(t) \]

\[ = \left(1 - \alpha_t^*\right)\left(1 - \alpha_{t-1}^*\right)Q(t) + \alpha_t^*R(t) \]

\[ = \left(1 - \alpha_t^*\right)^2 \left(1 - \alpha_{t-1}^*\right)^2 Q(t) + \left(1 - \alpha_t^*\right)\alpha_{t-1}^*R(t) + \alpha_t^*R(t) \]

\[ = \left(1 - \alpha_t^*\right)^3 \left(1 - \alpha_{t-1}^*\right)^3 Q(t) + \left(1 - \alpha_t^*\right)^2\alpha_{t-1}^2R(t) + \left(1 - \alpha_t^*\right)\alpha_{t-1}^*R(t) + \alpha_t^*R(t) \]

\[ = \ldots \]

By this expansion, we obtain

\[ Q(t + 1) = \prod_{j=i}^{t} \left(1 - \alpha_{t-j}^*\right) Q(1) + \sum_{k=0}^{t} \alpha_{t-k,i} Q_{t-k} \prod_{l=0}^{k-1} \left(1 - \alpha_{t-l-i}^*\right) R(t - k). \]  \hfill (28)

From this expression, we find that the impact of the reward at trial \( t - k \) on trial \( t + 1 \) is

\[ \alpha_{t-k,i}^* \prod_{l=0}^{k-1} \left(1 - \alpha_{t-l-i}^*\right). \]  \hfill (29)

This leads to Eq. (15).

Appendix B. Simulation settings

Here, we detail the settings for each simulation. All the simulations and statistical analysis were performed using R version 3.2.0 (R Core Team, 2015).

B.1. Regression analysis of simulated data generated from RL models

The procedures for the simulations that compare the RL models and the logistic regression model are as follows. First, we generated the choice data from the RL models that were applied to a probabilistic reversal learning task. As we mentioned in the introduction, we focus on the transient properties of model behavior. To investigate the transient properties, we used a task in which reward probability is volatile (temporally changing) so that the model behavior does not converge to a stable state. In this task, one of the options was associated with a high reward probability, 0.7, while other option was associated with a low reward probability, 0.3. With the probability for the chosen option, the reward was given (\( R(t) = 1 \)); otherwise, no reward was given (\( R(t) = 0 \)). After each 50-trial block, the contingencies of the two stimuli were reversed. We generated hypothetical data for 5000 sessions per condition, which resulted in 2,500,000 trials. Using the simulated data of the large number of trials, statistical estimation error can be neglected in the interpretation of the results (Katahira, 2015).

We subsequently fitted the parameters of the logistic regression model to the simulated data using the maximum likelihood method. Specifically, we used the “glm” function implemented in the R programming language, version 3.2.0 (http://cran.us.r-project.org). We set the history length of the reward history and choice history as \( M_r = 10 \) and \( M_c = 10 \), respectively.
B.2. Model-based analysis for factual and counter-factual learning task

The details of the simulations examining pseudo-positivity bias (in factual learning) and pseudo-confirmation bias (in counter-factual learning) are as follows. We generated hypothetical data based on the tasks with the same design as Palminteri et al. (2017). Specifically, the model performed two sessions, where one session consists of four types of choice contexts. Each context has different options (cues) from other contexts and other sessions. In each context, 24 choice trials were presented. In total, the model performed 192 choice trials (2 sessions × 4 choice contexts × 24 trials). In each trial, the model chose one option, resulting in winning (R = 1) or losing a point (R = −1) according to the outcome probability associated with each option. The associated winning probabilities \( R \times \text{text} \), \( \text{24 choice trials were presented. In total, the model performed 192 choice trials (2 sessions × 4 choice contexts × 24 trials). In each trial, the model chose one option, resulting in winning (R = 1) or losing a point (R = −1) according to the outcome probability associated with each option. The associated winning probabilities (option 1, option 2) were as follows: “Symmetric condition” \(- (.5, .5)\) for 24 trials (entire context), “Asymmetric condition 1” \(- (.75, .25)\) for 24 trials, “Asymmetric condition 2” \(- (.25, .75)\) for 24 trials, and “Reversal condition” \(- (.83, .17)\) for the first 12 trials and \(- (.17, .83)\) for the later 12 trials. In the simulation of factual learning, only the outcome of the chosen option was presented to the subject. In the simulation of counter-factual learning, the outcomes of both the chosen option and the unchosen option were presented. Because different cue stimuli were used for each choice context, the initial action values in the RL models were reset at the beginning of each choice context as \( Q_0(1) = Q_0(2) = 0 \).

