Elucidation of the cognitive computational process underlying the behavior of pursuing the unprofitable targets

（手に入らない対象を追求する行動の基盤にある認知計算過程の解明）

SUGAWARA, Michiyo

(菅原　通代)

Doctor of Informatics

Graduate School of Informatics, Nagoya University

(名古屋大学大学院情報学研究科　博士（情報学）)

2021

Declaration

The Doctor of Informatics thesis entitled “Elucidation of the information processing underlying the behavior of pursuing the unprofitable targets” is the author’s own original work. I declare that this thesis has not been submitted, either in the same form or in a different form, to this or any other university for a degree.

SUGAWARA, Michiyo

Acknowledgement

I would first like to express my sincere gratitude to my supervisor, Dr. Kentaro Katahira, for his support in completing this thesis. I also would like to express my gratitude to Prof. Dr. Hideki Ohira and Dr. Hiroki Tanabe who provided valuable comments. I also wish to thank laboratory members, for their support and providing many useful discussions. I'm also deeply grateful to my family and friends for encouragements. Finally, great appreciation goes to all the people who have given me support.

Summary

In studies of the learning processes underlying human choice behaviors, it is common to assume that values are updated based on the outcomes from choices (i.e., reinforcement learning). Although reinforcement learning has been shown to account for many aspects of human learning processes well, humans sometimes take the behavior that appears irrational to observers. In real-life, an easy-to-understand example is the partner selection. Some people continue to pursue same targets, even though they have been rejected over and over again. Why do such pursuing behaviors occur? In this thesis, we aimed to explain the information processing behind the pursuing behavior using a computational modeling assumed external factors (choice outcome) and internal factors (choice *per se*).

A good understanding of model properties is required for the correct interpretation of the results obtained from the computational modeling of behaviors. Thus, in study1 (Sugawara & Katahira, 2019; Sugawara & Katahira, *under revision*), we examined the usefulness of a hybrid model that included both external (choice outcome) and internal factors (choice *per se*). First, by conducting simulations, we demonstrated that the hybrid model can identify the true underlying process. Second, using the hybrid model, we showed that empirical data collected from a web-based experiment was governed by preceding choices (i.e., "choice perseverance") rather than the asymmetric value updating based on previous outcomes (i.e., "asymmetric learning"). This result was also supported by a model-neutral analysis. Finally, we applied the hybrid model to two open datasets in which asymmetric learning was reported. As a result, the asymmetric learning was validated in one dataset but not in another. These findings support the usefulness of the hybrid model to identify the genuine process underlying choice behaviors.

In study 2 (Sugawara & Katahira, *submitted*), to answer why some people pursue the hard-to-get target, we investigated whether the pursuit of the hard-to-get target which seldom respond in a positive manner was emerged from choice perseverance and/or asymmetric learning by using the hybrid model. All subjects in a web-based experiment conducted an avatar choice task which mimicked the partner selection. In this task, we defined "difficult" and "easy" avatars by manipulating the outcome probability. As a result, we found that some subjects repeatedly selected a difficult avatar (Pursuit group). Based on simulation, we clarified that higher choice perseverance could account for the pursuit of difficult avatars. Then, the hybrid model indicated that the Pursuit group had significantly higher choice perseverance than the No-pursuit group in the web-based experimental data. Moreover, although the baseline attractiveness was comparable among all avatars used in the choice task, the attractiveness of the difficult avatar significantly

increased only in the Pursuit group. Taken together, we concluded that people with high choice perseverance pursue the hard-to-get target, subsequently making the target more attractive.

Contents

Chapter 1 General Introduction

1.1 Significance of computational modeling research

The study of decision-making is the science of choice. The central issue of decision-making focuses on how we think and decide our behaviors. Behaviors are outputs that result from the brain processing environmental inputs, and by using the cutting-edge neuroscientific technologies, we can record numerous neural activities from behaving animals. However, to interpret what these brain activities represent, the theory which explains how the brain causes the behavior is necessary. Assuming organisms as an information processing system, various investigations has been conducted to model the mechanism of information processing from a stimulus-response relationship. Through the interaction with surrounding environments, we always learn the value of action and predict the consequences of its action. Computational approach is one of research tools to quantify the information processing underlying our decision-making. Especially, cognitive computation models formulate multiple and complicated cognitive processes, and allows us to develop theories via simulations and experiments. By estimating model parameters from the time-series of behaviors, cognitive computation models allow us to predict future behaviors. Furthermore, the estimated parameters provide some important insights for the principle behind behaviors. Thus, studying animal behaviors including humans give significant insights to understand why we do so in a daily life (Niv, 2020). Because cognitive computation models are worthful for the neuroimaging and neuropharmacological studies attempting to unravel an invisible information processing in neural system, these models have been used in various fields such as neuroscience, behavioral economics, psychiatry, and psychology.

We always do a lot of choices in everyday life. In the expected utility theory (von Neumann & Morgenstern, 1944), rationality refers to behave in a way that maximizes the expected utilities. However, people often behave irrationally in daily life. For example, despite wanting to lose weight, you eat the cake in front of you. You decided to get up early to study before the exam, but you would fall asleep twice. You know the deadline is looming, but you do nothing until the last minute. Why do humans continue to behaves looking irrational from observers? Especially, I have a special interest in pursuing behaviors toward the unprofitable targets. For instance, stalkers pursue a specific person who does not respond in a positive manner, and scientists passionately pursue specific topics that have not been successful for a long time. Why do humans take pursuing behaviors? Such pursuing behaviors are usually problematic. In a clinical psychology, behavior modification is the most important process, outcome, and goal for

cognitive behavioral therapy (Eysenck, 1959, 1987). However, to modify the behavior, it should understand how the brain cause the behavior by using cognitive computation modeling.

1.2 Reinforcement learning as cognitive behavioral modeling

In early 20th century, Edward L. Thorndike mentioned that animals come to respond more for behavior with the subsequently positive outcomes in trial and error learning (Thorndike, 1911). This behavior principle has been called as "the law of effect". In the middle of 20th century that behaviorism was dominant in psychology, many behavioral theories were proposed. Even today, these theories underpin the study of human behaviors (Watoson, 1930; Skinner, 1938; Rescorla & Wagner, 1972). Especially, B. F. Skinner sophisticated the law of effect and defined "reinforcement" which is that the increase in the frequency of action according to a reward obtained as an outcome resulted from the action (Skinner, 1938). The reinforcement learning explains the mechanism of learning adequate behavior by reinforcement through trial and error. The reinforcement learning theory developed by Skinner is the fundamental behavioral theory underling behavioral and cognitive-behavioral therapies. Furthermore, the reinforcement learning is a worthful framework for controlling actions, and has been formulated as a reinforcement learning model in the fields of artificial intelligence and engineering (Sutton & Barto., 1998).

Reinforcement learning model assumes that decision-makers (hereafter agents) have a computational model that is formulated to process information entered from the surrounding environment and determine their action according to this computational model. To decide next action, an agent iteratively corrects own internal model based on the outcomes resulted from actions generated from the internal model (**Figure 1.1**). This updating principle is formulated based on the Rescorla-Wagner model (Rescorla & Wagner, 1972). Rescorla-Wagner model describes the change in the strength of association between a conditioned stimulus (CS) and an unconditioned stimulus (US) in time series data, and is represented by an updating equation:

$$V(t+1) = V(t) + \alpha\big(R(t) - V(t)\big). \tag{1.1}$$

Here, t is the current number of trials. V is the strength of association. R is the strength of US. In this equation, α is the parameter that determines how much to update the strength of association from experience. The strength of association ($V(t+1)$) represents the expectation (prediction) for the next US, and $R(t) - V(t)$ represents the error between the actual and predicted US. Q-learning model, which is a typical reinforcement learning model, represents the expected reward resulted

from an action a in specific state s as Q value, and updates this Q value according to Rescorla-Wagner model:

$$Q_{s,a}(t+1) = Q_{s,a}(t) + \alpha\left(R_{s,a}(t) - Q_{s,a}(t)\right). \quad (1.2)$$

As Equation 1.2 shows, the reinforcement learning model is assuming that agents learn action-outcome associations with outcomes resulted from chosen actions on a trial-and-error basis. The bandit task is gold standard to investigate this model in humans. In this thesis, I used two-armed bandit task that has only two options.

Reinforcement learning model must formulate the action selection where an agent chooses one action among available actions. In the two-armed bandit task, Q-learning model calculated the probability of choosing an option (P_{option}) by the softmax function

$$P_{\text{option }1}(t) = \frac{1}{1 + \exp\left(-\beta\left(Q_{option\,1}(t) - Q_{option\,2}(t)\right)\right)} \quad (1.3)$$

where $Q_{option\,1}$ is the value of choosing an option, and $Q_{option\,2}$ is the value of choosing another option. The inverse temperature (β) determines the sensitivity of the choice probabilities to the difference between the Q values. Thus, as Equation 1.3 represents, reinforcement learning model assumes that an agent chooses one action based on the difference of the expected reward resulted from available actions.

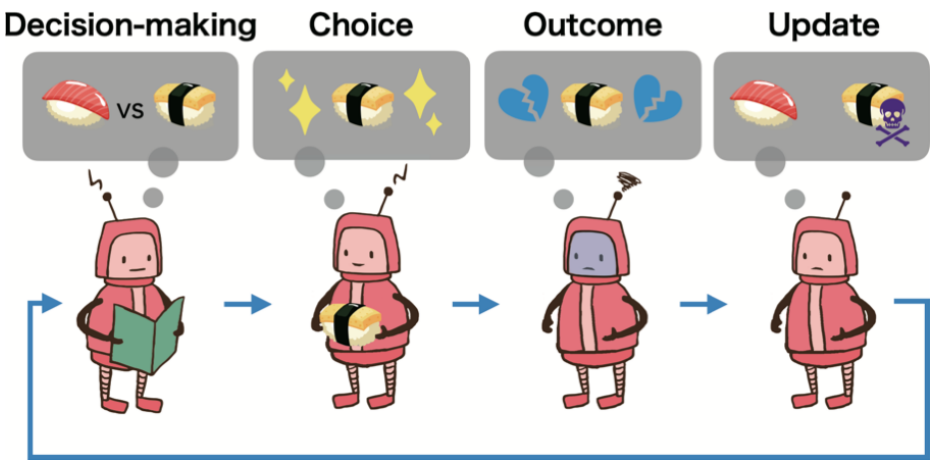


Figure 1.1 The concept of reinforcement learning

1.3 Reinforcement learning model for repetitive choice behaviors

By “the law of effect”, animals come to respond more for behavior with the subsequently positive outcomes in trial and error learning (Thorndike, 1911). The law of effect has been elaborated in the reinforcement learning framework, and has been widely used as the most versatile framework to explain the behavior of animals in various areas. However, the pursuing behaviors toward unprofitable target such as stalkers and scientists are deviated from the law of effect. How does cognitive computation modeling explain these pursuing behaviors? In this section, I introduce the extended reinforcement learning models explaining repetitive choice behaviors.

1.3.1 *Cognitive biases that influence decision-making*

One potential psychological process is a cognitive bias to explain repetitive choice behavior. Real-life decision making is subject to many cognitive biases (Erev et al., 2017). To explain cognitive biases influencing human choice behaviors in an everyday life, it is necessary to extend the reinforcement learning model. In the field of psychology, the cognitive biases such as the positive bias (Peeters, 1971; Sears, 1983) and the confirmation bias (Mynatt et al., 1977) that preferentially assign more attention to own desirable outcomes have been studied for a long time. The selective attention to desirable outcomes decreases the effect of undesirable outcomes on the future decision making, leading to repeat preceding choices.

To explain these cognitive biases, Palminteri, Lefebvre, et al. (2017) and Lefebvre et al. (2017) proposed an extended reinforcement learning model. As I mentioned in section 1.2, the learning rate is a model parameter that determines how much future reward prediction is updated from the current reward prediction error (RPE), which is the error between the actual rewards and the predicted rewards. Thus, the cognitive biases can be expressed by making the learning rate for the desired outcomes higher than the learning rate for the undesired outcomes (i.e., the asymmetric value updating). The extended model has two different learning rates for the positive RPE from the desired outcomes and the negative RPE from the undesired outcomes. The cognitive biases can be expressed by the asymmetry of these learning rates (i.e., the asymmetric value updating). Specifically, to express positivity and confirmation biases, the learning rate for positive RPE is greater than that for negative RPE.

1.3.2 *Choice perseverance*

Asymmetric value updating depends on the outcomes resulted from chosen outcomes. Can our behaviors be sufficiently explained only by action-outcome associations? Let you consider why you went to the school when you were a student in high school. Almost students do not expect pleasurable events every day. On the other hand, they always decide to go to school not because their parents punish them due to the absent from school. Usually, students are unaware of clear reasons why they go to school. In addition to the law of effect, (Thorndike, 1911) proposed the "law of exercise" that the preceding action is more likely to be selected later again. In the recent framework, the concept called "decision inertia" (Akaishi et al., 2014; Alós-Ferrer et al., 2016; Urai et al., 2017) and "choice perseverance" (Katahira, 2018; Sugawara & Katahira, *accepted*) captures the same behavioral phenomena referred as the law of exercise. As well as the positivity and confirmation biases, choice perseverance leads to repeat preceding actions. However, choice perseverance assumes that the intrinsic information from the current choice *per se* influences the subsequent choices, while positivity and confirmation biases depend on the extrinsic information from the desirable outcomes. In line with choice perseverance, students go to school because going to school per se today promotes to go to school tomorrow, rather than because students expect the desirable outcomes. If you go to school depending on choice perseverance, the absence from school makes it difficult to go to school. In the actual cases, the beginning of school refusal is often attributed to an absence due to a trivial cause such as a cold.

The choice perseverance in the reinforcement learning model is represented by the choice history independent of the outcome history (i.e., Q value):

$$C_a(t+1) = \begin{cases} C_a(t) + \tau\big(1 - C_a(t)\big) & \text{if } a \text{ is chosen} \\ C_a(t) + \tau\big(0 - C_a(t)\big) & \text{if } a \text{ is not chosen.} \end{cases} \quad (1.4)$$

As this equation represents, the choice history (C) is updated by the current choice. The decay rate (τ) determines the number of preceding choices influencing the current choice (Katahira, 2015, 2018). To reflect this choice history in the action selection, the softmax function indicated in Equation 1.3 is extended:

$$P_{\text{option }1}(t) = \frac{1}{1 + \exp(-\beta(Q_{option\,1}(t) - Q_{option\,2}(t)) - \varphi(C_{option\,1}(t) - C_{option\,2}(t)))}. \quad (1.5)$$

If the parameter of choice perseverance (φ) which determines how much the choice history influence subsequent choices is more positive, the agent more frequently repeats the same choices.

Therefore, while the positivity and confirmation biases maintain one's own beliefs by biasing the impact of past outcomes, the choice perseverance maintains one's own beliefs by taking over choice *per se*.

1.3.3 The main question and the end goal of the thesis

The purpose of this thesis is to elucidate the information processing underlying pursuing behaviors in humans by using cognitive computation modeling. While an outcome obtained from action is extrinsic information about the action, the action *per se* is also intrinsic information about the action. These two different information leads to repeat preceding choices. Previous studies explain the repetitive choice behaviors the model which formulate only one updating process from either extrinsic or intrinsic information. If the fitted model neglects the computational process which largely influences the choice behaviors, severe statistical bias could be produced. This statistical bias leads an erroneous interpretation about the estimated parameters, churning out incorrect scientific conclusions. To answer why people pursue unprofitable targets, it is necessary to assume that both extrinsic and intrinsic information influence our behaviors. Therefore, I firstly developed the hybrid model dissociate the extrinsic factors (i.e. the positive and confirmation bias) and the intrinsic factors (i.e. choice perseverance) in reinforcement learning paradigm. Then, by using the hybrid models, I elucidated the information processing underlying pursuing behaviors in humans by using cognitive computation modeling which can dissociate the effects of extrinsic and intrinsic information in the reinforcement learning paradigm.

1.4 Structure of this thesis

In **Chapter 2**, to reveal the cognitive processes underlying the pursuing behavior, I aimed to dissociate the extrinsic factors (i.e. the positive and confirmation bias) and the intrinsic factors (i.e. choice perseverance) in reinforcement learning paradigm. Specifically, this study adopted the hybrid model incorporating both asymmetric value updating and choice perseverance. First, I examined the usefulness of the hybrid model through the simulated data. Second, to reveal the genuine process underlying the empirical choice behavior in the reinforcement learning paradigm, I applied the hybrid model into the actual data collected by the web-based experiment. Finally, to re-evaluate the underlying cognitive processes in open datasets published in previous studies, I fitted the hybrid model and showed that the asymmetric value updating which was reported in the previous study could be accounted for by the choice perseverance. According to these examinations, I demonstrated the validity of the computational approach to distinguish between the effects of the choice outcomes and the choice *per se*.

In **Chapter 3**, to clarify the psychological factors behind the pursuing behavior toward the unprofitable target by using a hybrid model, I conducted web-experiment mimicking decision-making in a real-life. Based on the choice for the hard-to-get target that seldom responded in a positive manner, subjects were assigned two different groups: Pursuit and No-pursuit groups. Subjects in Pursuit group continuously chose the hard-to-get target, while subjects in No-pursuit group less chose that target. To clarify the cognitive process underlying the pursuing behavior toward the hard-to-get target, I applied the hybrid model established in Study 1 into the simulated and the actual choice data. Additionally, I also focused on psychological factors such as "preference" and "attractiveness" that are associated with choice behavior. Although it is generally known that choice is based on one's preference (Glimcher, 2009), it has been reported that choice *per se* makes the targets more attractive (Ariely & Norton, 2008; Brehm, 1956; Cockburn et al., 2014; Hornsby & Love, 2020; Izuma & Murayama, 2013; Koster et al., 2015; Nakao et al., 2016; Schonberg et al., 2014; Sharot et al., 2009). In a real-life, various outcomes are usually produced by our actions and influence preference as well as subsequent choices. Therefore, I examined how the cognitive process and attractiveness was linked with the pursuing behavior in the reinforcement learning paradigm.

Finally, **Chapter 4** will summarize the results of the studies in this dissertation and speculate on future directions.

Chapter 2 Dissociation between asymmetric value updating and perseverance in human reinforcement learning

2.1 Introduction

Repetitive choices are induced by either extrinsic (i.e., outcome) or intrinsic information (i.e., choice *per se*) resulted from own choice. To investigate the computational process underlying the pursuing behaviors, it is necessary to develop the computational model which can dissociate the impacts of choice *per se* and outcome. Thus, in this chapter, I extended the reinforcement learning (RL) model to formulate the impact of both extrinsic and intrinsic information, and demonstrate the usefulness of this extended model for the dissociation between the effect of extrinsic and intrinsic information on human choice behaviors.

As I mentioned in Chapter 1, RL models have been broadly used to model the choice behavior of humans and other animals (Daw et al., 2011; Redish & Johnson, 2008). Standard RL models suppose that agents learn action-outcome associations from outcomes on a trial-and-error basis (Barto, 1997). The learned action values are assumed to be updated according to the reward prediction error, which is the difference between the actual and expected rewards (Rescorla & Wagner, 1972; Sutton & Barto., 1998).

Although this mechanism is often assumed to underlie many background processes of human behavior, human decision making is subject to many biases (Erev et al., 2017). Several modeling studies investigating human choice behavior have reported that the magnitude of the value update is biased depending on the sign of the reward prediction error. This bias can be represented in RL models as asymmetric learning rates for positive and negative outcomes (Frank et al., 2007; Gershman, 2015; Niv et al., 2012).

Lefebvre et al. (2017) suggested that this learning asymmetry reflects positivity bias (the tendency to emphasize good outcomes) in factual learning in which feedback is given only for the option chosen by the subject. Refining this idea, Palminteri, Lefebvre, et al. (2017) reported that this learning asymmetry represents confirmation bias (the tendency to selectively process information that supports one's beliefs) in counterfactual learning in which feedback is given for both the chosen and unchosen options (Boorman et al., 2011; Fischer & Ullsperger, 2013). These learning asymmetries lead to choice repetition because the influences of the outcomes that

reinforce the choice (positive outcome for the chosen option and negative outcome for unchosen option) are enhanced, whereas those that weaken the choice are diminished (Katahira, 2018).

It has also been shown that our decisions depend on our choice history regardless of the choice outcome (Bertelson, 1965; Gold et al., 2008; Nakao et al., 2016; Schönberg et al., 2007). A positive dependency leads to the repetition of the same choices (hereafter, "perseverance"). Perseverance leads to behavior seemingly similar to that resulting from asymmetric learning rates. Katahira (2018) suggested the possibility that the estimation of asymmetric learning rates suffers from statistical artifacts caused by model misspecification. If an RL model without the perseverance factor is fitted to data that possess intrinsic autocorrelation (e.g., perseverance), the model tends to represent perseveration by asymmetric learning rates. Thus, a statistical bias that overestimates the difference in learning rates occurs. Due to this statistical bias, it is difficult to identify the cognitive process underlying human choice behavior. Nevertheless, the identification of computational processes, such as asymmetric value updating and perseverance, is crucial for interpreting neural mechanisms and investigating the association with personality traits in the fields of neuroscience, psychology, and psychiatry (Akaishi et al., 2014; Alós-Ferrer et al., 2016; Frank et al., 2007; Gershman et al., 2009; Huys et al., 2011; Kuzmanovic & Rigoux, 2017; Lefebvre et al., 2017; Niv et al., 2012).

The present study proposes methods to dissociate these computational processes from empirical behavioral data. Specifically, I address this issue by using a hybrid model (hereafter Hybrid model) incorporating asymmetric learning rates and perseverance. In the present study, I first conduct simulations to investigate how the Hybrid model works to identify the true underlying processes under various conditions. Then, I demonstrate how the Hybrid model can identify the underlying process in an empirical dataset with a relatively large sample size. Finally, to clarify the genuine process underlying open datasets collected from previous studies reporting asymmetric updating, I apply the Hybrid model to these datasets. According to a series of investigations, I conclude that the Hybrid model combining outcome-based and outcome-independent processes enables the detection of the genuine cognitive process underlying choice behavior while avoiding statistical artifacts.

2.2 Methods

2.2.1 Behavioral tasks

In this study, I used the same behavioral task in both the simulations and web-based experiment. The task was a modified version of the probabilistic instrumental learning task developed in previous studies (Lefebvre et al., 2017; Palminteri et al., 2015; Palminteri, Lefebvre, et al., 2017). The framework used in this task is generally called a two-armed bandit problem in which an agent (subject) sequentially explores the best choice among multiple options (Sutton & Barto., 1998). This task consisted of a factual block and a counterfactual block (**Figure 2.1a**). In the web-based experiment, half of the subjects started with the factual block, and the other half started with the counterfactual block. In each block, the agent experienced two sessions separated by a 20-s break. In each session, I selected eight abstract stimuli (Agathodaimon font) and generated four different pairs. In the second session, all stimuli were renewed such that the agent began learning anew (i.e., the option stimuli differed between the two sessions). The display positions of the stimuli were set to appear on the left and right in the same number of trials. These four stimulus pairs were distributed among the following three conditions: same (1 pair), different (2 pairs), and reversal (1 pair). Under the same condition, both stimuli were associated with a 50% reward probability (here, the reward was "+10 pt"). Under the different condition, one stimulus was associated with a 25% reward probability, and the other stimulus was associated with a 75% reward probability. Under the reversal condition, one stimulus was associated with a 17% reward probability, and the other stimulus was associated with an 83% reward probability during the first 12 trials, and then, these contingencies were reversed during the final 12 trials (**Figure 2.1b**). Each pair was presented in 24 trials per session. Thus, each session included 96 trials. The order of the trials was pseudo-randomized with the constraint that the same condition continued for no more than four times in a sequence. The agents were not given any explicit information regarding the reward probabilities. The agents were instructed to earn as many points as possible across experiments by trial and error.

The agents completed an 8-trial practice session before each block (factual or counterfactual learning) after the overall task description was provided. The stimuli used in the practice trials were not used in the main task. At the initiation of each trial, a fixation crosshair appeared for 500 ms. Following the fixation crosshair, one of four stimulus pairs was displayed for 2000 ms during which the agent had to choose one of the two stimuli by pressing either "F" (left option) or "J" (right option) on their keyboard. If the agent chose one option within 1500 ms, a red triangle was placed below the chosen option until the outcome presentation. If the agent did

not choose any option within 1500 ms, a warning message was displayed for 1500 ms, and the trial was considered missed ("-10 pt"). Then, the outcomes were displayed for 1500 ms ("+10 pt" or "-10 pt"). In the factual learning context, the agents were only shown the outcome of the chosen option. In the counterfactual learning context, the agents were shown the outcomes of both the chosen and unchosen options. Since this research involved subjects in a web-based experiment, some tasks reported in Palminteri, Lefebvre, et al. (2017) were modified. The main modifications were the inclusion of a time limit for the response and the use of a fixed duration for the feedback presentation. In previous experiments, the subjects responded and observed feedback at their own pace. These modifications aimed to control the entire duration of the experiment.

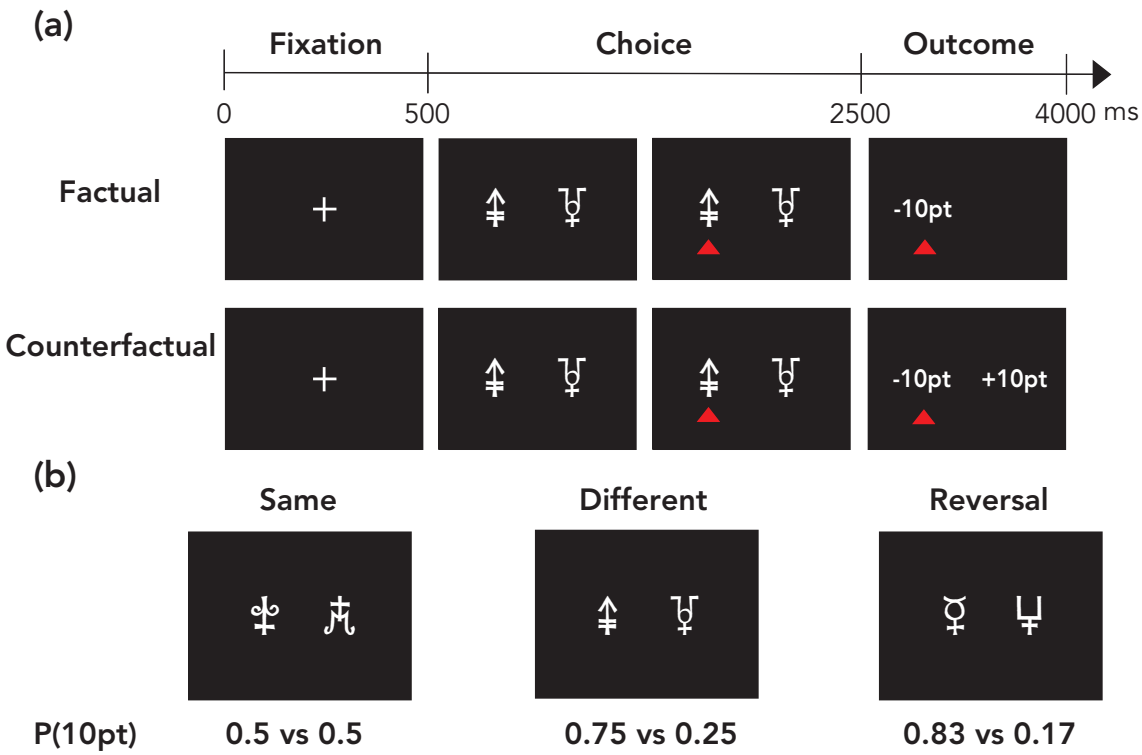


Figure 2.1 Experimental task. (a) There were two types of learning contexts in the present study. In the factual learning context, the subjects were shown only the outcome of the chosen option. In the counterfactual learning context, the subjects were shown the outcomes of both the chosen and unchosen options. (b) Task conditions: Under the same condition, the option pair had an identical reward contingency. Under the different condition, one option had a higher reward probability than the other option. Under the reversal condition, the reward probability was reversed between the options after the first 12 trials were completed.

2.2.2 Models

In this study, I mainly used three types of reinforcement learning models. All models were modifications of a typical Q-learning model. (1) The Asymmetry model has two independent learning rates, i.e., α_c^+ and α_c^- ,for positive and negative reward prediction errors (RPEs), respectively, to represent asymmetric value updating. (2) The Perseverance model includes the computational process of choice history independent of the outcome-based learning process. The computational process of choice history has the following two free parameters: decay rate (τ) and perseverance parameter (φ). (3) The Hybrid model has the features of both the Asymmetry and Perseverance models. In the counterfactual learning context, the models consider the impact of the forgone outcomes of the unchosen option and the impact of the obtained outcome of the chosen option. The details of the models are described below.

For the data from the factual learning task, I used the following six RL models: standard RL, Asymmetry, Perseverance (impulsive), Perseverance (gradual), Hybrid (impulsive), and Hybrid (gradual) models. The standard RL model is the most basic of all models considered in the present study. In the standard RL model, the action value of the chosen option in trial t, which is denoted by $Q_c(t)$, is updated according to the following equation:

$$Q_c(t+1) = Q_c(t) + \alpha\big(R_c(t) - Q_c(t)\big). \tag{2.1}$$

Here, the outcome (of the chosen option) in trial t is denoted by $R_c(t)$. $(R_c(t) - Q_c(t))$ represents the prediction error, which is subsequently denoted by δ_c. The learning rate (α) determines how much the model updates the action value with the prediction error. The initial action value of each option is set to zero. For the data from the factual learning task, only the Q value of the chosen option is updated because the agents are informed only of the outcome of the chosen option. Choice probability $P_c(t)$ is determined by the following softmax function:

$$P_c(t) = \frac{1}{1 + \exp\Big(-\beta\big(Q_c(t) - Q_u(t)\big)\Big)} \tag{2.2}$$

where Q_c is the Q value of the chosen option, and Q_u is the value of the unchosen option. The inverse temperature (β) determines the sensitivity of the choice probabilities to the difference between the Q values.

The Asymmetry model is extended from the standard RL model to allow the learning

rates to differ (α_c^+ , α_c^-) depending on the sign of the prediction error. Thus, the Q values are updated as follows:

$$Q_c(t+1) = \begin{cases} Q_c(t) + \alpha_c^+ \, \delta_c(t) & \text{if } \delta_c(t) \geq 0 \\ Q_c(t) + \alpha_c^- \, \delta_c(t) & \text{if } \delta_c(t) < 0. \end{cases} \tag{2.3}$$

Previous studies have shown that this model can be used to express positivity bias or confirmation bias (Lefebvre et al., 2017; Palminteri, Lefebvre, et al., 2017).

The Perseverance model uses the same update rule as the standard RL model (Equation 2.3). In the models that incorporate the perseverance factor, the choice trace $C(t)$ is defined to introduce the effect of a past choice to the choice probability (Gershman et al., 2009; Huys et al., 2011):

$$P_c(t) = \frac{1}{1 + \exp(-\beta(Q_c(t) - Q_u(t)) - \varphi(C_c(t) - C_u(t)))}. \tag{2.4}$$

The perseverance parameter (φ) is a parameter that controls for the tendency to repeat the choice of or avoid a recently chosen option. A high positive value of this parameter indicates that the agent frequently repeats the previous choice. The choice trace is computed using the following update rule (Akaishi et al., 2014; Katahira, 2018):

$$\begin{aligned} C_c(t+1) &= C_c(t) + \tau\big(1 - C_c(t)\big) \\ C_u(t+1) &= C_u(t) + \tau\big(0 - C_u(t)\big) \end{aligned} \tag{2.5}$$

where C_c and C_u denote the choice trace of the chosen and unchosen options, respectively. The decay rate determines the number of preceding choices in the choice history influencing the current choice(Katahira, 2015, 2018). In the Perseverance (impulsive) model, with the decay rate fixed at 1, only the immediately preceding choice influences the current choice. Most previous studies examining choice perseverance have incorporated the influence of only the immediate prior trial (Gillan et al., 2016; Huys et al., 2011). However, Katahira (2018) showed that the long-term choice history caused bias in the estimation of the asymmetric learning rates.

The Hybrid model is a model combining the Asymmetry and Perseverance models. This model incorporates not only the asymmetric learning rates but also the choice trace.

For the data from the counterfactual learning task, I used the following six RL models as described in the factual learning task: the standard RL, Asymmetry, Perseverance (impulsive), Perseverance (gradual), Hybrid (impulsive), and Hybrid (gradual) models. Here, all models are allowed to update the Q values of both the chosen and unchosen options because the agent was informed of both outcomes. The standard RL, Perseverance (gradual), and Perseverance (impulsive) models have the same parameters as the models used in the factual learning task because an identical learning rate is used to update the values of both the chosen and unchosen options regardless of the sign of the prediction error.

In the Asymmetry models, four different learning rates are defined to represent the asymmetric updating of the chosen (α_c^+ , α_c^-) and unchosen (α_u^+ , α_u^-) options. The Q values of the chosen and unchosen options are computed as follows:

For the chosen option

$$Q_c(t+1) = \begin{cases} Q_c(t) + \alpha_c^+ \, \delta_c(t) & \text{if } \delta_c(t) \geq 0 \\ Q_c(t) + \alpha_c^- \, \delta_c(t) & \text{if } \delta_c(t) < 0. \end{cases} \tag{2.3}$$

For the unchosen option:

$$Q_u(t+1) = \begin{cases} Q_u(t) + \alpha_u^+ \, \delta_u(t) & \text{if } \delta_u(t) \geq 0 \\ Q_u(t) + \alpha_u^- \, \delta_u(t) & \text{if } \delta_u(t) < 0 \end{cases} \tag{2.6}$$

where δ_u denotes the prediction error of the unchosen option.

In the counterfactual learning context, I also used the Hybrid (gradual) and Hybrid (impulsive) models, which are hybrid models combining the Asymmetry model and the Perseverance model, to examine the asymmetry of the learning rate while incorporating choice perseverance.

2.2.3 Parameter estimation and model comparison

Using the R function "solnp" in the Rsolnp package (Ghalanos & Theussl, 2015), I fit the parameters of each model with the maximum a posteriori (MAP) estimation and calculated the log marginal likelihood of each model using Laplace approximation (Kass & Raftery, 1995). In contrast to a likelihood, a marginal likelihood penalizes a complex model with extra parameters in the marginalization process. Because the marginal likelihood is proportional to the posterior

probability of the model, the model resulting in the highest marginal likelihood is the most likely one given a particular data set. Notably, this situation is only true if all models have an equal prior probability (i.e., all models are equally likely before the data are provided). This method incorporates the prior distributions of the parameters and can avoid extreme values in the estimates of the parameters (Daw, 2011; Katahira, 2016). The prior distributions and constraints were set following Palminteri, Lefebvre, et al. (2017). All learning rates were constrained to the range of $0 \leq \alpha \leq 1$ with a *Beta* (1.1, 1.1) prior distribution. The inverse temperature was constrained to the range of $\beta \geq 0$ with a *Gamma* (shape = 1.2, scale = 5.0) distribution. In the perseverance model, the decay rate was constrained to the range of $0 \leq \tau \leq 1$ with a *Beta* (1, 1) distribution (i.e., a uniform distribution), and the perseverance parameter was constrained to the range of $-10 \leq \varphi \leq 10$ with a *Norm* ($\mu = 0$, $\sigma^2 = 5$) distribution.

2.2.4 Simulations

To understand how the Hybrid model works, I conducted simulations that directly evaluated the amount of bias in the parameter estimates of the misspecified models. In the simulations, I first generated the choice data under the five simulated conditions (true models; **Table 2.1**) used to perform the probabilistic instrumental learning task (see the 'Behavioral tasks' section) and then fitted three models (the Asymmetry, Perseverance (gradual), and Hybrid (gradual) models) to the data.

In the factual learning context, the simulated conditions from the versions of the three models were set as follows: (i) a model with asymmetric learning rates assuming positivity bias ($\alpha_c^+ = 0.5,\ \alpha_c^- = 0.2, \beta = 0.3,\ \tau = 0.4,\ \varphi = 0$); (ii) a model with an asymmetric learning rate assuming negativity bias ($\alpha_c^+ = 0.2,\ \alpha_c^- = 0.5, \beta = 0.3,\ \tau = 0.4,\ \varphi = 0$); (iii) a model with a symmetric learning rate and perseverance ($\alpha_c^+ = \alpha_c^- = 0.5, \beta = 0.3,\ \tau = 0.4,\ \varphi = 1.5$); (iv) a model with an asymmetric learning rate and perseveration assuming positivity bias ($\alpha_c^+ = 0.5$, $\alpha_c^- = 0.2, \beta = 0.3,\ \tau = 0.4,\ \varphi = 1.5$); and (v) a model with an asymmetric learning rate and perseveration assuming negativity bias ($\alpha_c^+ = 0.2,\ \alpha_c^- = 0.5, \beta = 0.3,\ \tau = 0.4,\ \varphi = 1.5$). In the counterfactual learning context, the simulated conditions from the versions of the three models were set as follows: (i) a model with an asymmetric learning rate assuming confirmation bias ($\alpha_c^+ = 0.5,\ \alpha_c^- = 0.2,\ \alpha_u^+ = 0.2,\ \alpha_u^- = 0.5,\ \beta = 0.3,\ \tau = 0.4,\ \varphi = 0$); (ii) a model with an asymmetric learning rate assuming opposite confirmation bias ($\alpha_c^+ = 0.2,\ \alpha_c^- = 0.5,\ \alpha_u^+ = 0.5$, $\alpha_u^- = 0.2,\ \beta = 0.3,\ \tau = 0.4,\ \varphi = 0$); (iii) a model with a symmetric learning rate and perseverance ($\alpha_c^+ = \alpha_c^- = \alpha_u^+ = \alpha_u^- = 0.5,\ \beta = 0.3,\ \tau = 0.4,\ \varphi = 1.5$); (iv) a model with an asymmetric learning rate and perseveration assuming confirmation bias ($\alpha_c^+ = 0.5,\ \alpha_c^- = 0.2,$

$\alpha_u^+ = 0.2$, $\alpha_u^- = 0.5$, $\beta = 0.3$, $\tau = 0.4$, $\varphi = 1.5$); and (v) a model with an asymmetric learning rate and perseveration assuming opposite confirmation bias ($\alpha_c^+ = 0.2$, $\alpha_c^- = 0.5$, $\alpha_u^+ = 0.5$, $\alpha_u^- = 0.2$, $\beta = 0.3$, $\tau = 0.4$, $\varphi = 1.5$). All parameters were set according to the parameters of the empirical dataset obtained from the web-based experiment. The number of trials was set as 960 trials per session per block. Under each simulation condition, 100 virtual datasets were simulated.

Table 2.1 List of models and model selection results of the simulation data

Simulation condition (True model)	Learning context	Fit model	Learning rate (s)	Inverse temperature	Perseverance	# of free parameters	P(Fit model |True model)
Asymmetry with Confirmation bias ($\alpha_{c+} > \alpha_{c-}, \alpha_{u+} < \alpha_{u-}$)	factual	Asymmetry	α_c^+, α_c^-	β	–	3	**0.97**
		Perseverance	α	β	τ, φ	4	0.00
		Hybrid	α_c^+, α_c^-	β	τ, φ	5	0.03
	counterfactual	Asymmetry	$\alpha_c^+, \alpha_c^-, \alpha_u^+, \alpha_u^-$	β	–	5	**0.90**
		Perseverance	α	β	τ, φ	4	0.08
		Hybrid	$\alpha_c^+, \alpha_c^-, \alpha_u^+, \alpha_u^-$	β	τ, φ	7	0.02
Asymmetry with Opposite confirmation bias ($\alpha_{c+} < \alpha_{c-}, \alpha_{u+} > \alpha_{u-}$)	factual	Asymmetry	α_c^+, α_c^-	β	–	3	**0.96**
		Perseverance	α	β	τ, φ	4	0.00
		Hybrid	α_c^+, α_c^-	β	τ, φ	5	0.04
	counterfactual	Asymmetry	$\alpha_c^+, \alpha_c^-, \alpha_u^+, \alpha_u^-$	β	–	5	**0.98**
		Perseverance	α	β	τ, φ	4	0.00
		Hybrid	$\alpha_c^+, \alpha_c^-, \alpha_u^+, \alpha_u^-$	β	τ, φ	7	0.02
Perseverance and Symmetric learning rate ($\alpha_{c+} = \alpha_{c-} = \alpha_{u+} = \alpha_{u-}$)	factual	Asymmetry	α_c^+, α_c^-	β	–	3	0.00
		Perseverance	α	β	τ, φ	4	**0.89**
		Hybrid	α_c^+, α_c^-	β	τ, φ	5	0.11
	counterfactual	Asymmetry	$\alpha_c^+, \alpha_c^-, \alpha_u^+, \alpha_u^-$	β	–	5	0.02
		Perseverance	α	β	τ, φ	4	**0.97**
		Hybrid	$\alpha_c^+, \alpha_c^-, \alpha_u^+, \alpha_u^-$	β	τ, φ	7	0.01
Hybrid with Confirmation bias ($\alpha_{c+} > \alpha_{c-}, \alpha_{u+} < \alpha_{u-}$)	factual	Asymmetry	α_c^+, α_c^-	β	–	3	0.00
		Perseverance	α	β	τ, φ	4	0.02
		Hybrid	α_c^+, α_c^-	β	τ, φ	5	**0.98**
	counterfactual	Asymmetry	$\alpha_c^+, \alpha_c^-, \alpha_u^+, \alpha_u^-$	β	–	5	**0.40**
		Perseverance	α	β	τ, φ	4	0.35
		Hybrid	$\alpha_c^+, \alpha_c^-, \alpha_u^+, \alpha_u^-$	β	τ, φ	7	0.25
Hybrid with Opposite confirmation bias ($\alpha_{c+} < \alpha_{c-}, \alpha_{u+} > \alpha_{u-}$)	factual	Asymmetry	α_c^+, α_c^-	β	–	3	0.00
		Perseverance	α	β	τ, φ	4	0.02
		Hybrid	α_c^+, α_c^-	β	τ, φ	5	**0.98**
	counterfactual	Asymmetry	$\alpha_c^+, \alpha_c^-, \alpha_u^+, \alpha_u^-$	β	–	5	0.25
		Perseverance	α	β	τ, φ	4	0.00
		Hybrid	$\alpha_c^+, \alpha_c^-, \alpha_u^+, \alpha_u^-$	β	τ, φ	7	**0.75**

2.2.5 Web-based experiment

Experimental procedures

One hundred and fifty adults participated in the web-based experiment via CrowdWorks (https://crowdworks.jp/). I limited the subjects' age to over 18 years and paid approximately 700 yen (approximately $6) if the subjects completed all tasks and surveys without any interruption. Informed consent was obtained from all subjects by clicking 'I Agree' after reading the

information regarding the aim and procedures of this study. After the subjects provided their basic demographic information, including gender and age, and downloaded Inquisit player (Millisecond Software LLC, Seattle, USA), the subjects started the behavioral task (see the 'Behavioral tasks' section). The subjects were anonymized, and their privacy was protected. The study was approved by the Ethical Research Committee of Nagoya University, and the study was carried out in accordance with the relevant guidelines and regulations.

Seven subjects were excluded due to inappropriate task execution. Six of these subjects showed a false start rate greater than 30%. Thus, these subjects pressed any button before the choice options were presented in more than 30% of the trials. The other subject chose only the option that appeared on the right side across the experiments (even though each option randomly appeared on both sides). Thus, data from 143 subjects (58 females and 85 males) aged between 19 and 72 years (mean ± SD = 38.7 ± 9.6) were included in the subsequent analyses.

Performance evaluation

To evaluate whether the subjects successfully performed the behavioral tasks, I calculated the preferred response rate under the same condition and the correct rate under the different and reversal conditions as described in a previous study (Palminteri, Lefebvre, et al., 2017). The preferred response rate was calculated as the fraction of "preferred response," which was defined as the most frequently chosen option (i.e., the option chosen by the subject in more than 50% of the trials). The correct rate was defined as the fraction of trials in which the subjects chose the option associated with a higher reward probability.

I divided each session (24 trials) into four phases (6 trials per phase) to investigate the learning-related performance changes under each condition. Because the subjects completed two sessions in each learning context, I pooled the trials in each phase of the two sessions (i.e., 12 trials per pooled phase). Then, the correct rate under the different and reversal conditions and the preferred response rate under the same condition were calculated in each pooled phase.

Parameter correlation and parameter recovery

To validate our model-fitting results in the web-based experiment, I checked the correlations between the free parameters in each learning context and the capacity of recovering the model parameters using simulated data (Palminteri, Wyart, et al., 2017; Wilson & Collins, 2019). For the parameter recovery, I simulated the choice dataset under each condition of our

behavioral paradigm with model parameters corresponding to those estimated from our actual subjects (N = 143). The number of trials was set as 960 trials per session per block. These simulations were conducted using the model parameters estimated using the Asymmetry, Perseverance (gradual), and Hybrid (gradual) models. Thus, 143 virtual datasets were simulated per context and model. I fitted the same model used in the simulation to the simulated datasets. Then, the correlation coefficients between the true parameters used in the simulation and the estimated parameters in each context and model were calculated. Additionally, to determine the precision of the parameter recovery, I calculated the root mean squared error between the true value used to generate the data and the estimated value.

Model-neutral analysis of the web-based experiment

To assess the asymmetric value updating process underlying the empirical choice data collected in the web-based experiment, I conducted a model-neutral analysis as proposed in Katahira (2018). Because a detailed explanation was provided in the previous study (Katahira, 2018), here, I only briefly introduce the concept of this analysis. As reported in previous studies (Katahira, 2015, 2018), one behavioral consequence of asymmetric learning rates is that the impact of past outcomes depends on subsequent outcomes. Thus, an interaction likely exists between the outcome of one trial ago and the outcome of two trials ago that serves as a factor influencing the current choice. In this case, since the regression model assumes that the outcomes of one and two trials ago affect the current choice independently of each other, this point can be used to test the existence of learning rate asymmetry in the data. Thus, the regression coefficient of the interaction term is 0 if the true learning rate is symmetric ($\alpha^+ = \alpha^-$). Here, I only consider up to two trials ago and consider the condition under which the outcome is a binary value (1 = reward, 0 = no reward). Under this condition, the interaction term represents the degree of the asymmetric learning rates. A negative interaction indicates positive asymmetric updating ($\alpha^+ > \alpha^-$), while a positive interaction indicates negative asymmetric updating ($\alpha^+ < \alpha^-$).