For the factual learning task, we considered three RL models as the true models for generating data. In Case 1 and Case 2, the model had identical learning rates for both positive and negative RPEs \( \alpha^p_1 = \alpha^p_2 = 0.5 \). In Case 1, the model had impulsive autocorrelation \( \tau = 1.0, \varphi = 2.5 \), while in Case 2, the model had gradual autocorrelation \( \tau = 0.4, \varphi = 2.5 \). In Case 3, the true model had different learning rates \( \alpha^p_1 = 0.5, \alpha^p_2 = 0.2 \) but no autocorrelation factor \( \varphi = 0 \). Common true parameters for all cases were \( \beta = 3.0 \) and \( \alpha^f = 0.0 \) (no forgetting). For the counterfactual learning task, the true models were RL models of two types. In Cases 1 and 2, the true model possesses identical learning rates for all four conditions \( \alpha^c_1 = \alpha^c_2 = \alpha^u_1 = \alpha^u_2 = 0.5 \). In Case 1, the true model possesses impulsive autocorrelation \( \tau = 1.0, \varphi = 3.5 \), while in Case 2, the true model possesses gradual autocorrelation \( \tau = 0.4, \varphi = 3.5 \). The other true parameters were \( \beta = 1.5 \) and \( \alpha^f = 0.0 \) (no forgetting). In each case of each task, the simulations were run 20 times. The resulting data were regarded as those from 20 hypothetical subjects.

To estimate the model parameters, maximum likelihood estimation (MLE) was separately performed for data from each hypothetical subject. MLE searches a single parameter set to maximize the log-likelihood of the model for all trials. This maximization was performed using the rslnp 1.15 package, which implements the augmented Lagrange multiplier method with an SQP interior algorithm (Ghalanos & Theuissl, 2011).

To evaluate the fitted models, we used two standard model selection criteria: Akaike’s Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarz, 1978). They are defined as

\[
\text{AIC} = -2L + 2k, \quad (30)
\]

\[
\text{BIC} = -2L + k \log(T), \quad (31)
\]

where \( L \) is the maximum value of the log likelihood function for the model, \( k \) is the number of estimated parameters in the model, and \( T \) is the number of trials (here, \( = 192 \)). Both criteria introduce a penalty for the number of parameters in the model. For both criteria, a smaller value indicates a better model.

B.3. Model-neutral hypothesis test

The simulation that demonstrates the model-neutral hypothetical test is as follows. The data generation processes was basically the same as in B.1; the data were generated by the RL models that performed a probabilistic reversal learning task. The exception was that for each simulation, the choice data sets of the 20 hypothetical subjects, each consisting of 200 trials, were generated from an RL model. The simulation was performed 1000 times under each condition. The data from each simulation run were submitted to the proposed hypothesis test, as described in the main text.

References

Ahn, W. Y., Haines, N., & Zhang, L. (2017). Revealing neurocomputational mechanisms of reinforcement learning and decision-making with the hBayesDM package. *Computational Psychiatry*, 1, 2457.

Ahn, W. Y., Krawitz, A., Kim, W., Busemeyer, J. R., & Brown, J. W. (2011). A model-based fMRI analysis with hierarchical Bayesian parameter estimation. *Journal of Neuroscience, Psychology, and Economics*, 4(2), 95.

Akaieke, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.

Akiishi, K., Umeda, K., Nagase, A., & Sakai, K. (2014). Autonomous mechanism of internal choice estimate underlies decision inertia. *Neuron*, 87(1), 195–206.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. http://dx.doi.org/10.18637/jss.v067.i01.

Behrens, T., Woolrich, M., Walton, M., & Rushworth, M. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, 10(9), 1214–1221.

Cazé, R. D., & van der Meer, M. A. (2013). Adaptive properties of differential learning rates for positive and negative outcomes. *Biological Cybernetics*, 107(6), 711–719.

Corrado, G., & Doya, K. (2007). Understanding neural coding through the model-based analysis of decision making. *Journal of Neuroscience*, 27(31), 8178.