To create the data used in this analysis, I first sorted the trial sequences of each subject, each four-stimulus pair, and each session. From each sequence, I extracted all possible three successive trials (I refer to these as a 'triplet'), denoted by $(t+1)$-th trial, t-th trial, and $(t-1)$-th trial. Then, from the resulting triplets, I further selected only the triplets in which the same option was selected at both the t-th trial and $(t-1)$-th trial. Finally, the triplets containing time out (no response) trials were excluded, and the remaining triplets were included in the analysis. Then, I constructed a logistic regression model to predict the probability that a subject chooses the same option in the $(t+1)$-th trial as in the t-th trial, denoted by $p(stay(t+1))$. In the

factual learning condition, the regressors are the outcomes of the choice at the t-th trial ($Rc(t)$) and $(t-1)$-th trial ($Rc(t-1)$) and their interaction term $Rc(t) \times Rc(t-1)$. The regressors are coded as $Rc(t) = 1$: rewarded and $Rc(t) = 0$: unrewarded. The logistic regression model is as follows:

$$\log\frac{p\big(stay(t+1)\big)}{p\big(switch(t+1)\big)} = b_0 + b_1 Rc(t) + b_2 Rc(t-1) + b_{12} Rc(t)Rc(t-1) \qquad (2.7)$$

where $p(switch(t+1))$ is the probability that the subject switches the choice at the $(t+1)$-th trial. The intercept b_0 represents the overall tendency to repeat the same choice, which may absorb the effects of the choices and rewards at trials before the $(t-1)$-th trial. In the counterfactual learning context, the logistic regression model also includes the regressors of the outcomes of the unchosen options at the t-th trial ($Ru(t)$) and the $(t-1)$-th trial ($Ru(t-1)$) and their interaction ($Ru(t) \times Ru(t-1)$). Thus, the logistic regression model is expressed as follows:

$$\log\frac{p(stay(t+1))}{p(switch(t+1))} = b_0 + b_1 Rc(t) + b_2 Rc(t-1) + b_3 Ru(t) + b_4 Ru(t-1)$$

$$+ b_{12} Rc(t)Rc(t-1) + b_{34} Ru(t)Ru(t-1) \qquad (2.8)$$

The hypothesis test based on the null hypothesis $b_{12} = 0$ and/or $b_{34} = 0$ was conducted by using mixed-effects models ("glmer" function) implemented with the lme4 package (Bates et al., 2019) in the R programming language. Within-subject factors (intercept, main effects, and their interaction) were included as random effects, i.e., allowed to vary across subjects.

2.2.6 Additional open data analysis

To clarify the genuine process underlying the empirical choice data collected in previous studies reporting asymmetric updating, I also applied the Hybrid model to two open datasets. Dataset 1 comprised the open data reported by Palminteri, Lefebvre, et al. (2017). Since our research was carried out according to this previous study, the experimental procedures were mostly the same. Although the authors of the previous study (Palminteri, Lefebvre, et al., 2017) analyzed the influence of choice perseverance, they did not examine the influence of the gradual perseverance factor ($\tau < 1$). Thus, using these open data, tests were performed by comparing the models, including those incorporating the gradual perseverance factor. Furthermore, I used

Dataset 2 reported by Niv et al. (2012) to investigate whether the asymmetric learning rates observed in another learning task could be explained by choice perseverance. I also applied the Hybrid model to these previous data and compared the model fitting and learning rate parameters. More detailed information regarding each dataset is as follows.

Dataset1 (Palminteri et al., 2017; https://figshare.com/authors/ /2803402)

In Palminteri, Lefebvre, et al. (2017), the asymmetric learning rates were examined in both factual and counterfactual learning contexts. As mentioned above, our web-based study was carried out using largely the same procedures as those used in this previous study. However, in the previous task, the subjects responded and observed feedback at their own pace. Furthermore, the previous study employed a between-subjects design in which each subject performed the task in either a factual ($N = 20$) or counterfactual ($N = 20$) learning context.

Dataset2 (Niv et al., 2012: http://www.princeton.edu/~nivlab/data/NivEtAl2012JNeuro/)

In Niv et al. (2012), Asymmetry models were used to explain risk-seeking/aversion behaviors in a factual learning context. A negative outcome learning rate higher than a positive outcome learning rate leads to risk aversion, whereas the opposite pattern leads to risk seeking. Their task included the following six option pairs that differed in risk and expected rewards: 20¢ (100%) versus 0 (50%) /40¢ (50%), 40¢ (100%) versus 0 (50%) /40¢ (50%), 20¢ (100%) versus 40¢ (100%), 0¢ (100%) versus 0 (50%) /40¢ (50%), 0¢ (100%) versus 20¢ (100%), and 0¢ (100%) versus 0¢ (100%). The experiment involved two types of trials. In the 'choice trials,' the subjects were required to choose between two stimuli, whereas in the 'forced trials,' the subjects were presented only one of five stimuli and had to choose the presented stimulus ($N = 16$). Similar to the analyses of the web-based data, I compared the estimated parameters among the Asymmetry, Hybrid (impulsive), and Hybrid (gradual) models.

2.2.7 Statistical tests

For the model comparison, one-way repeated-measures analysis of variance (rmANOVA) was conducted to compare the log marginal likelihoods of the models in each learning context (factual and counterfactual learning). I also investigated the difference in learning rates (α_c^+, α_c^-) within each model. For the Asymmetry and Hybrid models in the factual learning context, the difference in the learning rates was compared by a paired t-test. For the Asymmetry and Hybrid models in the counterfactual learning context, two-way rmANOVAs with Valence

(positive or negative) and Choice (chosen or unchosen) were performed to test for differences in the four learning rates (α_c^+, α_c^-, α_u^+, and α_u^-). The degree of biases in the learning rates were compared across the models by using a one-way rmANOVA in each learning context. Additionally, the degree of the perseverance parameter (φ) was compared across the models using a one-way rmANOVA in each learning context. To correct for the violation of the sphericity assumption, Greenhouse-Geiser's adjustment of the degrees of freedom was used in all rmANOVAs when appropriate. The post hoc pairwise comparisons were performed based on Shaffer's correction for multiple comparisons. For the simulation, the differences between the true and estimated parameters were evaluated by using a one-sample *t*-test with the true parameters. To control for the multiple comparison issue, the significance of the one-sample *t*-tests was tested with Bonferroni correction. In the parameter correlation analysis, I estimated the Pearson's correlation coefficients between the model parameters of the Perseverance (gradual) and Hybrid (gradual) models in the factual and counterfactual learning contexts. Additionally, in the parameter recovery, I estimated the Pearson's correlation coefficients between the model parameters estimated from the empirical dataset and the simulated dataset. The significance of the correlation coefficients was tested with Bonferroni correction to avoid multiple comparison issues. These analyses were executed using R version 3.5.1 statistical software (http://cran.us.r-project.org).

2.3 Results

2.3.1 Model identifiability and the usefulness of the Hybrid model

By conducting simulations, I investigated the identifiability of the three models (i.e., Asymmetry, Perseverance, and Hybrid) in each learning context, whether pseudo-asymmetric learning rates and pseudo-perseverance occurred by fitting mismatched models, and whether the Hybrid model could distinguish asymmetric value updating from choice perseveration.

To determine the identifiability of the models, I applied the three models to simulated data from the following versions of the three models: Asymmetry model assuming positivity/confirmation bias, Asymmetry model assuming negativity/opposite confirmation bias, Perseverance model, Hybrid model assuming positivity/confirmation bias, and Hybrid model assuming negativity/opposite confirmation bias. Then, I compared these models using log marginal likelihood (LML). Except for the simulated data from the Hybrid model assuming confirmation bias in the counterfactual context (rmANOVA, $F(1.94,192.33) = .39$, $p = .67$), the true model was deemed the best model (rmANOVA, $Fs >= 143.27$, $ps < .001$; **Table 2.1**).

Katahira (2018) demonstrated that by fitting the Asymmetry model to simulated data generated from the Perseverance model, the pseudo-asymmetry of the learning rates was observed. However, whether pseudo-perseverance might appear when the Perseverance model is fitted to the simulated data generated from the true Asymmetric model remains unclear. To examine this question, I fitted the Perseverance model to the simulated data from the Asymmetry model assuming positivity/confirmation bias and the Asymmetry model assuming negativity/opposite confirmation bias. In both cases, a higher perseverance parameter was observed despite the lack of perseverance ($\varphi = 0$) in the true model in the factual (**Figure 2.2d**, one-sample *t*-test, $t(99) = 12.60$, $p < .001$; **Figure 2.2h**, one-sample *t*-test, $t(99) = -4.76$, $p < .001$) and counterfactual (**Figure 2.3d**, one-sample *t*-test, $t(99) = 107.97$, $p < .001$; **Figure 2.3h**, one-sample *t*-test, $t(99) = -58.49$, $p < .001$) contexts. Although the Asymmetry model obviously captured true learning rate biases in the simulated data from the Asymmetry model assuming positivity/confirmation bias (**Figure 2.2a**, paired *t*-test, $t(99) = 56.07$, $p < .001$; **Figure 2.3a**, rmANOVA, $F(1,99) = 4842.84$, $p < .001$) and the Asymmetry model assuming negativity/opposite confirmation bias (**Figure 2.2e**, paired *t*-test, $t(99) = -52.45$, $p < .001$; **Figure 2.3e**, rmANOVA, $F(1,99) = 55408.02$, $p < .001$), I also replicated the previous finding by showing that pseudo-asymmetry of learning rates occurred when the Asymmetry model was fitted to the simulated data from the Perseverance model (**Figure 2.2i**, paired *t*-test, $t(99) = 16.82$, $p < .001$; **Figure 2.3i**, rmANOVA, $F(1,99) = 1754.38$, $p < .001$). These results indicate that an inadequate model causes either pseudo-asymmetric learning rates or pseudo-perseverance.

Finally, I investigated whether the Hybrid model could dissociate these underlying processes (i.e., asymmetric value updating and perseverance). Our results clearly demonstrate that the Hybrid model could capture the genuine process underlying choice behavior. When the Hybrid model was fitted to the simulated data generated from the true Asymmetry model assuming positivity/confirmation and negativity/opposite confirmation bias, the bias of learning rates was captured by the Hybrid model (**Figure 2.2b** and **2.2c**, Asymmetry vs Hybrid with paired *t*-test , $t(99) = 1.61$, $p = .11$; **Figure 2.2f** and **2.2g**, Asymmetry vs Hybrid with paired *t*-test, $t(99) = -1.35$, $p = .18$; **Figure 2.3b** and **2.3c**, Asymmetry vs Hybrid with paired *t*-test, $t(99) = 1.07$, $p = .29$; **Figure 2.3f** and **2.3g,** Asymmetry vs Hybrid with paired *t*-test, $t(99) = .05$, $p = .96$), and the pseudo-perseverance induced by fitting the Perseverance model was controlled (**Figure 2.2d**, paired *t*-test, $t(99) = 11.90$, $p < .001$; **Figure 2.2h**, paired *t*-test, $t(99) = -4.50$, $p < .001$; **Figure 2.3d**, paired *t*-test, $t(99) = 24.56$, $p < .001$; **Figure 2.3h**, paired *t*-test, $t(99) = -36.16$, $p < .001$). When the Hybrid model was fitted to the simulated data generated from the true Perseverance model, the pseudo-bias of learning rates induced by fitting the Asymmetry model was controlled (**Figure 2.2j** and **2.2k**, Asymmetry vs Hybrid with paired *t*-test, $t(99) = 20.56$, $p < .001$; **Figure**

2.3j and **2.3k**, Asymmetry vs Hybrid with paired *t*-test, $t(99) = 17.88$, $p < .001$), while the perseverance parameter was captured by the Perseverance model (**Figure 2.2l**, paired *t*-test, $t(99) = -.49$, $p = .63$; **Figure 2.3l**, paired *t*-test, $t(99) = .82$, $p = .42$). Furthermore, when the Hybrid model was fitted to the simulated data generated from the true Hybrid model assuming positivity/confirmation and negativity/opposite confirmation bias, the Hybrid model identified the true parameters related to both asymmetric updating and perseverance in each learning context (**Figure 2.4** and **2.5**). These data confirm that the Asymmetry, Perseverance, and Hybrid models were identifiable. Given that the advantage of the Hybrid model was validated, I subsequently applied the empirical data collected in the web-based experiment and open data from previous studies.

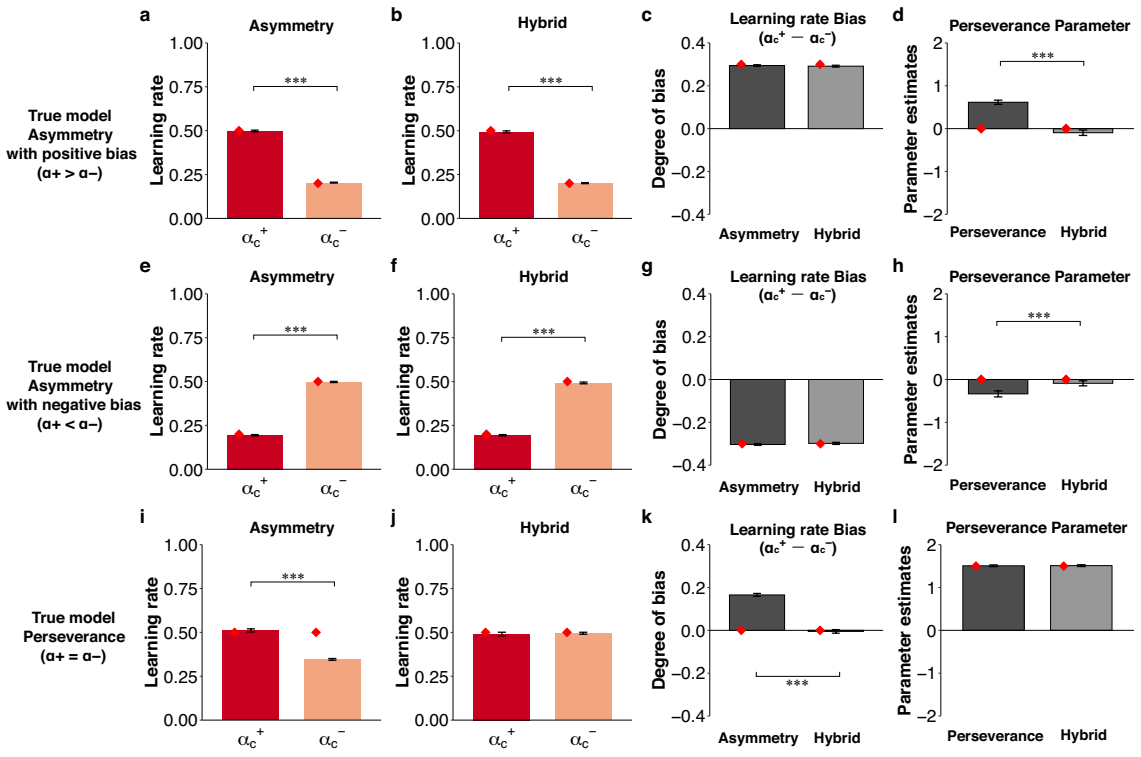


Figure 2.2 The results of the simulation in the factual learning context. (a - d) The results of the true model with asymmetric learning rates assuming positivity bias ($\alpha_c^+ = 0.5, \alpha_c^- = 0.2$). (e - h) The results of the true model with asymmetric learning rates assuming negativity bias ($\alpha_c^+ = 0.2, \alpha_c^- = 0.5$). (i - l) The results of the true model assuming symmetric learning rates ($\alpha_c^+ = \alpha_c^- = 0.5$) and choice perseverance ($\varphi = 1.5$). (a, e, i) The first column indicates the learning rates in the Asymmetry model. (b, f, j) The second column indicates the learning rates (α_c^+ , α_c^-) in the Hybrid (gradual) model. (c, g, k) The third column shows the degree of learning rate bias ($\alpha_c^+ - \alpha_c^-$). (d, h, l) The final column shows the perseverance parameter (φ) in the Perseverance

(gradual) and Hybrid (gradual) models. ***p < .001, **p < .01 and *p < .05. The error bars represent the standard error of the mean. The diamonds denote the ground-truth value of the parameters used in the data generation.

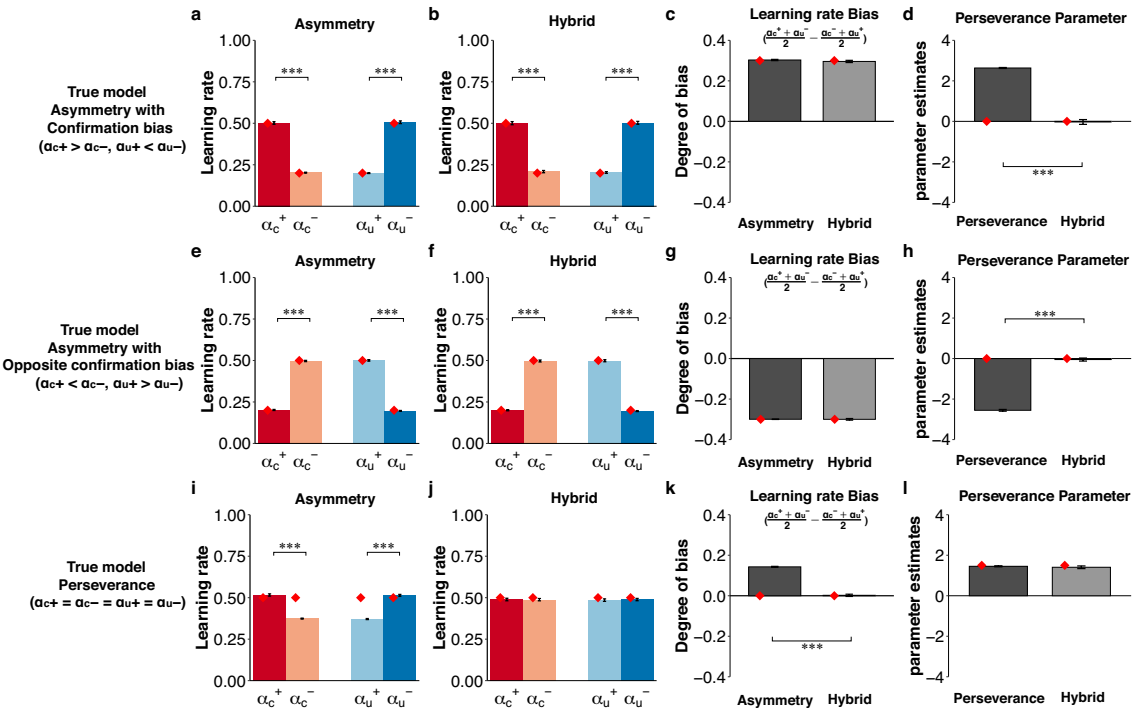


Figure 2.3 The results of the simulation in the counterfactual learning context. (a - d) The results of the true model with asymmetric learning rates assuming confirmation bias ($\alpha_c^+ = 0.5, \alpha_c^- = 0.2, \alpha_u^+ = 0.2, \alpha_u^- = 0.5$). (e - h) The results of the true model with asymmetric learning rates assuming opposite confirmation bias ($\alpha_c^+ = 0.2, \alpha_c^- = 0.5, \alpha_u^+ = 0.5, \alpha_u^- = 0.2$). (i - l) The panel shows the results of the true model with symmetric learning rates ($\alpha_c^+ = \alpha_c^- = \alpha_u^+ = \alpha_u^- = 0.5$) and choice perseverance (φ = 1.5). (a, e, i) The first column indicates the learning rates (α_c^+ , α_c^- , α_u^+ , α_u^-) in the Asymmetry model. (b, f, j) The second column indicates the learning rates in the Hybrid (gradual) model. (c, g, k) The third column indicates the degree of confirmation bias ($\frac{\alpha_c^+ + \alpha_u^-}{2} - \frac{\alpha_c^- + \alpha_u^+}{2}$). (d, h, l) The final column shows the perseverance parameter (φ) in the Perseverance (gradual) and Hybrid (gradual) models. ***p < .001, **p < .01 and *p < .05. The error bars represent the standard error of the mean. The diamonds denote the ground-truth value of the parameters used in the data generation.

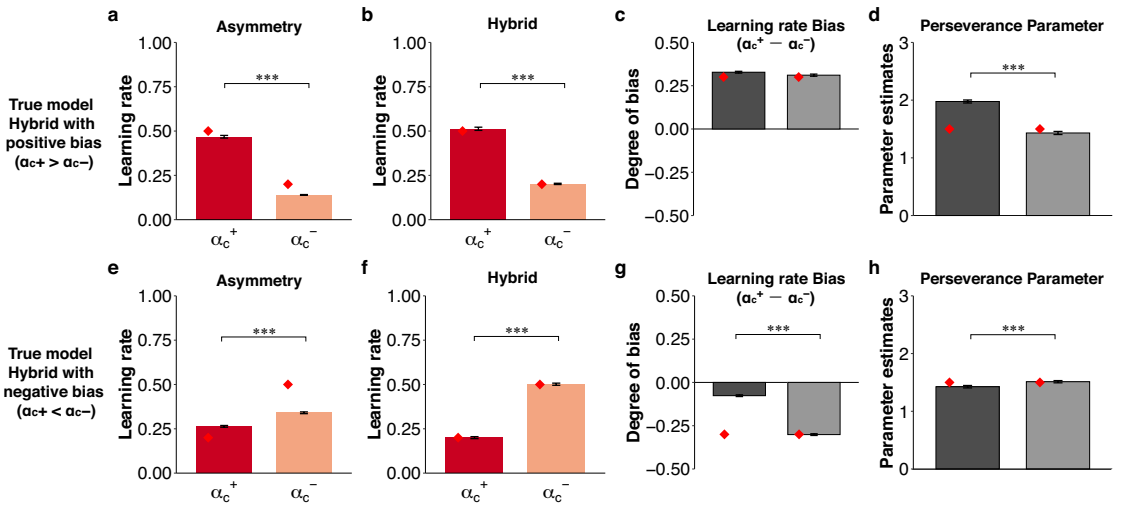


Figure 2.4 The results of the simulations generated from the Hybrid model (gradual) in the factual learning context. (a - d) The results of the true model with asymmetric learning rates assuming positivity bias ($\alpha_c^+ = 0.5, \alpha_c^- = 0.2$) and choice perseverance. (e - h) The results of the true model with asymmetric learning rates assuming negativity bias ($\alpha_c^+ = 0.2, \alpha_c^- = 0.5$) and choice perseverance. The first and second columns indicate the learning rates (α_c^+ and α_c^-) in the Asymmetry (a, e) and Hybrid (gradual) (b, f) models, respectively. (c, g) The third column shows the degree of learning rate bias ($\alpha_c^+ - \alpha_c^-$). (d, h) The final column shows the perseverance parameter (φ) in the Perseverance (gradual) and Hybrid (gradual) models. ***$p < .001$, **$p < .01$ and *$p < .05$. The error bars represent the standard error of the mean. The diamonds denote the ground-truth value of the parameters used in the data generation.

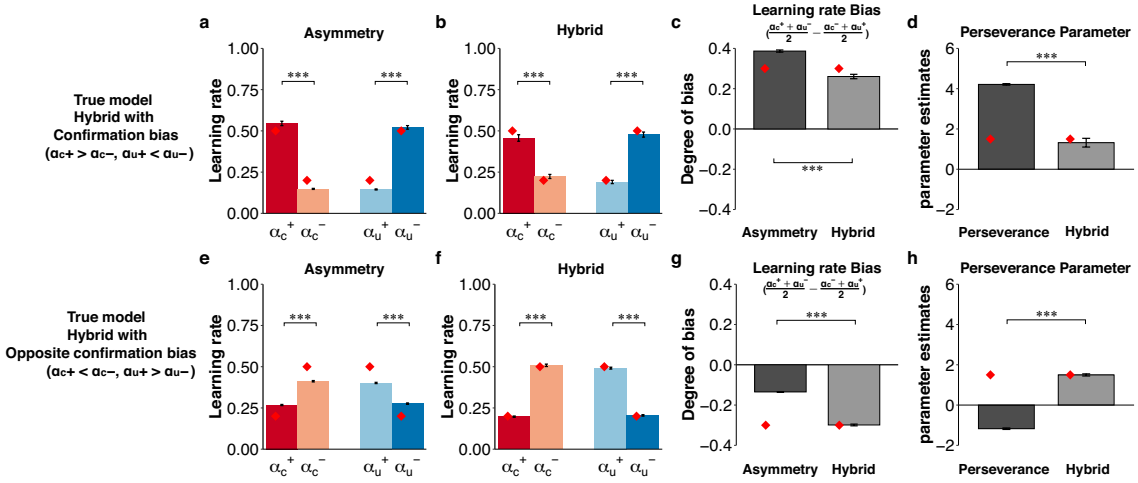


Figure 2.5 The results of the simulations generated from the Hybrid (gradual) model in the counterfactual learning context. (a - d) The results of the true model with asymmetric learning

rates assuming confirmation bias ($\alpha_c^+ = 0.5, \alpha_c^- = 0.2, \alpha_u^+ = 0.2, \alpha_u^- = 0.5$) and choice perseverance. (e - h) The results of the true model with asymmetric learning rates assuming opposite confirmation bias ($\alpha_c^+ = 0.2, \alpha_c^- = 0.5, \alpha_u^+ = 0.5, \alpha_u^- = 0.2$) and choice perseverance. The first and second columns indicate the learning rates (α_c^+, α_c^-, α_u^+, and α_u^-) in the Asymmetry (a, e) and Hybrid (gradual) (b, f) models, respectively. (c, g) The third column shows the degree of learning rate bias ($\frac{\alpha_c^+ + \alpha_u^-}{2} - \frac{\alpha_c^- + \alpha_u^+}{2}$). (d, h) The final column shows the perseverance parameter (φ) in the Perseverance (gradual) and Hybrid (gradual) models. ***$p < .001$, **$p < .01$ and *$p < .05$. The error bars represent the standard error of the mean. The diamonds denote the ground-truth value of the parameters used in the data generation.

2.3.2 Application of the Hybrid model to empirical data

Our subsequent aim was to evaluate the extent to which asymmetric updating and choice perseverance influence actual human choice behavior. To reliably achieve this goal, I conducted a web-based experiment to obtain a relatively large sample size (N = 143 per context; see details in the 'Methods' section).

Behavioral performance in the web-based experiment

To determine whether the subjects in the web-based experiment adequately learned the probabilistic instrumental learning task, I conducted a two-way repeated-measures analysis of variance (rmANOVAs) with the effects of Context (factual and counterfactual) and Phase (1st to 4th) of the preferred response rate under the same condition and the correct rate under the different and reversal conditions (**Figure 2.6**).

Under the same condition, a significant main effect of Phase was observed (F(2.82,400) = 3.74, p = .01), whereas the main effect of Context and the interaction were not significant (Context: F(1,142) = 2.21, p = .14; Context × Phase: F(2.71,385.31) = 3.74, p = .71). The post hoc comparisons showed that the preferred rate significantly increased from the 1st to 3rd phase ($p < .01$). Under the different condition, the main effect of Phase and the interaction were significant (Phase: F(2.76,392.35) = 30.07, $p < .001$; Context × Phase: F(2.69,382.62) = 4.50, $p < .01$), and no main effect of Context was observed (F(1,142) = 2.68, $p = .10$). Significant simple effects of Phase were observed in both the factual and counterfactual contexts (factual: F(2.71,384.91) = 9.82, $p < .001$; counterfactual: F(2.63,373.86) = 28.26, $p < .001$). Additionally, significant simple effects of Context were found in the 2nd and 3rd phases

(2nd: $F(1,142) = 6.79$, $p < .05$; 3rd: $F(1,142) = 7.25$, $p < .001$). The post hoc comparisons showed that the correct rate significantly increased from the 1st to 2nd ($p < .05$) and from the 3rd to 4th ($p < .05$) phases in the factual learning context, whereas the correct rate increased only from the 3rd to 4th phases ($p < .05$) in the counterfactual learning context. These results suggest that the subjects successfully learned the reward probabilities of the presented options. Under the reversal condition, the main effect of Phase was significant ($F(2.16,306.59) = 102.5$, $p < .001$), but the main effect of Context and the interaction were not (Context: $F(1,142) = .002$, $p = .97$; Context × Phase: $F(2.3,326.03) = 1.56$, $p = .21$). The post hoc comparisons indicated that the correct rate significantly increased in the 2nd phase, decreased in the 3rd phase, and then increased again in the 4th phase ($ps < .001$). This profile confirmed that the subjects detected the reversal in the reward probability.

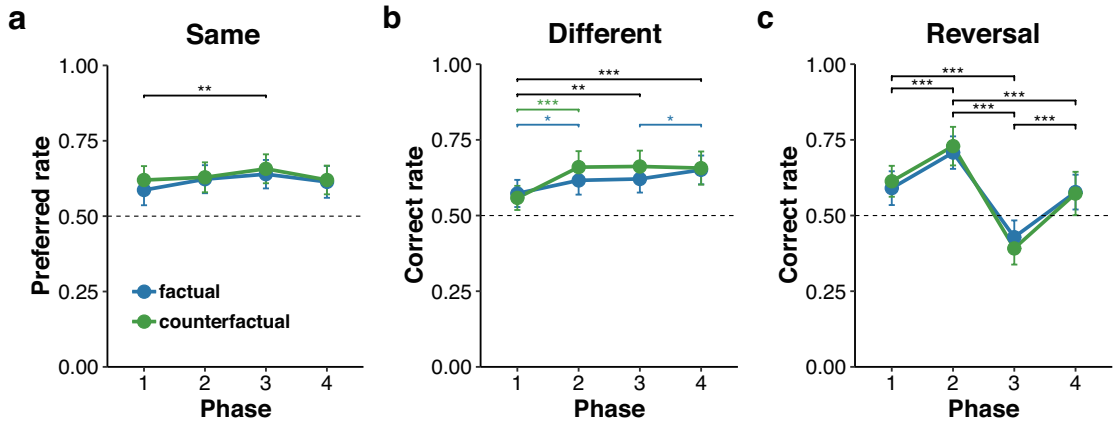


Figure 2.6 Behavioral performances across the three conditions in the factual and counterfactual contexts of the web-based experiment. I divided each learning session into four subphases. Each curve shows the average performances of 143 participants in the four phases. The error bars represent the standard errors of the mean. (a) Under the same condition, performance was measured as the preferred rate. Under both the different (b) and reversal (c) conditions, performance was measured as the correct rate. Blue and green denote the factual and counterfactual learning contexts, respectively. ***$p < .001$, **$p < .01$, and *$p < .05$.

Model comparisons using web-based experimental data

In addition to the three models used in the simulation (Asymmetry, Perseverance, and Hybrid models), a standard RL model was fitted to the empirical datasets as a benchmark for the model comparisons. Furthermore, I used two variants of the Perseverance and Hybrid models.

The original models used in the simulation have a gradual decay rate ($0 \leqq \tau \leqq 1$) in which several preceding choices influence the current choice. In addition to these Perseverance and Hybrid models with a gradual decay rate, I also fitted Perseverance and Hybrid models with an impulsive decay rate ($\tau = 1$) in which only the immediately preceding choice influences the current choice because this type of decay rate was included in the models reported in a previous study (Palminteri, Lefebvre, et al., 2017). Thus, I applied six models (i.e., RL, Asymmetry, Perseverance (impulsive), Hybrid (impulsive), Perseverance (gradual), and Hybrid (gradual) models) to the subjects' choice behavior and then compared these models using log marginal likelihood (LML; **Table 2.2**).

In the factual learning context, the Perseverance model was the best among the six models (rmANOVA; $F(1.73,246.04) = 17.69$, $p < .001$; post hoc comparison: $ps < .05$) but was comparable with the Hybrid (gradual) model (post hoc comparison: $p > .99$). In the counterfactual learning context, the log marginal likelihood was decreased in the order of Perseverance (gradual), Hybrid (gradual), Asymmetry, Perseverance (impulsive), Hybrid (impulsive), and RL models (rmANOVA; $F(1.74,246.92) = 31.09$, $p < .001$). The Perseverance (gradual) model was the best among the six models (post hoc comparisons: $ps < .05$). These results indicate that the preceding choices greatly influenced the current choice in both learning contexts.

Table 2.2 List of models and model selection results of the web-based experiment data

Learning context	Model	Learning rate (s)	Inverse temperature	Perseverance	# of free parameters	Log marginal likelihood (SD)
Factual learning	RL	α	β	–	2	-123.99 (15.64)
	Asymmetry	α_c^+, α_c^-	β	–	3	-120.33 (19.36)
	Perseverance (inpulsive)	α	β	τ=1, φ	3	-121.12 (20.20)
	Perseverance (gradual)	α	β	τ, φ	4	**-118.41 (21.83)**
	Hybrid (implusive)	α_c^+, α_c^-	β	τ=1, φ	4	-120.51 (21.08)
	Hybrid (gradual)	α_c^+, α_c^-	β	τ, φ	5	-118.69 (22.47)
Counterfactual learning	RL	α	β	–	2	-121.27 (20.98)
	Asymmetry	α_c^+, α_c^-, α_u^+, α_u^-	β	–	5	-115.28 (25.83)
	Perseverance (inpulsive)	α	β	τ=1, φ	3	-116.97 (25.50)
	Perseverance (gradual)	α	β	τ, φ	4	**-113.18 (27.21)**
	Hybrid (implusive)	α_c^+, α_c^-, α_u^+, α_u^-	β	τ=1, φ	6	-117.27 (26.64)
	Hybrid (gradual)	α_c^+, α_c^-, α_u^+, α_u^-	β	τ, φ	7	-114.54 (27.01)

Parameter estimates using web-based experiment data

To empirically confirm that the Hybrid model can evaluate the degree of asymmetric value updating by controlling the pseudo-bias of learning rates, I compared the estimated learning rates among the three models (Asymmetry, Hybrid (impulsive), and Hybrid (gradual) models; **Table 2.3**). I predicted that if the bias of learning rates estimated by fitting the Asymmetry model was pseudo-bias, this bias should disappear by fitting the Hybrid (gradual) model. This prediction was confirmed in the factual learning context. The positivity bias of learning rates ($\alpha_c^+ > \alpha_c^-$) observed in the Asymmetry model (**Figure 2.7a**; $t(142) = 4.70$, $p < .001$) disappeared by fitting the Hybrid (gradual) model (**Figure 2.7c**; $t(142) = -.54$, $p = .59$) but not by fitting the Hybrid (impulsive) model (**Figure 2.7b**; $t(142) = 2.26$, $p = .03$). Indeed, the degree of positivity bias decreased in the order of the Asymmetry, Hybrid (impulsive), and Hybrid (gradual) models (**Figure 2.7d**; $F(1.42,202.94) = 37.26$, $p < .001$; post hoc comparisons: all $ps < .001$). In the counterfactual learning context, our prediction was also confirmed. According to a previous study (Palminteri, Lefebvre, et al., 2017), confirmation bias in RL is characterized as follows: the learning rates of the outcome that supports one's choice (i.e., learning rate of the positive outcome of the chosen option (α_c^+) and negative outcome of the unchosen option (α_u^-) are higher than the learning rates that do not support one's choice (i.e., learning rate of the negative outcome of the chosen option (α_c^-) and positive outcome of the unchosen option (α_u^+)). The confirmation bias of the learning rates observed in the Asymmetry model (**Figure 2.7f**; a two-way repeated-measures ANOVA; interaction: $F(1,142) = 155.21$, $p < .001$) was diminished by fitting the Hybrid (gradual) model (**Figure 2.7h**; interaction: $F(1,142) = .85$, $p = .36$) but not by fitting the Hybrid (impulsive) model (**Figure 2.7g**; interaction: $F(1,142) = 83.73$, $p < .001$). The degree of confirmation bias ($\frac{\alpha_c^+ + \alpha_u^-}{2} - \frac{\alpha_c^- + \alpha_u^+}{2}$) was significantly decreased in the order of the Asymmetry, Hybrid (impulsive), and Hybrid (gradual) models (**Figure 2.7i**; $F(1.28, 182.05) = 59.70$, $p < .001$; post hoc comparisons: all $ps < .001$).

Furthermore, to confirm that the Hybrid model can evaluate the degree of choice perseverance by controlling pseudo-perseverance, I examined the perseverance parameter (φ) in the four models (Perseverance (impulsive), Perseverance (gradual), Hybrid (impulsive), and Hybrid (gradual) models; **Table 2.3**). In the factual learning context, the perseverance parameters in the Perseverance (gradual) and Hybrid (gradual) models were comparable (**Figure 2.7e**; rmANOVA: $F(1.62,230.24) = 38.90$, $p < .001$; post hoc comparisons: $p = .12$) but significantly higher than those in the Perseverance (impulsive) and Hybrid (impulsive) models ($ps < .001$). Similarly, in the counterfactual learning context, the perseverance parameters in the Perseverance

(gradual) and Hybrid (gradual) models were comparable (**Figure 2.7j**; rmANOVA: $F(1.77, 251.69) = 111.47$, $p < .001$; post hoc comparisons: $p = .13$) but significantly higher than those in the Perseverance (impulsive) and Hybrid (impulsive) models ($ps < .001$).

Taken together, these results indicate that choice perseverance mainly governed choice behavior in the web-based experiment. This result also highlights that the Hybrid model allowed us to clarify a genuine process underlying the empirical choice data.

Table 2.3 List of models and parameter results of the web-based experiment data

	Model	α	α_c^+	α_c^-	α_u^+	α_u^-	β	τ	φ
Factual learning	RL	0.36 (0.32)	–	–	–	–	0.32 (0.48)	–	–
	Asymmetry	–	0.43 (0.30)	0.26 (0.32)	–	–	0.19 (0.20)	–	–
	Perseverance (inpulsive)	0.35 (0.32)	–	–	–	–	0.27 (0.36)	1.00	0.47 (0.46)
	Perseverance (gradual)	0.45 (0.32)	–	–	–	–	0.19 (0.26)	0.41 (0.30)	1.10 (1.26)
	Hybrid (implusive)	–	0.41 (0.31)	0.32 (0.34)	–	–	0.17 (0.20)	1.00	0.33 (0.35)
	Hybrid (gradual)	–	0.45 (0.34)	0.47 (0.33)	–	–	0.13 (0.15)	0.39 (0.30)	0.98 (1.35)
Counterfactual learning	RL	0.18 (0.22)	–	–	–	–	0.51 (1.04)	–	–
	Asymmetry	–	0.41 (0.30)	0.15 (0.21)	0.15 (0.23)	0.40 (0.33)	0.18 (0.23)	–	–
	Perseverance (inpulsive)	0.23 (0.27)	–	–	–	–	0.36 (0.60)	1.00	0.59 (0.54)
	Perseverance (gradual)	0.35 (0.29)	–	–	–	–	0.21 (0.33)	0.33 (0.27)	1.62 (1.32)
	Hybrid (implusive)	–	0.37 (0.30)	0.16 (0.23)	0.17 (0.25)	0.34 (0.31)	0.19 (0.26)	1.00	0.21 (0.37)
	Hybrid (gradual)	–	0.39 (0.34)	0.36 (0.33)	0.35 (0.33)	0.36 (0.33)	0.17 (0.23)	0.38 (0.32)	1.49 (1.36)

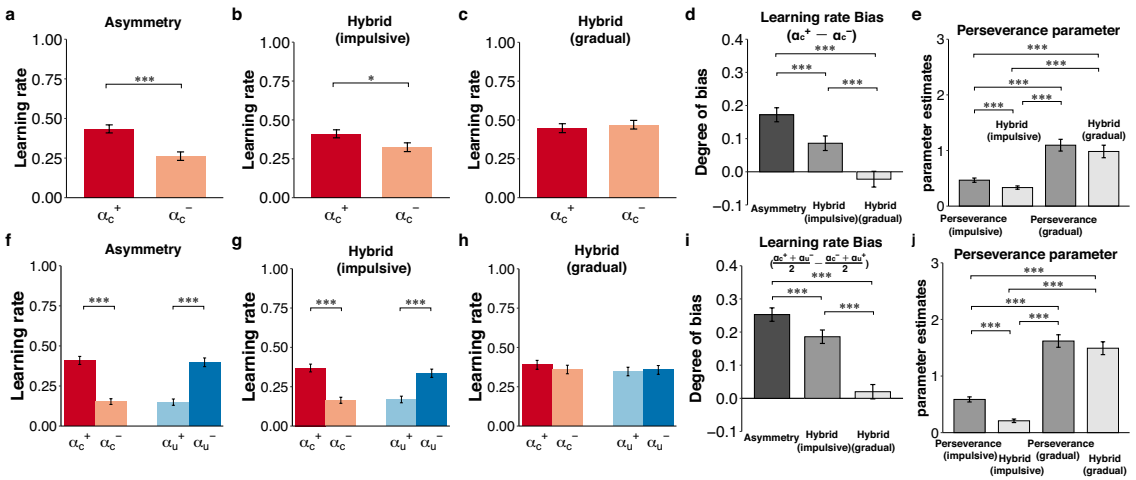


Figure 2.7 The results of the web-based experiment. The rows show the results of the web-based experiment in the factual and counterfactual learning contexts. The first to third columns

represent the learning rates in the Asymmetry (a, f), Hybrid (impulsive) (b, g), and Hybrid (gradual) models (c, h). (d, i) The fourth column indicates the degree of learning rate bias (i.e., positivity bias in the factual context and confirmation bias in the counterfactual context). (e, j) The final column shows the perseverance parameter (φ) in the Perseverance (impulsive), Perseverance (gradual), Hybrid (impulsive), and Hybrid (gradual) models. ***$p < .001$, **$p < .01$ and *$p < .05$. The error bars represent the standard error of the mean.

Parameter recovery using web-based experimental data

While it is important to identify the perseverance parameter (φ) in the Perseverance (gradual) and Hybrid (gradual) models, it is possible that the perseverance parameter and inverse temperature parameter (β) represent a trade-off (see Equation 2.4 in the 'Methods' section). To determine the identifiability of these parameters, I calculated the correlation between the estimated parameters and further performed parameter recovery under both the factual and counterfactual conditions (see 'Methods' section). The correlation analysis ensured that the perseverance parameter was not significantly correlated with the inverse temperature parameter in both learning contexts in both the Perseverance (gradual) and Hybrid (gradual) models (**Figure 2.8**; *ps* > .99). The parameter recovery also indicated that all parameters were well recovered in the factual (**Figure 2.9**; $0.84 < r < 0.98$, all *ps* < .001 with Bonferroni correction) and counterfactual (**Figure 2.9**; $.75 < r < .95$, all *ps* < .001 with Bonferroni correction) learning contexts. These results confirm that the parameter optimization procedure used in this study allowed us to identify the free parameters in each model.

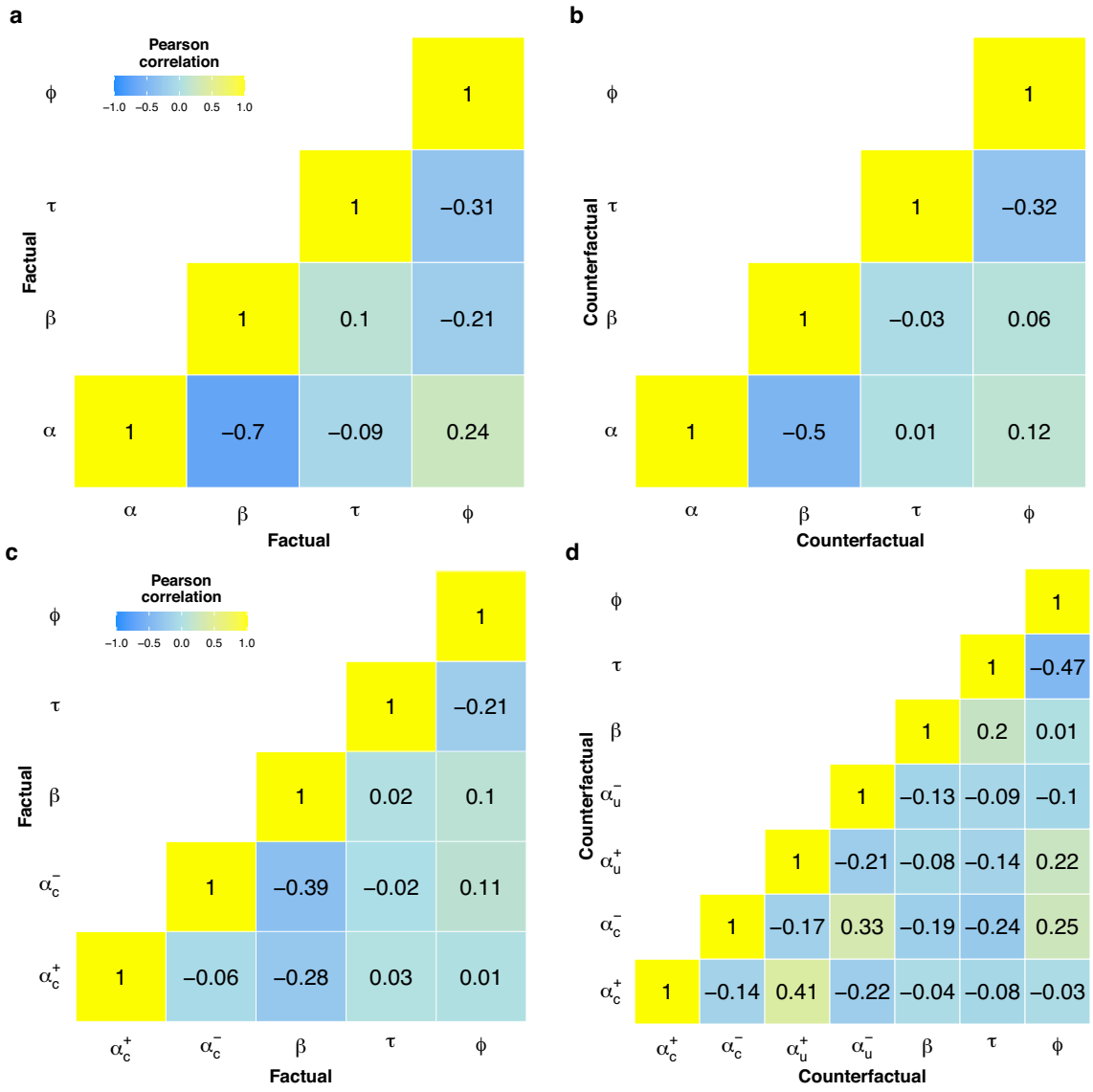


Figure 2.8 Parameter correlation in the web-based experiment. Correlation matrices indicating Pearson's correlations between the estimated parameters. The top row was estimated from the Asymmetry model in the factual (a) and counterfactual (b) learning contexts. The bottom row was estimated from the Hybrid (gradual) model in the factual (c) and counterfactual (d) learning contexts. The color and value in each cell represent Pearson's correlation coefficients.