Corrado, G., Sugre, L. P., Seung, H. S., & Newsome, W. T. (2005). Linear-nonlinear-Poisson models of primate choice dynamics. *Journal of the Experimental Analysis of Behavior*, 84(3), 581–617.

Daw, N., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans’ choices and striatal prediction errors. *Neuron*, 69(6), 1204–1215.

Daw, N., O'Doherty, J., Dayan, P., Seymour, B., & Dolan, R. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095), 876–879.

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences*, 104(1), 16311–16316.

Gershman, S. J. (2015). Do learning rates adapt to the distribution of rewards? *Psychonomic Bulletin & Review*, 22(5), 1320–1327.

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, 71, 1–6.

Gershman, S. J., Pesaran, B., & Daw, N. (2009). Human reinforcement learning subdivides structured action spaces by learning effector-specific values. *Journal of Neuroscience*, 29(43), 13524–13531. http://dx.doi.org/10.1523/JNEUROCIR.2460-09.2009.

Ghalanos, A., & Theussl, S. (2011). Rslnp: General non-linear optimization using augmented Lagrange multiplier method, Version 1.15.

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal directed control. *eLife*, 5(MARCH2016), 1–24. http://dx.doi.org/10.7554/eLife.11305.

Huys, Q. J., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R. J., et al. (2011). Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. *PLoS Computational Biology*, 7(4), e1002028.

Huys, Q. J., Moutoussis, M., & Williams, J. (2011). Are computational models of any use to psychiatry? *Neural Networks*, 24(6), 544–551.

Ito, M., & Doya, K. (2009). Validation of decision-making models and analysis of decision variables in the rat basolateral amygdala. *Journal of Neuroscience*, 29(31), 9861.

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.

Katahira, K. (2016). How hierarchical models improve point estimates of model parameters at the individual level. *Journal of Mathematical Psychology*, 73, 37–59.

Katahira, K., Bai, Y., & Nakao, T. (2017). Pseudo-learning effects in reinforcement learning model-based analysis: A problem of misspecification of initial preference. *PsychArXiv*. http://dx.doi.org/10.17605/OSF.IO/A6HZQ.

Katahira, K., Yuki, S., & Okamoto, K. (2017). Model-based estimation of subjective values using choice tasks with probabilistic feedback. *Journal of Mathematical Psychology*, 79, 29–43.
Kovach, C. K., Daw, N., Rudrauf, D., Tranel, D., O’Doherty, J. P., & Adolphs, R. (2012). Anterior prefrontal cortex contributes to action selection through tracking of recent reward trends. *Journal of Neuroscience, 32*(25), http://dx.doi.org/10.1523/JNEUROSCI.5468-11.2012.

Kuzmanovic, B., & Rigoux, L. (2017). Valence-dependent belief updating: Computational alidation. *Frontiers in Psychology, 8.*

Lau, B., & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior, 84*(3), 555–579.

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour, 1*, 0067.

Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience, 14*(2), 154–162.

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience, 5.*

Nassar, M. R., & Gold, J. I. (2013). A healthy fear of the unknown: Perspectives on the interpretation of parameter fits from computational models in neuroscience. *PloS Computational Biology, 9*(4), e1003015.

Niv, Y., Edlund, J., Dayan, P., & O’Doherty, J. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience, 32*(2), 551–562.

Palminteri, S., Lefebvre, G., Klöff, E. J., & Blakemore, S. J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS Computational Biology, 13*(8), e1005684.

R Core Team, (2015). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.

Seymour, B., Daw, N., Roiser, J. P., Dayan, P., & Dolan, R. (2012). Serotonin selectively modulates reward value in human decision-making. *Journal of Neuroscience, 32*(17), 5833–5842. http://dx.doi.org/10.1523/JNEUROSCI.0053-12.2012.

Toyama, A., Katahira, K., & Ohira, H. (2017). A simple computational algorithm of model-based choice preference. *Cognitive, Affective, & Behavioral Neuroscience, 17*, 764–783.

Yechiam, E., Bussemeyer, J., Stout, J., & Bechara, A. (2005). Using cognitive models to map relations between neuropsychological disorders and human decision-making deficits. *Psychological Science, 16*(12), 973–978.