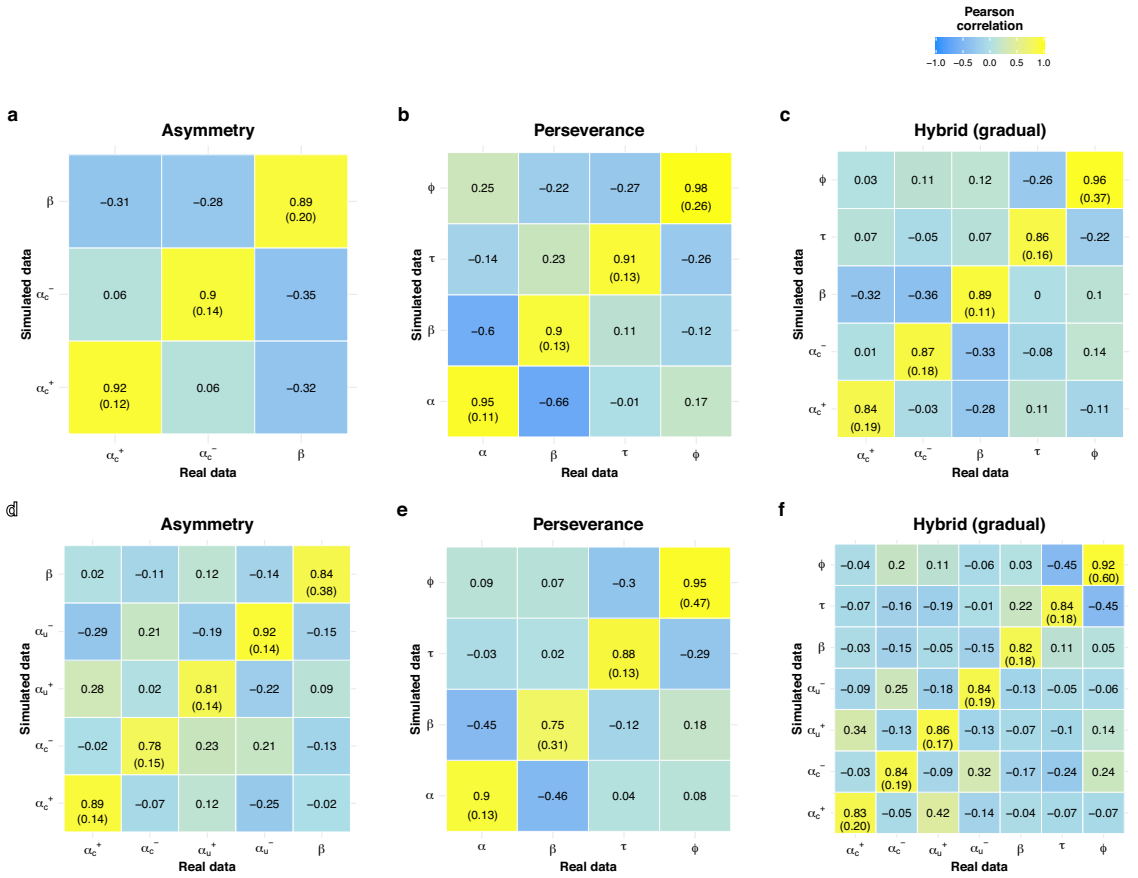


Figure 2.9 Parameter recovery in the web-based experiment. The correlation matrix represents the Pearson's correlation coefficients between the parameters estimated from the empirical data (x-axis) and the simulated data (y-axis). (a - c) The results of the Asymmetry, Perseverance (gradual), and Hybrid (gradual) models in the factual learning context. (d - f) The results of the Asymmetry, Perseverance (gradual), and Hybrid (gradual) models in the counterfactual learning context. The color and value in each cell represent the Pearson's correlation coefficients. Additionally, in the diagonal elements, the root mean squared errors between the true value used in the data generation and the estimated value are noted in the parentheses.

Model-neutral analysis

Katahira (2018) proposed the use of a model-neutral analysis to examine the existence of the asymmetric value updating process without the RL model framework. This analysis utilizes the fact that the asymmetric learning rate induces an interaction effect between past outcomes on the current choice. The merit of a model-neutral analysis is that it does not assume a specific

functional form regarding how past experience influences the reward, while RL model fitting does make this assumption. Thus, there is a possibility that the absence of the asymmetric learning rate in the RL model fitting is due to a mismatch of the functional form. To examine this possibility, I performed a model-neutral analysis (see details in Methods) of the empirical choice data. Consequently, no evidence of asymmetric value updating was observed, which is consistent with our RL model-based analysis.

In the factual learning context, the logistic regression model included the following three terms: the outcome of the chosen option at the t-th trial ($\boldsymbol{Rc_t}$), the chosen outcome at the $(t-1)$-th trial ($\boldsymbol{Rc_{t-1}}$), and the interaction between these past outcomes ($\boldsymbol{Rc_t}\times\boldsymbol{Rc_{t-1}}$). The regression coefficients of $\boldsymbol{Rc_t}$ and $\boldsymbol{Rc_{t-1}}$ were significant and positive ($\boldsymbol{Rc_t}$: $\beta = .89$, $p < .001$; $\boldsymbol{Rc_{t-1}}$: $\beta = .40$, $p < .001$, **Table 2.4**). However, the interaction was not significant ($\boldsymbol{Rc_t}\times\boldsymbol{Rc_{t-1}}$: $\beta = -.08$, $p = .30$), indicating that evidence of asymmetry in value updating during the underlying learning process was lacking. In addition, the intercept was significant and positive ($\beta = .27$, $p < .001$), suggesting a tendency to repeat choices.

In the counterfactual context, the logistic regression model including the following six terms: the chosen outcome at the t-th trial ($\boldsymbol{Rc_t}$), the chosen outcome at the $(t-1)$-th trial ($\boldsymbol{Rc_{t-1}}$), the interaction between these past outcomes ($\boldsymbol{Rc_t}\times\boldsymbol{Rc_{t-1}}$), the outcome of the unchosen option at the t-th trial ($\boldsymbol{Ru_t}$), the unchosen outcome at the $(t-1)$-th trial ($\boldsymbol{Ru_{t-1}}$), and the interaction between these latter two outcomes ($\boldsymbol{Ru_t}\times\boldsymbol{Ru_{t-1}}$). The regression coefficients of $\boldsymbol{Rc_t}$ and $\boldsymbol{Rc_{t-1}}$ were significant and positive ($\boldsymbol{Rc_t}$: $\beta = .59$, $p < .001$; $\boldsymbol{Rc_{t-1}}$: $\beta = .27$, $p < .001$, **Table 2.4**). However, the regression coefficients of $\boldsymbol{Ru_t}$ and $\boldsymbol{Ru_{t-1}}$ were significant but negative ($\boldsymbol{Ru_t}$: $\beta = -.40$, $p < .001$; Rut-1: $\beta = -.19$, $p < .001$). Furthermore, neither interactions (between the outcomes of the chosen options or between those of the unchosen options) were significant ($\boldsymbol{Rc_t}\times\boldsymbol{Rc_{t-1}}$: $\beta = -.011$, $p = .19$; $\boldsymbol{Ru_t}\times\boldsymbol{Ru_{t-1}}$: $\beta = -.09$, $p = .27$), further indicating that evidence of asymmetric value updating is lacking. The intercept was significant and positive ($\beta = .98$, $p < .001$).

Table 2.4 Regression coefficients of the logistic regression model in the model-neutral analysis

Learning context	Effect	*Beta* (*SE*)	*z-value*	*p-value*
Factual learning	Intercept	0.27 (0.07)	3.78	< 0.001 ***
	Rc_t	0.89 (0.08)	11.74	< 0.001 ***
	Rc_{t-1}	0.40 (0.06)	6.56	< 0.001 ***
	$Rc_t \times Rc_{t-1}$	-0.08 (0.08)	-1.04	0.30
Counterfactual learning	Intercept	0.98 (0.09)	10.93	< 0.001 ***
	Rc_t	0.59 (0.08)	7.76	< 0.001 ***
	Rc_{t-1}	0.27 (0.06)	4.72	< 0.001 ***
	Ru_t	-0.40 (0.07)	-5.88	< 0.001 ***
	Ru_{t-1}	-0.19 (0.06)	-3.11	< 0.01 **
	$Rc_t \times Rc_{t-1}$	-0.11 (0.08)	-1.32	0.19
	$Ru_t \times Ru_{t-1}$	-0.09 (0.08)	-1.1	0.27

Note: Rc_t = chosen outcome at t-th trial; Rc_{t-1} = chosen outcome at (t - 1)-th trial; Ru_t = unchosen outcome at t-th trial; and Ru_{t-1} = unchosen outcome at (t - 1)-th trial.

2.3.3 Application of the Hybrid model using open data

As shown above, the Hybrid model allowed us to identify a genuine process underlying the empirical choice data. Here, to reconsider the processes underlying open datasets collected by previous studies reporting asymmetric value updating, I re-analyzed these open datasets by applying the Hybrid model. Similar to the web-based experiment, I fitted six models (the RL, Asymmetry, Perseverance (impulsive), Perseverance (gradual), Hybrid (impulsive), and Hybrid (gradual) models) to these open datasets and compared the parameter estimates.

Dataset 1 (Palminteri et al., 2017)

Dataset 1 comprised the open data reported by Palminteri, Lefebvre, et al. (2017), who examined the asymmetric learning rates in both the factual and counterfactual learning contexts. The model comparisons (**Table 2.5**) showed that the Perseverance (gradual) model was the best

among the models in both the factual (F(1.11,21.16) = 6.41, p = .02) and counterfactual (F(1.39,26.32) = 26.58, p < .001) learning contexts.

In the factual learning context, by fitting the Asymmetry model, I replicated the finding showing that the learning rate of the positive outcome (α_c^+) was significantly higher than that of the negative outcome (α_c^-) (**Figure 2.10a**; paired t-test, t(19) = 2.36, p = .03), supporting positivity bias. However, this positivity bias ($\alpha_c^+ - \alpha_c^-$) was decreased by fitting the Hybrid (impulsive) model (**Figure 2.10b**; paired t-test, t(19) = 1.35, p = .19) and was diminished by fitting the Hybrid (gradual) model (**Figure 2.10c**; paired t-test, t(19) = .15, p = .88). Indeed, the degree of positivity bias was significantly smaller in the order of the Asymmetry, Hybrid (impulsive), and Hybrid (gradual) models (**Figure 2.10d**; F(1.53,28.98) = 15.95, p < .001; post hoc comparisons, all ps < 4.47×10^{-3}). Although the degree of the perseverance parameter significantly differed among the models (**Figure 2.10e**; rmANOVA, F(2.05, 39.04) = 16.09, p < .001; post hoc comparisons, all ps < .047), the perseverance parameters (φ) estimated in the Perseverance and Hybrid models were above zero, leading to repeat preceding choices.

In the counterfactual learning context, I also replicated the finding showing that the learning rate of positive RPE was greater than that of negative RPE in terms of the chosen outcomes (i.e., $\alpha_c^+ > \alpha_c^-$), but the opposite was observed in terms of the unchosen outcomes (i.e., $\alpha_u^+ < \alpha_u^-$) (**Figure 2.10f;** two-way rmANOVA, interaction: F(1,19) = 124.88, p < .001), indicating confirmation bias. Although this confirmation bias was also observed by fitting the Hybrid (impulsive) (**Figure 2.10g**; F(1,19) = 53.45, p < .001) and Hybrid (gradual) models (**Figure 2.10h**; F(1,19) = 5.58, p = .03), a significant difference in the learning rates was not observed between the positive and negative RPE of both the chosen and unchosen outcomes in the Hybrid (gradual) model (ps > .15). The degree of confirmation bias in the Hybrid (gradual) model was significantly smaller than that in the Hybrid (impulsive) model (**Figure 2.10i**; rmANOVA, F(1.21, 23) = 7.10, p = .010; post hoc comparisons, p = .04). The perseverance parameter in the Hybrid (gradual) model was smaller than that in the Perseverance (gradual) model (**Figure 2.10j**; rmANOVA, F(1.69, 32.16) = 25.98, p < .001; post hoc comparison, p = 1.43×10^{-3}) but remained positive.

According to these results, the view claimed in the previous study (i.e., the existence of asymmetry in the learning rate) was not supported. In contrast, our results suggest that the choice behavior in Dataset 1 was mainly governed by choice perseverance rather than asymmetric value updating in both the factual and counterfactual learning contexts.

Dataset 2 (Niv et al., 2012)

Dataset 2 comprised the open data reported by Niv et al. (2012), who applied the Asymmetry model to explain risk-seeking/aversion behaviors in the factual learning context. In contrast to Dataset 1, the Asymmetry model was better than the Hybrid (impulsive) and Hybrid (gradual) models (**Table 2.5**; rmANOVA, $F(1.43,21.52) = 4.29$, $p = .04$; post hoc comparisons, $ps = 3.01\times10^{-3}$) but did not significantly differ from the RL, Perseverance (impulsive), and Perseverance (gradual) models ($ps > .41$).

As Niv et al. (2012) reported, in the Asymmetry model, the learning rate of positive RPE (α_c^+) was significantly lower than that of negative RPE (α_c^-) (**Figure 2.10k**; $t(15) = -3.07$, $p = 7.80\times10^{-3}$). This negativity bias were also observed in the Hybrid (impulsive) (**Figure 2.10l**; $t(15) = -3.06$, $p = 7.93\times10^{-3}$) and Hybrid (gradual) models (**Figure 10m**; $t(15) = -2.84$, $p = .012$). The degree of negativity bias ($\alpha_c^+ - \alpha_c^-$) was comparable among these models (**Figure 2.10n**; rmANOVA, $F(1.01,15.19) = .75$, $p = .40$). Additionally, the perseverance parameter (φ) in the Hybrid (gradual) model was almost zero and did not significantly differ from that in the Perseverance (impulsive), Perseverance (gradual), and Hybrid (impulsive) models (**Figure 2.10o**; rmANOVA, $F(1.55, 23.23) = 0.87$, $p = .41$). Thus, our results based on the Hybrid model support the asymmetric value updating process claimed in a previous study (Niv et al., 2012).

Table 2.5 Models and model selection results of Dataset 1 (Palminteri et al., 2017) and Dataset 2 (Niv et al., 2012)

Learning context	Model	Learning rate (s)	Inverse temperature	Perseverance	# of free parameters	LML (SD) Dataset1	LML (SD) Dataset2
Factual learning	RL	α	β	–	2	-99.23 (24.09)	-67.62 (6.69)
	Asymmetry	α_c^+, α_c^-	β	–	3	-90.21 (26.09)	**-67.60 (8.57)**
	Perseverance (inpulsive)	α	β	τ=1, φ	3	-91.34 (27.25)	-69.57 (7.03)
	Perseverance (gradual)	α	β	τ, φ	4	**-87.50 (28.12)**	-68.44 (7.68)
	Hybrid (implusive)	α_c^+, α_c^-	β	τ=1, φ	4	-90.17 (27.12)	-69.66 (8.85)
	Hybrid (gradual)	α_c^+, α_c^-	β	τ, φ	5	-88.55 (27.44)	-69.35 (8.66)
Counterfactual learning	RL	α	β	–	2	-89.18 (25.91)	–
	Asymmetry	α_c^+, α_c^-, α_u^+, α_u^-	β	–	5	-76.09 (29.89)	–
	Perseverance (inpulsive)	α	β	τ=1, φ	3	-80.01 (29.87)	–
	Perseverance (gradual)	α	β	τ, φ	4	**-75.82 (29.35)**	–
	Hybrid (implusive)	α_c^+, α_c^-, α_u^+, α_u^-	β	τ=1, φ	6	-78.15 (30.24)	–
	Hybrid (gradual)	α_c^+, α_c^-, α_u^+, α_u^-	β	τ, φ	7	-77.11 (30.35)	–

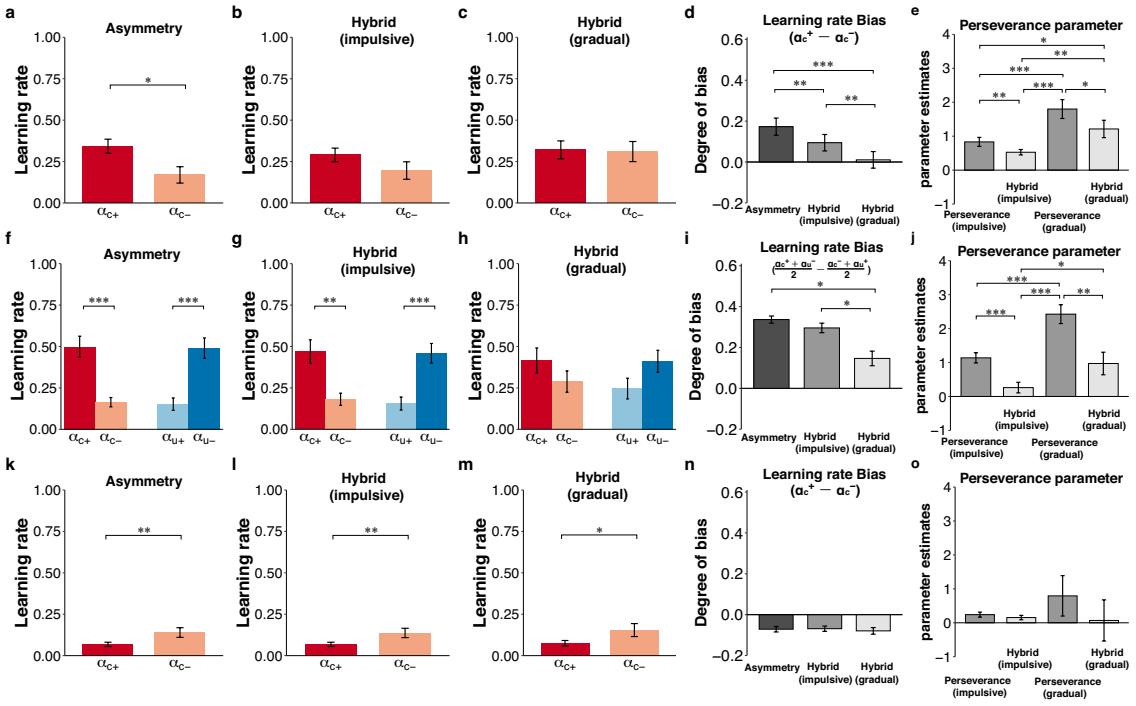


Figure 2.10 The results of open datasets 1 and 2. (a - e) The results of open dataset 1 (Palminteri et al., 2017) in the factual learning context. (f - j) The results of open dataset 1 in the counterfactual learning context. (k - o) The results of open dataset 2 (Niv et al., 2012). The first to third columns indicate the learning rates (α_c^+ and α_c^- in the factual learning context; α_c^+, α_c^-, α_u^+, and α_u^- in the counterfactual context) in the Asymmetry (a, f, k), Hybrid (impulsive) (b, g, l), and Hybrid (gradual) (c, h, m) models. (d, i, n) The fourth column shows the degree of learning rate bias. (e, j, o) The final column shows the perseverance parameter (φ) in the Perseverance (impulsive), Perseverance (gradual), Hybrid (impulsive), and Hybrid (gradual) models. ***$p < .001$, **$p < .01$ and *$p < .05$. The error bars represent the standard error of the mean.

2.4 Discussion

This study considered a method to dissociate two factors underlying human choice behavior, i.e., asymmetric learning and choice perseverance. By using these methods, I attempted to identify the processes underlying human choice behavior. In the simulation, I replicated previous findings (Katahira, 2018) showing that pseudo-asymmetric updating was induced when a model without perseverance (Asymmetry model) was fit to simulated data from a model with symmetric updating and perseverance (Perseverance (gradual) model). In contrast, when a model

without perseverance was fitted to the simulated data generated from a model with true asymmetric updating, pseudo-perseverance was observed. As Katahira (2018) mentioned, these results show that asymmetric updating and choice perseverance result in similar choice behavior statistical properties. Therefore, it is important to investigate how to dissociate these processes underlying choice behavior. In this study, I considered the Hybrid model, which incorporating both asymmetric updating and perseverance components, and I tested the capability of the Hybrid model using simulated and empirical datasets. The simulations showed that the Hybrid model could identify the following true parameters in the simulated dataset generated from all hypothetical models: optimistic asymmetric updating, pessimistic asymmetric updating, and symmetric updating with perseverance. The Hybrid model also identified the true parameters of the simulated dataset from a hypothetical model containing asymmetric updating and perseverance. These results support the advantage of the Hybrid model in distinguishing the processes underlying choice behavior.

Palminteri, Lefebvre, et al. (2017) claimed that asymmetric value updating underlies choice behavior in a probabilistic instrumental learning task. Their candidate models also included the Perseverance model and showed that an asymmetric learning rate model attained a better fit than the Perseverance model (Palminteri, Lefebvre, et al., 2017). However, their Perseverance model only considered impulsive perseverance (the influence of only the most recent choice under the same condition). As Katahira (2018) noted, a model that considers only impulsive perseverance is insufficient for avoiding statistical bias in estimates of the learning rate. Thus, there is a possibility that the overlooked influence of a more distant past induces pseudo-asymmetric learning rates.

To determine whether learning asymmetry or perseverance is dominant in choice behavior in a probabilistic instrumental learning task while addressing the above issue, I applied the Hybrid model (with gradual perseverance) to the empirical data. To obtain the empirical data, I mainly focused on data collected in a web-based experiment involving relatively large samples (N = 143 per context; compared with the previous study, N = 20 per context) to improve the statistical robustness. As previously reported (Lefebvre et al., 2017; Palminteri, Lefebvre, et al., 2017), I replicated the asymmetry in learning rates in both factual and counterfactual learning contexts in a model without the perseverance factor (Asymmetry model). The learning rates of the chosen outcomes when the outcomes were positive were greater than those when the outcomes were negative, whereas the opposite pattern was observed in the learning rates of the unchosen outcomes. Such asymmetry was interpreted as "confirmation bias" in a previous study (Palminteri, Lefebvre, et al., 2017). However, I found that this asymmetry in learning rates disappeared when

the Hybrid model was fitted, including the gradual perseverance factor ($\tau < 1$) and the asymmetric learning rate (Hybrid (gradual) models. Moreover, the model-neutral analysis did not support the existence of asymmetric value updating. These findings support our previous claim that model misspecification in which perseverance is not considered in the model can cause the erroneous detection of asymmetry in the learning rates of choice behavior (Katahira, 2018). Our results also highlight the merit of the Hybrid model in identifying the underlying process in empirical data.

I also showed that when the Hybrid model, which included impulsive perseverance ($\tau = 1$), was fitted, the asymmetric learning rates were significant in both contexts. Furthermore, I demonstrate that this residual asymmetry of learning rates disappeared when using the Hybrid model that incorporated gradual perseverance ($\tau < 1$). Indeed, similar results were obtained using open data in a previous study (Palminteri, Lefebvre, et al., 2017). These findings suggest that the superiority of the asymmetric learning model over the perseverance model in the previous study was due to an insufficient length of the choice history.

Furthermore, I demonstrated that the Hybrid model could identify asymmetric updating in empirical data obtained in a different type of task. In the open data reported by Niv et al. (2012), the asymmetry in the learning rates remained after controlling for choice perseverance. The factor inducing asymmetry in value updating in the context of reinforcement learning remains unclear. It is possible that the structural differences in the instrumental learning tasks might contribute to the discrepancy between the two datasets of open data in the influence of choice perseverance. In Niv et al. (2012), the existence of forced choices might have weakened the effect of choice history (Alós-Ferrer et al., 2016). Furthermore, the existence of certain options that vary the risk level between the options might lead to asymmetric value updating. Future studies should investigate the psychological source of asymmetric learning rates.

In conclusion, I demonstrate the utility of the Hybrid model with multiple computational components in dissociating the cognitive process underlying human choice behavior. The proposed model used in this study contributes to a deeper understanding of the neural mechanisms of and individual differences in these cognitive components in instrumental learning.

Chapter 3 Pursuit of overtly unprofitable targets: computational substrates and its psychological effects

3.1 Introduction

Repetitive choice behaviors are induced by either extrinsic or intrinsic information resulted from own choices (i.e., chosen outcomes and choice *per se*). Through Chapter 2, I developed the hybrid model incorporating these information processing based on the reinforcement learning model, and demonstrated that this hybrid model could dissociate the impact of each information in human choice behaviors. Thus, in Chapter 3, I aimed to investigate the computational process underlying the pursuit of the hard-to-get target by using the hybrid model developed in Chapter 2.

Life is a series of choices. For example, we are often faced with the selection of a partner in the real world. Although selecting a person with whom one could build a good relationship is critical to enriching one's life, humans sometimes direct their unrequired passion toward the person who hardly responds in a positive manner.

Reinforcement learning is the predominant framework to account for choice behavior in organisms (Doya, 2007; Thorndike, 1911). From a traditional reinforcement learning perspective, choice behaviors depend on previously obtained outcomes (Daw & Tobler, 2013; Sutton & Barto., 1998). According to this outcome-dependent process, the option that is never reinforced is rarely chosen. Thus, it is difficult to explain pursuit of hard-to-get targets by reinforcement learning. From the computational perspective of reinforcement learning, previous studies have reported that asymmetric value updating (Frank et al., 2007; Samuel J Gershman, 2015; Lefebvre et al., 2017; Niv et al., 2012; Palminteri, Lefebvre, et al., 2017) and choice perseverance (Akaishi et al., 2014; Alós-Ferrer & Shi, 2015; Erev et al., 2013; Lau & Glimcher, 2005; Schönberg et al., 2007) lead to repetitive choices. Asymmetric value updating is able to facilitate the impact of positive outcomes and to inhibit the impact of negative outcomes, subsequently leading to repeating the previous choice. On the other hand, choice perseverance leads us to repeat the past choice independent of past outcomes. It is possible that the reinforcement learning model with asymmetric value updating or choice perseverance accounts for the pursuit of hard-to-get targets. The purpose of this study is to investigate whether the pursuit of a hard-to-get target is accounted for by asymmetric value updating, choice perseverance or both.

Preference is another important factor in decision-making. Most people might believe that the pursued target is the most preferred for the decision-maker. In general decision theory, preferences are thought to be stable over time (Glimcher, 2009), and we can infer others' preferences from their choices. However, many studies have reported that the choice *per se* increases the preference for the chosen target (Ariely & Norton, 2008; Brehm, 1956; Cockburn et al., 2014; Hornsby & Love, 2020; Izuma & Murayama, 2013; Koster et al., 2015; Nakao et al., 2016; Schonberg et al., 2014; Sharot et al., 2009). Through this choice-induced reevaluation, the chosen target becomes more preferred, which will often lead to choosing the same option again. Therefore, it is reasonable that even when the target is not the most attractive at a baseline, if the target is continuously chosen, the target comes to be recognized as more attractive.

Based on the above evidence, I hypothesized that cognitive computation processes, such as asymmetric value updating or choice perseverance, lead us to repeatedly choose the hard-to-get target, subsequently increasing the attractiveness of the selected target. Here, to investigate this hypothesis, an avatar choice task that mimicked partner selection was used in a web-based experiment. To control the baseline attractiveness of the avatars presented in the choice task, I selected avatars based on preference ratings before the choice task. Additionally, by manipulating outcome probabilities, I established hard-to-get and easy-to-get avatars. Subjects rated the attractiveness of the avatars again after the choice task to examine whether attractiveness was altered via the choice task. As a consequence, I found that subjects with higher choice perseverance pursued the hard-to-get avatar which rarely provided positive reactions, resulting in the increased attractiveness of the hard-to-get avatar.

3.2 Methods

3.2.1 Subjects

One hundred fifty subjects were recruited via CrowdWorks (https://crowdworks.jp/). Due to the nature of our task, I only recruited subjects who were at least 18 years old and were romantically interested in females. The study was approved by the ethical research committee at Nagoya University, and the study was carried out in accordance with the relevant guidelines and regulations.

3.2.2 Web-based experimental procedure

Informed consent was obtained from all subjects by clicking 'I Agree' after reading the information about the aims and procedures of this study. After they completed the survey for basic demographic information, including gender and age, they downloaded the Inquisit player (Millisecond Software LLC, Seattle, USA) and started a series of behavioral tasks (see the details below). To protect subjects' privacy, all data were anonymized. If the subject completed the entire task and survey without interruption, I paid 550 yen (approximately $5).

Two subjects were excluded from the following analyses because they omitted more than 30% of choice trials. Thus, the data from 148 subjects (age: range = 18-65 years, M = 38.07, SD = 11.03) were analyzed.

3.2.3 Behavioral tasks

In the web-based experiment, the subjects performed two tasks: an avatar evaluation task and an avatar choice task. First, to investigate the baseline attractiveness of 48 avatars, the subjects performed the avatar evaluation task. In this task (**Figure 3.1a**), three pictures of an avatar with different facial expressions (positive, neutral, and negative expressions) were displayed on the computer screen. They were asked to rate the subjective attractiveness of the presented avatar on a 9-point scale (1: not at all attractive, 9: very attractive) by pushing the numeric keys on their PCs. The order of presentation of the avatars was randomized across subjects.

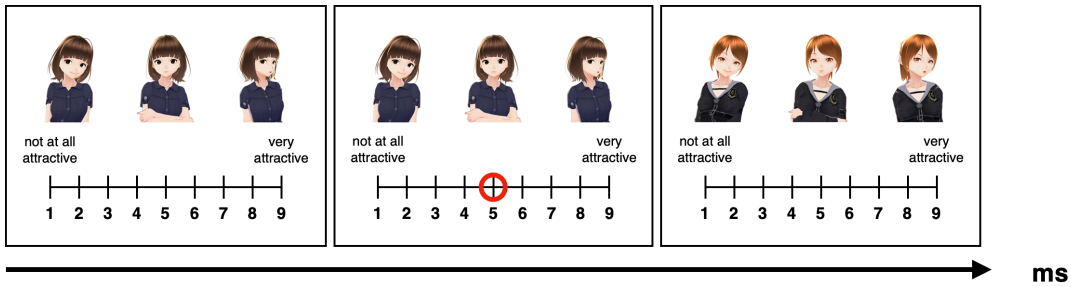
After the avatar evaluation task was completed, subjects performed the avatar choice task. To minimize the difference in baseline attractiveness of the avatars used in the avatar choice task as much as possible, eight avatars were selected based on the attractiveness rated in the preceding avatar evaluation task in the following manner. (i) The avatars rated as 6 or 7 points were selected. If the number of avatars rated as 6 or 7 points was less than eight, (ii) among the avatars rated less than 5 points, the avatar with the highest point was selected. If the total number of selected avatars was still less than eight, (iii) among the avatars rated more than 8 points, the avatar with the lowest point was selected. Then, (ii) and (iii) were repeated in a sequence until eight avatars were selected.

The avatar choice task consisted of two sessions. Four pairs were made from the eight selected avatars, and then two pairs (pair A and B) were used in each session. In each trial (**Figure 3.1b**), subjects were required to choose one of two avatars presented on the screen for 3,000 ms.

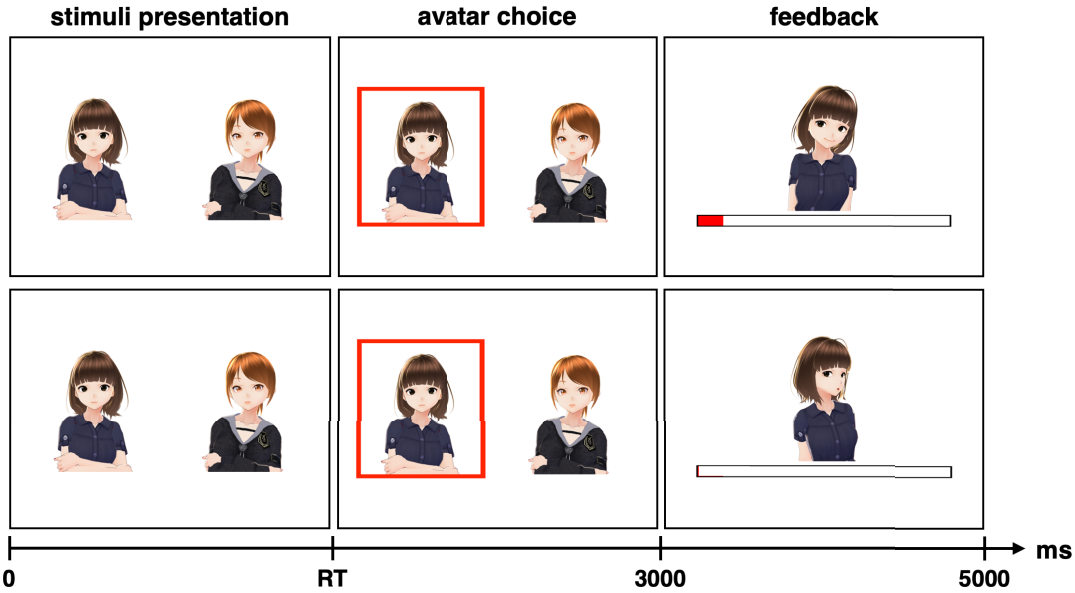
The presented position of the avatars was randomized across trials. After the subject selected one avatar, the selected avatar was highlighted with a red frame until the 3,000 ms had elapsed. Then, the visual and auditory stimuli associated with the reaction of the selected avatar was presented 2,000 ms. Specifically, the positive reaction was a smiling facial expression and a happy voice, while the negative reaction was a disappointed facial expression and a boring voice. At the display of the reaction, a horizontal bar, which represented the accumulated likability from the avatar to the subject, was presented below the avatar. The bar increased when the avatar expressed a positive reaction, while the bar did not change when the avatar expressed a negative reaction. The subjects were asked to maximize the likability from avatars throughout the task. In the first trial of each avatar pair, the reaction of the selected avatar was always negative. In the following trials, the ratio of positive and negative reactions was determined by the first choice to minimize the influence of first impressions (Shteingart et al., 2013). For pairs A and B, the reaction probability (positive/negative) of the initially chosen avatar was set at 0.1/0.9 and 0.9/0.1, respectively. Based on this probability, I referred to the initially chosen avatar in pair A as the “difficult” avatar and the initially chosen avatar in pair B as the “easy” avatar. For the unchosen avatars of both pairs (called "$neutral_{diff}$" and “$neutral_{easy}$”) in the first trial, the reaction probability was set at 0.5/0.5. These probabilities were fixed across the task so that subjects were required to learn the probability of each avatar (**Figure 3.1c**). In each session, subjects completed 80 trials (40 trials/pair). The order of pairs was randomized for each subject and session.

To investigate whether the attractiveness of the avatars was altered after the choice task, the subjects again rated the attractiveness of the 48 avatars presented at the initial evaluation task after they had completed the avatar choice task.

(a) Avatar evaluation task



(b) Avatar choice task



(c) Determination of reaction probability

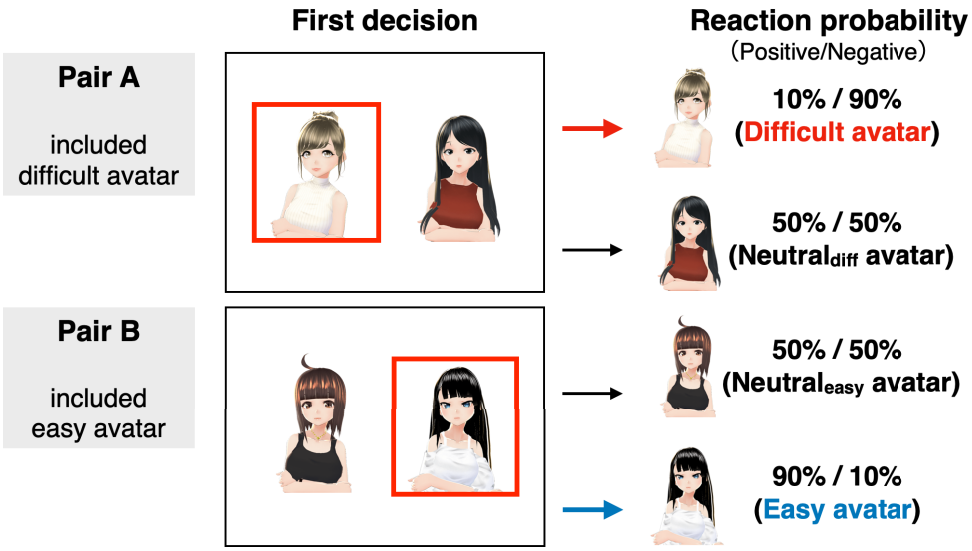


Figure 3.1 Behavioral task. (**a**) Avatar evaluation task. Subjects were asked to rate the attractiveness of 48 avatars on a 9-point scale. (**b**) Avatar choice task. This task required participants to choose one of the two avatars displayed on the screen and to maximize the likeability from the avatar represented as a length of red bar. After they chose an avatar (RT is response time), the reaction from the chosen avatar was displayed. Positive reactions increase the likeability, while negative reactions did not alter the likeability. (**c**) The outcome probability for each avatar was determined based on the first choice in the avatar choice task. In pair A, the initially chosen avatar was rarely associated with positive reactions in the subsequent trials (positive/negative = 0.1/0.9; i.e., difficult avatar). On the other hand, the initially chosen avatar in pair B was frequently associated with positive reactions in the subsequent trials (positive/negative = 0.9/0.1; i.e., easy avatar).

3.2.4 Behavioral analyses

I calculated the CP of avatars presented in the avatar choice task by dividing the number of choices by the number of trials (40 per avatar). Based on the CP of the difficult avatar (CP_{diff}), the subjects were divided into two groups: The Pursuit group (CP_{diff} was more than 0.5) and the No-pursuit group (CP_{diff} was less than 0.5). To confirm the difference in CP of the difficult and easy avatars between groups, two-way ANOVA with Group (Pursuit vs No-pursuit) and Avatar (difficult vs easy) was conducted.

To examine whether the baseline attractiveness differed among the avatars, including difficult, easy, $neutral_{diff}$, $neutral_{easy}$, and the unused avatars that were not presented in the choice task in the two groups, two-way mixed-design ANOVA with Group (Pursuit vs No-pursuit) and Avatars (difficult, easy, $neutral_{diff}$, $neutral_{easy}$, and unused) was performed. Additionally, to investigate whether attractiveness changed after the avatar choice task, I calculated the difference in attractiveness before and after the avatar choice task for the avatars. Then, two-way mixed-design ANOVA with Group (Pursuit vs No-pursuit) and Avatars (difficult, easy, $neutral_{diff}$, $neutral_{easy}$, and unused) was performed. To examine whether the attractiveness of each avatar was changed, the degree of changes in attractiveness was compared with zero by using one-sample *t* tests. The issue of multiple comparisons for one-sample *t* tests was corrected with Bonferroni's method. Moreover, to examine whether the choice *per se* increased the attractiveness of the chosen avatar, a general linear model (GLM) analysis was performed. In this model, the change in attractiveness was a dependent variable. The changes in attractiveness of the avatars used in the choice task were pooled across all subjects. The number of choices, the number of positive reactions, and an interaction were independent variables.

All analyses were executed by using R version 4.0.2 statistical software (http://cran.us.r-project.org). Post hoc pairwise comparisons for significant effects were conducted based on Shaffer's correction for multiple comparisons. The statistical threshold for significance was set at 0.05 for all behavioral analyses.

3.2.5 Models

In Chapter3, I mainly used three types of reinforcement learning models: (i) asymmetric, (ii) perseveration, and (iii) hybrid model. Although the models described below are identical with those used for the factual learning condition in Chapter 2, I redescribed the models to promote understanding. All models were modified based on a typical Q-learning model (called the "RL model"):

$$\delta(t) = R(t) - Q_i(t) \tag{3.1}$$

$$Q_i(t+1) = Q_i(t) + \alpha\delta(t) \tag{3.2}$$

Throughout this chapter, I basically consider cases with only two options (i = 1 or 2). The model assigns each option i an expected outcome $Q_i(t)$, where t is the index of the trial. The initial Q-values are set to zero (i.e., $Q_1(1) = Q_2(1) = 0$). The model updates the Q-values depending on the outcome of the choice (i.e., the reaction of the chosen avatar). The actual outcome at trial t is denoted by $R(t)$. I typically consider a binary outcome case whereby I set $R(t) = 1$ if a positive reaction is given and $R(t) = 0$ if a negative reaction is given. The learning rate α determines how much the model updates the action value depending on the reward prediction error, $\delta(t)$. Here, I denote the option that is chosen at trial t by $act(t)$ (= 1 or 2). Based on the set of Q-values, the model assigns the probability of choosing option 1 using the soft max function:

$$P(act(t) = 1) = \frac{1}{1 + \exp(-\beta[Q_1(t) - Q_2(t)])} \tag{3.3}$$

where β is called the inverse temperature parameter, which determines the sensitivity of the choice probabilities to differences in Q-values.

Based on the RL model, the asymmetric model assumes two independent learning rates:

$$Q_i(t+1) = \begin{cases} Q_i(t) + \alpha^+\delta(t) \text{ if } \delta(t) \geq 0 \\ Q_i(t) + \alpha^-\delta(t) \text{ if } \delta(t) < 0 \end{cases} \tag{3.4}$$

where α^+ adjusts the amplitude of value changes from one trial to the next when prediction errors are positive (when the actual reward $R(t)$ is better than the expected outcome $Q(t)$); the changes with α^- are vice versa.

The perseveration model is also based on the RL model and adds the computational process of choice history independent of the outcome-based learning process (Akaishi et al., 2014; S. J. Gershman et al., 2009; Schönberg et al., 2007):

$$C_i(t+1) = C_i(t) + \tau(I(act(t) = i) - C_i(t)) \tag{3.5}$$

The choice trace $C_i(t)$ is defined to introduce the effect of past choice into the CP. The initial values of $C_i(t)$ are set to zero (i.e., $C_1(1) = C_2(1) = 0$). The indicator function $I(\cdot)$ takes on a value of 1 if the statement is true and 0 if the statement is false. The decay rate τ is a free parameter that determines the number of preceding choices in the choice history influencing the current choice. When the choice is binary, the probability of choosing option 1 is implemented by the following:

$$P(act(t) = 1) = \frac{1}{1 + \exp(-\beta[Q_1(t) - Q_2(t)]) - \varphi[C_1(t) - C_2(t)])} \tag{3.6}$$

where the weight of choice history (φ) is a parameter that controls the tendency to repeat previous choices or avoid previously chosen options. A high positive value of this parameter indicates that the agent frequently repeats the previous choice.

Finally, the hybrid model has features of both the asymmetric and perseveration models. This model incorporates not only asymmetric learning rates but also choice traces (equations 3.4, 3.5, and 3.6). A previous study demonstrated that this hybrid model allows us to separately evaluate asymmetric learning rates and choice perseverance (Katahira, 2018; Sugawara & Katahira, *accepted*).

3.2.6 Simulation

To investigate what computational process contributes to pursuing the difficult avatar, I simulated the choices from the agents with the hybrid model. In particular, I systematically varied the free parameters of the hybrid model and evaluated CP_{diff} and CP_{easy} based on the

simulated choice data. The task structure used in a simulation was identical to that in the web-based experiment.

The hybrid model has five parameters: learning rates for positive and negative reward prediction error (α^+, α^-), inverse temperature (β), decay rate (τ), and weight of choice history (φ). Because I was interested in the degree of asymmetric learning rates, the difference in learning rates ($\alpha_{bias} = \alpha^+ - \alpha^-$) was calculated as the learning rate bias. In case 1, to examine the parameters related to the impact of past outcomes, the learning rate bias ($-1 \leqq \alpha_{bias} \leqq 1$, interval = 0.1) and inverse temperature ($0 \leqq \beta \leqq 10$, interval = 1) were varied, but the decay rate ($\tau = 0.5$) and the weight of choice history ($\varphi = 1$) were fixed. In case 2, to examine the parameters related to the impact of past choice, the decay rate ($0 \leqq \tau \leqq 1$, interval = 0.1) and the weight of choice history ($0 \leqq \varphi \leqq 10$, interval = 1) were varied, but the learning rate bias ($\alpha_{bias} = 0$) and inverse temperature ($\beta = 2$) were fixed. I hypothesized that the increased CP_{diff} was accounted for by the higher choice perseverance, which was represented as the greater weight of choice history. Thus, I further examined the interaction between the weight of choice history and parameters related to the impact of past outcome on the CP. In case 3, the learning rate bias ($-1 \leqq \alpha_{bias} \leqq 1$, interval = 0.1) and the weight of choice history ($0 \leqq \varphi \leqq 10$, interval = 1) were varied, while the inverse temperature ($\beta = 2$) and the decay rate ($\tau = 0.5$) were fixed. In case 4, the inverse temperature ($0 \leqq \beta \leqq 10$, interval = 1) and the weight of choice history ($0 \leqq \varphi \leqq 10$, interval = 1) were varied, while the learning rate bias ($\alpha_{bias} = 0$) and the decay rate ($\tau = 0.5$) were fixed. In the simulation, 100 virtual datasets were simulated for each parameter setting.

3.2.7 Parameter estimation and model selection procedure

I fitted the four models mentioned above (i.e., RL, asymmetric, perseveration, and hybrid models) to the choice data derived from the avatar choice task. The RL model was included as a benchmark for model estimation. Using the R function "solnp" in the Rsolnp package (Ghalanos & Maintainer, 2015), I fit the parameters of each model with the maximum a posteriori (MAP) estimation and calculated the log marginal likelihood for each model using the Laplace approximation (Daw, 2011; Katahira, 2016). If all the models have equal prior probability, because the marginal likelihood is proportional to the posterior probability of the model, the model resulting in the highest marginal likelihood is the most likely one given a data set. Note that this study used the negative log marginal likelihood (i.e., lower values indicate a better fit). The prior distributions and constraints were set following previous studies (Palminteri, Lefebvre, et al., 2017;Sugawara & Katahira, *accepted*). All the learning rates were constrained to the range of $0 \leq \alpha \leq 1$ with a *Beta* (1.1, 1.1) prior distribution. The inverse temperature was constrained to

the range of $\beta \geq 0$ with a *Gamma* (shape = 1.2, scale = 5.0) distribution. In the perseverance model, the decay rate was constrained to the range of $0 \leq \tau \leq 1$ with a *Beta* (1, 1) distribution (i.e., a uniform distribution), and the choice trace weight was constrained to the range of $-10 \leq \varphi \leq 10$ with a *Norm* ($\mu = 0$, $\sigma^2 = 5$) distribution.

For the model comparisons, two-way mixed-design ANOVA with Group (Pursuit and No-pursuit) and Model (RL, asymmetric, perseveration, and hybrid) was conducted to compare the log marginal likelihoods. Additionally, I compared the estimated model parameters. For the learning rates (α_c^+, α_c^-), two-way mixed-design ANOVA with Group and Valence was performed. To correct for the violation of the sphericity assumption, Greenhouse-Geiser's adjustment of the degrees of freedom was used for the within-subject factor when appropriate. Post hoc pairwise comparisons were performed based on Shaffer's correction for multiple comparisons. For the bias of learning rates, inverse temperature, decay rate, and weight of choice history, the group difference was evaluated by using a two-sample *t* test. All analyses were executed by using R version 4.0.2 statistical software (http://cran.us.r-project.org). The statistical threshold for significance was set at 0.05 for all behavioral analyses.

3.3 Results

3.3.1 The results of behavior and subjective evaluation

Choice probability in the avatar choice task

To characterize subjects who pursued the difficult avatar despite the frequent negative reactions, I focused on the choice probability for the difficult avatar (CP_{diff}) in the avatar choice task. Sixty-eight of 148 subjects showed that CP_{diff} was greater than 0.5 (i.e., they chose the difficult avatar in more than half of the trials). Thus, I divided subjects into two different groups based on CP_{diff}. The subjects with a CP_{diff} value greater than 0.5 were assigned to the Pursuit group ($n = 68$), while the subjects with a CP_{diff} value lower than 0.5 were assigned to the No-pursuit group ($n = 80$). **Figure 3.2** shows the averaged CP for the difficult and easy avatars in the Pursuit and No-pursuit groups. The group difference in CP was significantly different between difficult and easy avatars (two-way mixed-design analysis of variance (ANOVA); Group×Avatar interaction: $F(1,146) = 144.29$, $p < .001$). CP_{diff} was significantly higher in the Pursuit group than in the No-pursuit group (post hoc pairwise comparison; $p < .001$), whereas CP_{easy} was comparable between groups ($p = .05$). These results confirmed that subjects in the Pursuit group behaved

differently only toward the difficult avatar, which was a pattern that was not observed in the subjects in the No-pursuit group.

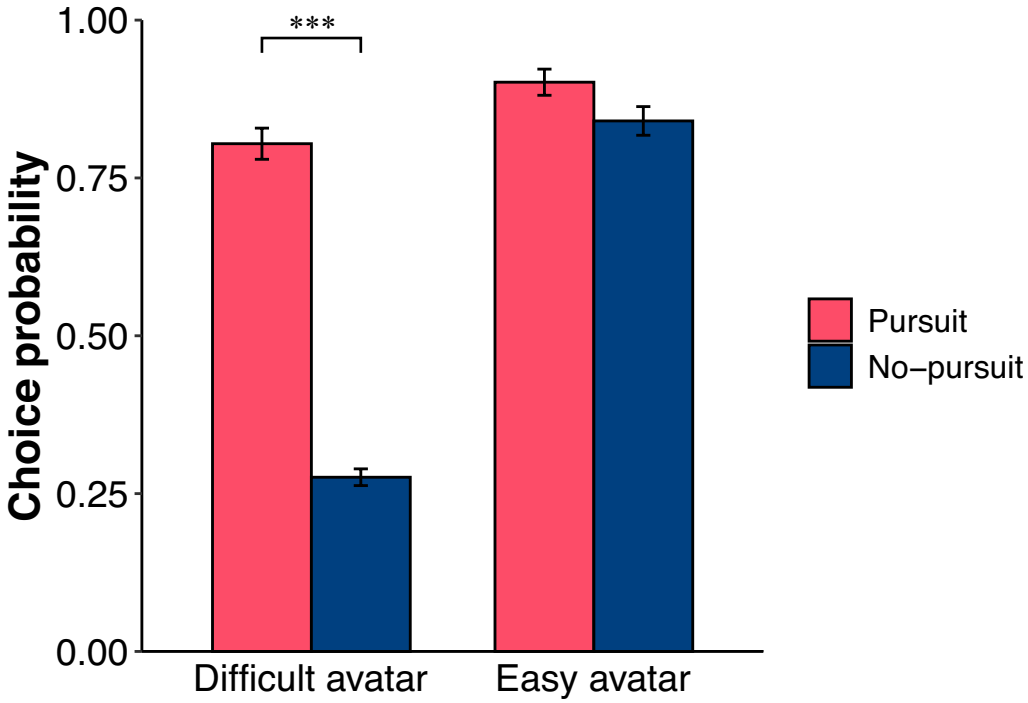


Figure 3.2 Choice probability for difficult and easy avatars in the two groups. In this study, subjects were assigned two different groups based on the choice probability for the difficult avatar. Error bars represent the standard error of the mean. Asterisk denotes a significant group difference: ***$p < .001$ (Shaffer's corrected).

Attractiveness of avatars before and after the avatar choice task

If the baseline attractiveness was higher for the initially chosen avatar than for the unchosen avatar, this difference in baseline attractiveness might have affected whether or not the subject pursued the difficult avatar. However, our results indicated that the attractiveness of the avatars used in the choice task was not different between groups and that the paired avatars were rated at the same level of attractiveness in both groups. I compared the attractiveness rated before the choice task between avatars and groups. The baseline attractiveness in any type of avatar (i.e., difficult, easy, $\text{neutral}_{\text{diff}}$, $\text{neutral}_{\text{easy}}$, and unused avatars) was not significantly different between groups (two-way mixed-design ANOVA; Group×Avatar interaction: $F(1.4, 204.05) = .32$, $p = .64$; main effect of Group: $F(1,146) = .28$, $p = .60$). On the other hand, the baseline attractiveness

was significantly different among avatars (main effect of Avatar: $F(1.4, 204.05) = 82.20$, $p < .001$). The avatars used in the choice task were rated as more attractive than the unused avatars (post hoc pairwise comparisons; *ps* < .001; Shaffer corrected). In addition, the avatars in pair B, including easy and $neutral_{easy}$ avatars, had significantly higher attractiveness scores than the avatars in pair A, including difficult and $neutral_{diff}$ avatars (post hoc pairwise comparisons; *ps* < .022; Shaffer corrected), with the exception of the comparison between difficult and $neutral_{easy}$ avatars ($p = .65$). However, the paired avatars had comparable attractiveness in both groups (*ps* > .11).

I investigated how the attractiveness of the avatars changed through the choice task. The change in avatar attractiveness was calculated by subtracting the score before the choice task from the score after the choice task (**Figure 3.3**). The interaction between groups and the types of avatar was significant (Group×Avatar interaction: $F(3.52, 513.71) = 14.62$, $p < .001$). The attractiveness of the unused avatars was not changed after the choice task in either group (one-sample t test, Pursuit: $t(67) = -.50$, $p > .99$, No-pursuit: $t(79) = .48$, $p > .99$; simple main effect of Group: $F(1,146) = .44$, $p = .51$). The attractiveness of the easy and $neutral_{easy}$ avatars did not differ between groups (simple main effect of Group; easy: $F(1,146) = .016$, $p = .90$, $neutral_{easy}$: $F(1,146) = .71$, $p = .40$). In both groups, the easy avatar was rated as more attractive (one-sample *t* test, Pursuit: $t(67) = 5.35$, $p < .001$; No-pursuit: $t(79) = 5.31$, $p < .001$), while the $neutral_{easy}$ avatar was rated as less attractive after the choice task (one-sample *t* test, Pursuit: $t(67) = -4.03$, $p = 1.45\times10^{-3}$; No-pursuit: $t(79) = -4.08$, $p = 1.07\times10^{-3}$). On the other hand, the attractiveness of the difficult avatar increased in the Pursuit group (one-sample *t* test, $t(67) = 3.02$, $p = 3.59\times10^{-2}$), while it did not change in the No-pursuit group (one-sample *t* test, $t(79) = -1.82$, $p = .72$; simple main effect of Group; $F(1,146) = 12.33$, $p < .001$). In contrast, the attractiveness of the $neutral_{diff}$ avatar, which was paired with the difficult avatar, decreased in the Pursuit group (one-sample *t* test, $t(67) = -4.65$, $p < .001$) but did not change in the No-pursuit group (one-sample *t* test, $t(79) = 2.35$, $p = .21$; simple main effect of Group; $F(1,146) = 27.55$, $p < .001$). These results indicated that both difficult and easy avatars were more attractive after the choice task in the Pursuit group, while only easy avatars were more attractive in the No-pursuit group.

The increased attractiveness of the difficult and easy avatars in the Pursuit group raised the question of what events occurred in the choice task to increase the attractiveness of avatars. To answer this question, I conducted a linear mixed-effect model analysis with the number of choices and positive reactions as independent variables and the changes in attractiveness as the dependent variable (see Methods section). The number of choices had a significant effect only on the change in attractiveness observed after the choice task ($F(1,663.85) = 31.88$, $p < .001$). On

the other hand, the main effect of the number of positive reactions ($F(1,1020.23) = .080$, $p = .78$) and the interaction were not significant ($F(1,1015.09) = 3.00\times10^{-3}$, $p = .99$). Thus, the changes in attractiveness depended on the choice *per se* rather than reactions in the choice task.

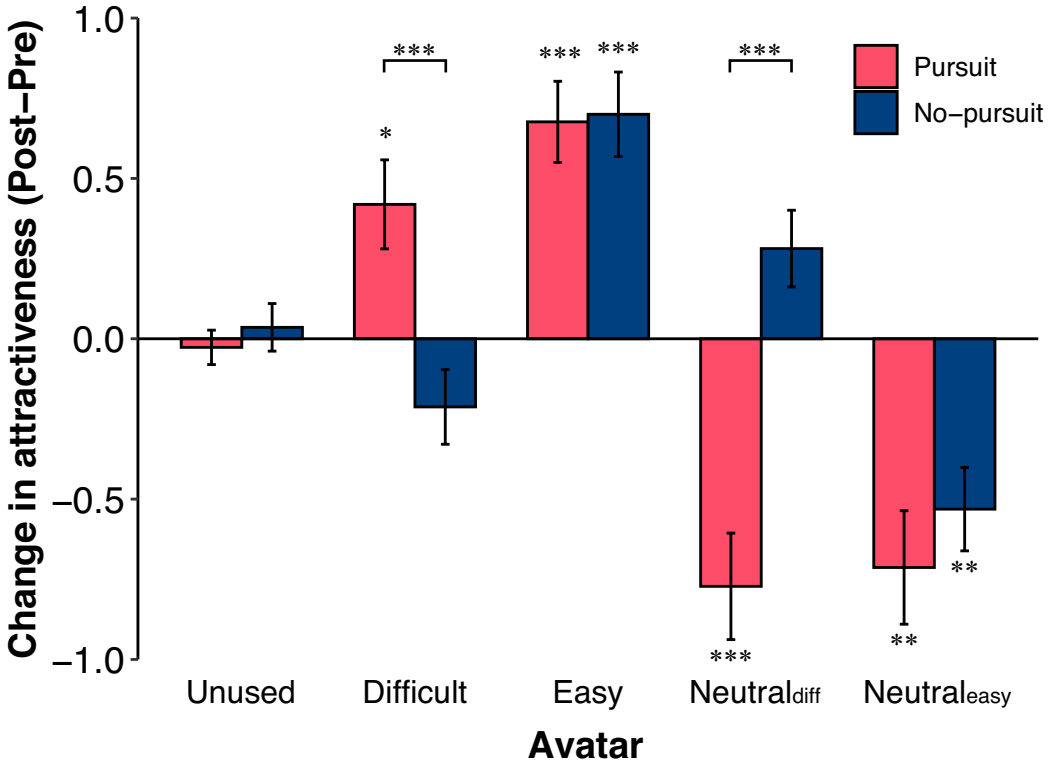


Figure 3.3 Changes in attractiveness ratings after the avatar choice task. The figure shows the changes in attractiveness ratings of the five types of avatar in the two groups. Changes in attractiveness ratings were calculated by subtracting the score at the prechoice rating from that at the postchoice rating. The unused avatars were not used in the avatar choice task (i.e., 40 avatars). The other types of avatar (i.e., difficult, easy, $\text{neutral}_{\text{diff}}$, and $\text{neutral}_{\text{easy}}$) were used in the avatar choice task. The error bars represent the standard error of the mean. Asterisks denote significance levels: ***$p < .001$, **$p < .01$, and *$p < .05$.

3.3.2 Simulation

I found that some subjects (i.e., the Pursuit group) pursued the difficult avatar despite very few positive reactions (**Figure 3.2**). This behavioral phenomenon raised the question of what cognitive process makes these subjects pursue the difficult avatar. To answer this question, I used several variants of reinforcement learning models to determine what accounted for this choice

behavior. Previous studies have reported that asymmetric value updating (Lefebvre et al., 2017; Palminteri, Lefebvre, et al., 2017) and choice perseverance (Akaishi et al., 2014) can lead to repetitive choice of a previously selected option. I previously demonstrated that a hybrid model allows us to distinguish between asymmetric value updating and choice perseverance (Sugawara & Katahira, *accepted*). Thus, I conducted a simulation to investigate what parameters implemented in the hybrid model could account for the behavioral pattern shown in the Pursuit group. In particular, the hybrid model has five free parameters: learning rates for positive and negative reward prediction errors (α^+ and α^-), inverse temperature (β), decay rate of choice history (τ), and choice trace weight (φ) (see Methods section). The degree of asymmetric value updating is denoted by the difference in the two learning rates (i.e., $\alpha_{bias} = \alpha^+ - \alpha^-$). Thus, I simulated the agent's choice behavior by manipulating these four parameters (α_{bias}, β, τ, and φ) under the same task structure as the web-based experiment (see Methods section).

In case 1, α_{bias} and β were varied, while τ (= 0.5) and φ (= 1.0) were fixed. The asymmetric learning rates quadratically decreased CP_{diff} (**Figure 3.4a**) but quadratically increased CP_{easy} (**Figure 3.4b**). Moderate positivity bias ($\alpha_{bias} = 0.4$) induced the least CP_{diff}, while moderate negativity ($\alpha_{bias} = -0.6$) bias induced the most CP_{easy}. The inverse temperature produced a linear decrease in CP_{diff} and a linear increase in CP_{easy}. In any combination, CP_{diff} was less than 0.5, indicating that these parameters did not account for the behavioral pattern observed in the Pursuit group ($CP_{diff} > 0.5$).

In case 2, τ and φ were varied, while α_{bias} (= 0) and β (= 2.0) were fixed. For the difficult avatar, CP_{diff} values in the condition with moderate decay rate ($\tau > 0.2$) and higher perseverance factor ($\varphi > 6.0$) reached over 0.7 (**Figure 3.4c**). Meanwhile, CP_{easy} did not depend on these parameters and was always over 0.7 (**Figure 3.4d**). In the higher perseverance condition, the behavioral pattern was similar to the Pursuit group in the web-based experiment.

To further examine whether the effect of perseverance trades off with the effect of the value-related parameters (i.e., α_{bias} and β), I covaried either α_{bias} (case 3) or β (case 4) with φ. In case 3, although CP_{diff} was modulated by the asymmetric learning rates (α_{bias}) in the condition with lower perseverance ($\varphi < 6.0$), the condition with higher perseverance ($\varphi > 6.0$) showed higher CP_{diff} (**Figure 3.4e**) and CP_{easy} (**Figure 3.4f**). Likewise, in case 4, in the condition with higher perseverance ($\varphi > 6.0$), CP_{diff} (**Figure 3.4g**) and CP_{easy} (**Figure 3.4h**) were not affected by inverse temperature and showed higher probability ($CP > 0.7$). Therefore, these results suggested that a higher perseveration generates the behavior pattern shown in the Pursuit group.

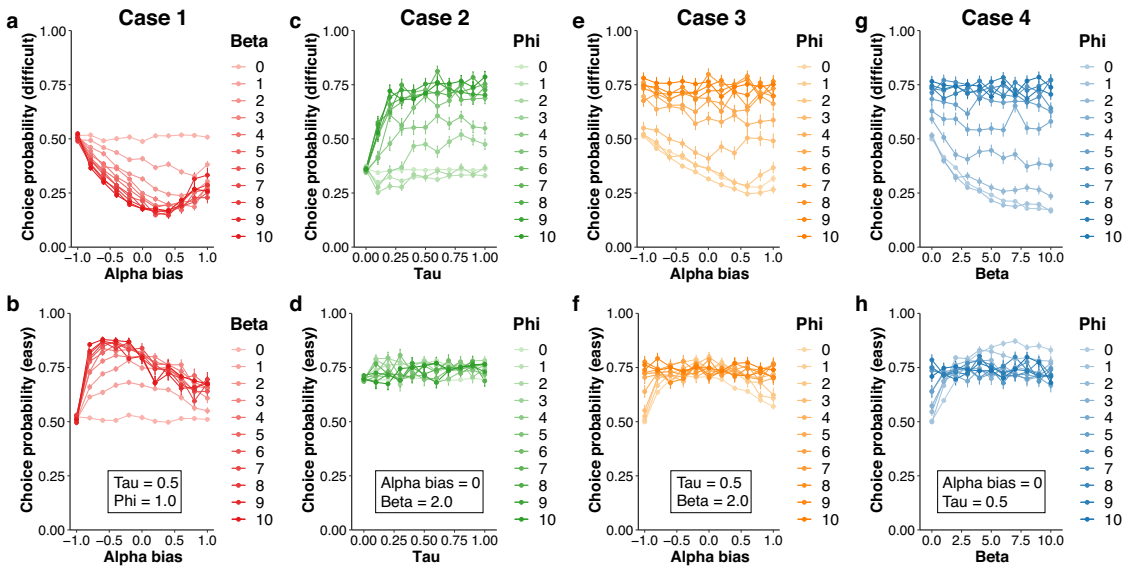


Figure 3.4 The results of the simulation in the hybrid model. The simulation of the agent's choice behavior was generated by manipulating four parameters (α_{bias}, β, τ, and φ) included in the hybrid model. The upper and lower rows show the choice probability for difficult and easy avatars, respectively. In case 1, the bias of learning rates ($\alpha_{bias} = \alpha^{+} - \alpha^{-}$) and the inverse temperature (β) were varied, while the decay rate ($\tau = 0.5$) and the weight of choice history ($\varphi = 1.0$) were fixed (**a**, **b**). In case 2, τ and φ were varied, while α_{bias} (= 0) and β (= 2.0) were fixed (**c**, **d**). In case 3, α_{bias} and φ were varied, while φ (= 1.0) and β (= 2.0) were fixed (**e**, **f**). In case 4, β and φ were varied, while α_{bias} (= 0) and τ (= 0.5) were fixed (**g**, **h**).

3.3.3 Model selection

By conducting a simulation, I indicated that choice perseverance could account for the increased choices for the difficult avatar with fewer positive reactions (**Figure 3.4**). Therefore, to investigate the hypothesis that the behavioral pattern shown in the Pursuit group was due to choice perseverance, I fitted computational models to the choice data derived from the web-based experiment. I used four variants of RL models to examine the benchmark of model fitting: (1) a standard Q-learning model (hereafter, the RL model), (2) the asymmetric model, (3) the perseveration model, and (4) the hybrid model (see Methods section). The results revealed that the perseveration model was the best for the Pursuit group, while the asymmetric model was the best for the No-pursuit group (**Table 3.1**). There were no differences among the models in the No-pursuit group ($F(1.05, 83.28) = .12$, $p = .75$), but there were differences in the Pursuit group ($F(1.57, 105.27) = 91.00$, $p < .001$). For the Pursuit group, there was no significant difference between the perseveration and hybrid models (post hoc comparison; $p = .64$; Shaffer corrected),

but the RL and asymmetric models, which did not include the choice history process, were much worse than the perseveration and hybrid models, which did include the choice history process (*ps* < .001).

Table 3.1 Models and model selection results

Model	Learning rate (s)	Inverse temperature	Perseveration	# of free parameters	Pursuit group LML (SD)	No-pursuit group LML (SD)
RL	α ($\alpha^+ = \alpha^-$)	β	−	2	-51.94 (27.52)	-53.83 (27.50)
Asymmetric	α^+, α^-	β	−	3	-47.09 (29.21)	-53.71 (27.47)
Perseveration	α ($\alpha^+ = \alpha^-$)	β	τ, φ	4	-37.89 (33.87)	-54.17 (27.23)
Hybrid	α^+, α^-	β	τ, φ	5	-38.23 (34.33)	-54.14 (27.29)

Furthermore, to examine whether the group difference of the underlying cognitive process was manifested in both pair A (including the difficult avatar) and B (including the easy avatar), I separated the choice data of pair A and B and then fitted four models into the separated datasets (**Table 3.2**). The results showed that there was a significant interaction between Group and Model in pair A ($F(1.58,230.41) = 41.58$, $p < .001$) but not in pair B ($F(1.51,220.14) = 1.19$, $p = .30$). Although the simple main effect of Model in pair A was significant in both groups (Pursuit: $F(1.42,94.92) = 44.90$, $p < .001$; No-pursuit group: $F(1.61,127.04) = 4.12$, $p = .026$), post hoc pairwise comparisons did not show any differences among models in the No-pursuit group (*ps* > .14). In contrast, there was a significant difference between all models in the Pursuit group (*ps* < .001) with the exception of the comparison between the perseveration and hybrid models ($p = .19$).

Table 3.2 Models and model selection results in each pair

Condition	Model	Learning rate (s)	Inverse temperature	Perseveration	# of free parameters	Pursuit group LML (SD)	No-pursuit group LML (SD)
Pair A	RL	α ($\alpha^+ = \alpha^-$)	β	−	2	-35.21 (12.23)	-34.37 (13.64)
includes DIFFICULT avatar	Asymmetric	α^+, α^-	β	−	3	-30.30 (15.06)	-34.19 (14.00)
	Perseveration	α ($\alpha^+ = \alpha^-$)	β	τ, φ	4	-22.30 (21.06)	-37.14 (15.91)
	Hybrid	α^+, α^-	β	τ, φ	5	-23.51 (23.17)	-36.56 (15.16)
Pair B	RL	α ($\alpha^+ = \alpha^-$)	β	−	2	-16.57 (16.54)	-19.96 (15.95)
includes EASY avatar	Asymmetric	α^+, α^-	β	−	3	-15.74 (16.80)	-19.54 (16.36)
	Perseveration	α ($\alpha^+ = \alpha^-$)	β	τ, φ	4	-17.33 (19.56)	-19.35 (16.48)
	Hybrid	α^+, α^-	β	τ, φ	5	-15.75 (19.61)	-19.50 (17.66)

These results indicated that choice behaviors in the Pursuit group depended on choice history, while choice behaviors in the No-pursuit group depended on past choice outcomes. Furthermore, this group difference in the impact of choice history was observed only in the specific context involving avatars with relatively few positive reactions.

3.3.4 Parameter estimation

To directly examine what computational process elicited the difference in choice behavior between the two groups, I compared the model parameters estimated from the hybrid model between groups. Although the hybrid model was not the best for the Pursuit and No-pursuit groups (**Table 3.1**), in Chapter 2, I proved that the hybrid model can distinguish between the elements and correctly derive the parameters, even if the true model only contains the element of either asymmetric value update or choice perseverance (Sugawara & Katahira, *accepted*).

The Pursuit group had a higher learning rate (α) than the No-pursuit group (**Figure 3.5a**; $F(1,146) = 16.46$, $p < .001$). Positive learning rates were higher than negative learning rates in both groups ($F(1,146) = 42.85$, $p < .001$). The interaction was not significant ($F(1,146) = 1.67$, $p = .20$). Furthermore, the difference between the positive learning rate minus the negative learning rate was calculated as the learning rate bias. There was no significant difference in the learning rate bias between groups (**Figure 3.5b**; $t(146) = -1.29$, $p = .20$). The inverse temperature (β) was significantly lower in the Pursuit group than in the No-pursuit group (**Figure 3.5c**; $t(146) = 7.45$, $p < .001$). While the decay rate (τ) was not significantly different between groups (**Figure 3.5d;** $t(146) = 1.28$, $p = 0.20$), the choice trace weight (φ) was significantly higher in the Pursuit group than in the No-pursuit group (**Figure 3.5e;** $t(146) = -8.48$, $p < .001$). These results indicated that the Pursuit group placed greater weight on past choice than the No-pursuit group, while past outcomes had a greater influence on choice in the No-pursuit group than in the Pursuit group.

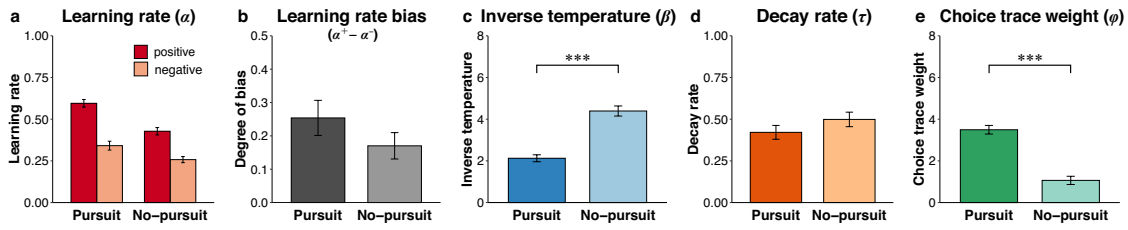


Figure 3.5 Estimated parameters with the hybrid model. The figure shows estimated parameters by fitting the hybrid model to the choice data derived from the web-based experiment. (**a**) The learning rates for the positive and negative reward prediction errors (α^+ and α^-). (**b**) The

learning rate bias calculated by subtracting the negative learning rate from the positive learning rate ($\alpha^{+} - \alpha^{-}$), indicating the degree of asymmetric value updating. (**c**) The inverse temperature (β) represents the sensitivity to value differences in decision-making. (**d**) The decay rate (τ) indicates how far past choices are incorporated into the next choice. The weight of choice history (φ) represents the sensitivity to differences of the choice history in the decision-making. Error bars represent the standard error of the mean. Asterisks denote significant group differences: ***p < .001 and **p < .01.

To examine whether this group difference in choice perseverance was observed in a specific context, I compared the model parameters in each separated dataset. Regarding the learning rates with both pair A (**Figure 3.6a**) and B (**Figure 3.6b**), the main effect of valence was significant (A: $F(1,146) = 78.36$, $p < .001$; B: $F(1,146) = 42.88$, $p < .001$), whereas the interaction was not significant (A: $F(1,146) = .06$, $p = .81$; B: $F(1,146) = 2.10$, $p = .15$). While the Pursuit group had a higher learning rate than the No-pursuit group with pair A ($F(1,146) = 8.84$, $p = 3.44\times10^{-3}$), no significant difference was shown with pair B ($F(1,146) = 2.93$, $p = .09$). The learning rate bias was not significantly different with either pair (**Figure 3.6c**; A: $t(146) = .24$, $p = .81$, **Figure 3.6d**; B: $t(146) = -1.45$, $p = .15$). The inverse temperature was significantly lower in the Pursuit group than in the No-pursuit group with both pairs (**Figure 3.6e**; A: $t(146) = 7.64$, $p < .001$, **Figure 3.6f**; B: $t(146) = 3.05$, $p = 2.69\times10^{-3}$). The decay rate was not significantly different between groups (**Figure 3.6g**; A: $t(146) = -.37$, $p = .71$, **Figure 3.6h**; B: $t(146) = -1.85$, $p = .07$) with both pairs. Importantly, while the choice trace weight was significantly higher in the Pursuit group than in the No-pursuit group with pair A (**Figure 3.6i**; $t(146) = -8.41$, $p < .001$), there was no significant difference with pair B (**Figure 3.6j**; $t(146) = -1.30$, $p = .20$). The increased weighting for past choices shown in the Pursuit group was noticeable only in the context that included the difficult avatar. The results suggested that increased weight for past choices (i.e., higher choice perseverance) causes the pursuit of the hard-to-get avatar.

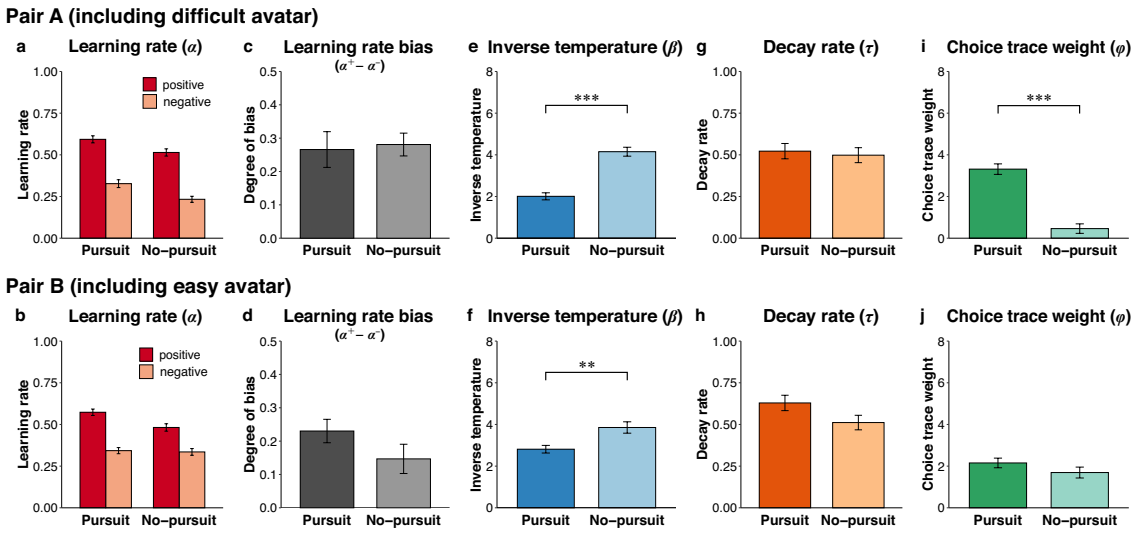


Figure 3.6 Estimated parameters with the hybrid model in each pair. The figure shows estimated parameters by fitting the hybrid model to the choice datasets separated by avatar pair. Upper and lower rows indicate estimated parameters for pair A (including the difficult avatar) and B (including the easy avatar), respectively. Error bars represent the standard error of the mean. Asterisks denote significant group differences: ***$p < .001$ and **$p < .01$.

3.4 Discussion

The present study investigated why people pursue hard-to-get targets. In the web-based experiment, the subjects performed the avatar evaluation and the avatar choice tasks. By manipulating outcome probabilities, I established the difficult avatar as one that rarely had positive reactions and the easy avatar as one that frequently had positive reactions. In most subjects, easy avatars that usually had positive reactions were more frequently chosen than paired avatars that had positive and negative reactions at the same frequency. Nevertheless, some of the subjects (i.e., the Pursuit group) frequently chose difficult avatars that seldom had positive reactions as well as easy avatars. Thus, I confirmed that some people pursue hard-to-get targets. The attractiveness of the avatars after the choice task was changed in accordance with the number of choices. Subsequently, following the choice task, the Pursuit group rated the difficult avatar as more attractive, while the No-pursuit group rated this avatar as less attractive. Then, I used a computational modeling approach to reveal the cognitive process mediating the pursuit of the difficult avatar. In a simulation, I demonstrated that a higher weight for choice history (i.e., choice perseverance) led to repetitive selection of not only the easy avatar but also the difficult avatar. To confirm this finding in the empirical data, I fitted the hybrid model proposed in a previous study (Sugawara & Katahira, *accepted*) to the choice data derived from the web-based experiment.

Consistent with the simulation results, the weight placed on choice history was significantly higher in the Pursuit group than in the No-pursuit group. According to these findings, I concluded that higher choice perseverance leads to repetitive choice of hard-to-get targets, consequently increasing the attractiveness of the selected target.

The pursuit of hard-to-get avatars shown in the Pursuit group was not explained by the traditional reinforcement learning theory that argues that the action probability is increased if the action is associated with positive outcomes (Sutton & Barto., 1998; Thorndike, 1911), despite the subjects in the current experiment having to maximize the likability from the avatar. Another possible explanation is that the Pursuit group preferred the hard-to-get avatar over the alternative avatar because the baseline preference influenced decision-making (Glimcher, 2009). However, in both groups, the baseline attractiveness did not differ between the paired avatars used in the avatar choice task. Thus, differences in baseline attractiveness did not account for the pursuit of hard-to-get avatars. The other possibility is that Pursuit group incidentally repeated to choose the hard-to-get avatar because they failed to update the values of its avatar from the obtained outcomes. However, the learning rates (α) in Pursuit group were significantly higher than those in Non-pursuit group, though the positivity bias was observed in both groups. These results might suggest that Pursuit group normally updates the values from the obtained outcomes. Why is the Pursuit group insensitive for the learned values? The answer for this question is the lower inverse temperature (β) in the Pursuit group compared to the Non-pursuit group. Higher inverse temperature makes us more sensitive to the difference in values for the options. Thus, even if they are able to learn the values from the obtained outcomes, their choices do not depend on the values when they are insensitive to the difference in values. Nevertheless, the only low inverse temperature is not enough to explain the repetitive choice behaviors. In the case of this study, the difference in values is obvious because the reward probabilities for difficult and easy avatars were extreme (i.e., 0.1 and 0.9, respectively). Considering that all participants showed learning rates over zero, it could be plausible that participants even in the Pursuit group were able to capture the difference in values. In addition, it is difficult to explain the repetitive choices for the hard-to-get avatar only by low inverse temperature. On the other hand, for the outcome-independent processing, our results indicated that the Pursuit group largely weighted the preceding choice history, representing the higher value of choice perseverance parameter (φ). If the choice perseverance parameter is higher, participants frequently repeat the preceding choice. In contrast, if this parameter is near zero, the participants' choice is independent of preceding choices. Choice perseverance reflects Thorndike's law of exercise stating that producing an action makes it more likely to be selected on future occasions (Thorndike, 1911). Although the law of exercise captures the key feature of habits in which behavioral repetition automatizes the behavior (Perez &

Dickinson, 2020), habituation is due to reward-based learning mechanisms (Miller et al., 2019). Because the hard-to-get avatar seldom made positive reactions, the pursuit of the hard-to-get avatar could not be accounted for by habituation. Unlike habituation, choice perseverance and the law of exercise are independent of choice outcomes. It is reasonable that the pursuit of the hard-to-get avatar is accounted for by choice perseverance.

Why were there two groups of subjects who did or did not pursue hard-to-get targets? To answer this individual difference, I should consider how choice perseverance emerges. Akaishi et al. (2014), which proposed the basis of the perseverance model used in this study, demonstrated that choice perseverance, which they called choice inertia, is accounted for by an autonomous learning mechanism of beliefs about the environmental state. Through this learning process, choice *per se* updates the choice likelihood estimate in each state. In the perseverance model, the weight of choice history (φ) determines how the choice likelihood estimate influences subsequent decisions independent of choice outcomes. According to this study (Akaishi et al., 2014), the weight of choice history reflected how the decision depended on one's own belief. Therefore, it could be speculated that subjects who pursued the hard-to-get target (i.e., Pursuit group) give weight to their own belief rather than the outcome history.

I also found that the group difference in the weight of choice history was observed only in the choice context including difficult avatars but not in the context including easy avatars. This finding suggested that the weight of one's own belief is modulated by the choice context even in people who pursued hard-to-get avatars. Akaishi et al. (2014) also showed that the tendency to repeat the same choice depends on ambiguity in the decision environment. That is, choices made in a more ambiguous state have more impact on subsequent trials. From this perspective (Akaishi et al., 2014), the weight of choice history might depend on the perceived ambiguity in each choice context. In the choice context including an easy avatar, the way to maximize positive reactions was obvious. On the other hand, I speculate that in the choice context including a difficult avatar, the perceived ambiguity varied across subjects because it is difficult to maximize positive reactions compared with the context including an easy avatar. Subjects who shift the behavioral goal to minimizing negative reactions might perceive the context as less ambiguous, while subjects who stick to positive reactions from previously selected avatars might perceive the context as highly ambiguous. Nevertheless, the psychological factors modulating choice perseverance remain unclear. To further understand the psychological mechanisms underlying the pursuit of hard-to-get targets, future studies should investigate such modulating factors in the decision-making used in the present study.

Our results showed that the increase in attractiveness depended on the number of choices rather than the number of positive reactions. This choice-dependent reevaluation has been reported (Brehm, 1956; Egan et al., 2007; Lieberman et al., 2001). Brehm (1956) reported that after the choice between two similar valued options, the selected option was evaluated as better than the unchosen option. Sharot et al. (2009) showed that hedonic-related neural activity in the caudate nucleus for the selected option was enhanced after a decision was made in a free choice task, suggesting that imagination during the decision process activates the hedonic-related brain region and conveys pleasure expected from the simulated event. This choice-induced reevaluation modifies the hedonic response to the selected option. From the view of imagination-related pleasure (Sharot et al., 2009), participants feel two types of pleasure in the avatar choice task used in this study: one induced by the imagination during the decision process and another induced by the obtained outcome. Subjects with higher choice perseverance focus on the decision process rather than the obtained outcome. Thus, it is possible that their preferences are more strongly affected by the pleasure from imagination during the decision process, which consequently increases attractiveness of the hard-to-get avatar. Moreover, choice-induced reevaluation was observed even in amnesic patients who did not remember the option they chose (Lieberman et al., 2001), younger children, and capuchin monkeys (Egan et al., 2007). According to this evidence, persons with higher choice perseverance repetitively select the hard-to-get avatar due to the weight placed on past choices, which subsequently increases the attractiveness of the selected avatar through automatic choice-induced reevaluation.

In the present study, I demonstrated that persons with higher choice perseverance pursue the target that rarely responded with positive reactions and rate the selected target as more attractive via the choice-induced reevaluation mechanism. In contrast, persons with less choice perseverance select the target depending on past positive reactions and rate the selected target as more attractive via the reinforcement learning mechanism.

Chapter 4 General discussion

4.1 Summary of present findings

The aim of this thesis was to elucidate the information processing underlying seemingly irrational pursuing behaviors by using the computational modeling which can dissociate the effects of extrinsic and intrinsic information in the reinforcement learning paradigm. To accomplish this aim, I firstly validated whether the hybrid computational model including both extrinsic and intrinsic information processing could identify the genuine process underlying the choice behaviors (Study 1; Sugawara & Katahira, 2019; Sugawara & Katahira, *accepted*). Subsequently, by using this hybrid model, I unveiled the cognitive process underlying the pursuit of the unprofitable target (Study 2; Sugawara & Katahira, *under revision*).

In Chapter 2, to validate whether computational modeling could dissociate the effects of "obtained outcomes" and "choice *per se*", I tested the usefulness of the hybrid model incorporating the asymmetric value updating and the choice perseverance that lead the repetitive choice behavior in the reinforcement learning context. Through a simulation, I demonstrated that the hybrid model is able to capture true model parameters even when the effects of the asymmetric value updating and the choice perseverance were mixed. By applying this hybrid model into actual choice behaviors collected from the web-based experiment, I empirically confirmed that the effect of choice perseverance was overlooked and was mistaken for the effect of the asymmetric value updating. Indeed, I re-analyzed the open data collected from previous study reporting the asymmetric value updating, and revealed that this open data was well accounted for by the choice perseverance rather than the asymmetric value updating. In parallel, another open data was accounted for by the asymmetric value updating rather than the choice perseverance. According to these findings, I validated the usefulness of the hybrid model to identify the genuine cognitive process underlying choice behaviors in the reinforcement learning context.

In Chapter 3, the hybrid model established in Study 1 was used to reveal the genuine cognitive process behind seemingly irrational behaviors such as the pursuit of the unprofitable target in spite of the lack of positive outcomes. Through the web-based experiment, I found that some subjects pursued the hard-to-get target that seldom returned positive reactions, while the other subjects did not choose such target. By using the hybrid model established in Study 1, I demonstrated that subjects who pursued the hard-to-get target (Pursuit group) make their choice depending on preceding choices (i.e., higher choice perseverance) compared to the subject who did not pursue such target (No-pursuit group). Furthermore, the attractiveness of the hard-to-get

target was increased in the Pursuit group after the choice task, indicating that the hard-to-get avatar becomes more attractive by the choice-dependent reevaluation. Taken together, I concluded that people with high choice perseverance pursue the hard-to-get target, making the target more attractive.

4.2 Psychological process mediating the choice perseverance

Choice perseverance is that the current choice *per se* influences subsequent choices. This psychological process is well documented in various terms: "the law of exercise" (Thorndike, 1911), choice bias (Akaishi et al., 2014; Urai et al., 2019), and choice perseverance (Katahira, 2018; Sugawara & Katahira, 2019; *accepted*). It is still unclear how choice perseverance influences subsequent choices. Such choice perseverance is observed not only in reinforcement learning context but not in perceptual decision-making (Akaishi et al., 2014; Bonaiuto et al., 2016; Urai et al., 2019). According to previous studies, choice perseverance is mediated by active cognitive processing (Akaishi et al., 2014; Urai et al., 2019) rather than by residual activity of preceding decision (Bonaiuto et al., 2016). By using the drift diffusion model (Ratcliff & McKoon, 2008), Urai et al. (2019) demonstrated that the speed of perceptual evidence accumulation (i.e., drift rate) is faster for the previously chosen stimulus than for the unchosen stimulus, suggesting that choice history might direct attentional resources toward the previously chosen interpretation of current sensory input. Likewise, Akaishi et al. (2014) formulated such decision bias as the simple updating process of choice history (see Equation 3.4). Afterward, Katahira, (2018) incorporated the updating process of choice history into the reinforcement learning model. Thus, the perseverance model used in this thesis is based on this updating process.

In perceptual decision-making, choice perseverance seems to bias the interpretation of sensory input. What is the interpretation of sensory input in the reinforcement learning context such as a two-armed bandit task? In Chapter 3, we found that the hard-to-get avatar was perceived as more attractive in subjects that pursued this avatar. This phenomenon is well known as choice-dependent re-evaluation in preference studies (Brehm, 1956; Egan et al., 2007; Lieberman et al., 2001). Although the mechanism of the choice-dependent reevaluation is still debating, Sharot et al. (2009) proposed that expectations that accompanied with choice aroused hedonic experiences, resulting in the choice-dependent reevaluation. During decision-making, we expect the desirable outcome resulted from the chosen option, and experience the pleasure from the expected outcomes. People enjoy the moments leading up to reward, that is, expectation makes us happy before the expected future comes true (Iigaya et al., 2020; Loewenstein, 1987). This hedonic experience induced by expectation might alter the affective evaluation for the options. The choice-

dependent reevaluation is observed even in amnesic patients (Lieberman et al., 2001), suggesting that this reevaluation is implicit. Thus, I hypothesize that the expected outcomes implicitly increase the value for the chosen option, leading the choice perseverance. Because this working hypothesis depends on the expected outcomes, the choice perseverance is not affected by the obtained outcomes. Based on this hypothesis, I propose a dual-updating model which has two types of value updating processes from actual outcomes and expected outcomes (Figure 4.1). In this model, I assume the value-updating process from the expected outcomes which is simulated during decision-making. Through two types of value-updating processes, agents represent two independent values: experience-dependent value (Q) and simulation-dependent value (Q'). During subsequent decision-making, these two values are integrated by the same computational process. If this dual-updating model is true, I expect that the Pursuit group accurately recognizes the actual reward probability due to representing the experience-dependent value. Future studies should investigate whether this proposed model could explain the pursuit of the hard-to-get avatar, and how this model relates to perseverance model through simulations and experimental investigations.

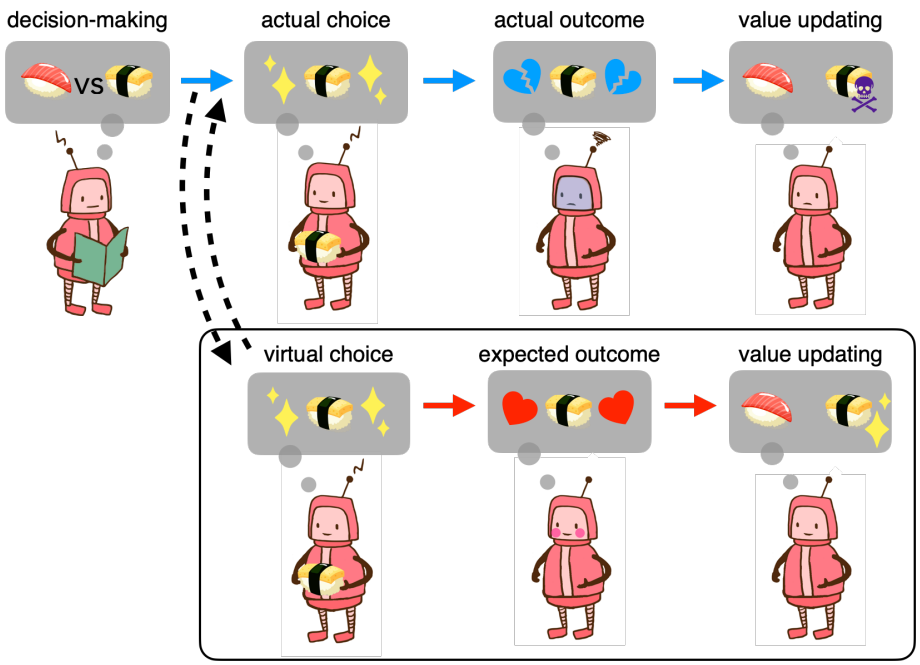


Figure 4.1 A dual-updating model. This model assumes two types of value-updating processes. One is the updating from actual outcomes which is formulated in standard reinforcement learning models. Another is the updating from expected outcomes which is simulated during decision-making. I assume that the values resulted from these two types of updating processes are independently represented in our brain, and are integrated to decide subsequent choices.

The value-updating from the expected outcomes in the proposed model is not limited to the social context addressed in Chapter 3, because the expectation of the desired outcomes is observed in non-social situations (e.g., lottery, auction, and shopping). Thus, future studies should investigate whether the pursuit of the hard-to-get object could be observed in non-social situations. In addition, value-dependent decision-making is generally investigated in non-human species. If the value-updating from the expected outcomes is observed across species, the neural mechanisms of this novel updating process could be extensively investigated by using cutting-edge neuroscientific technologies such as opto-genetics, chemo-genetics, and neuropharmacological interventions. It should be investigated whether the pursuit of the hard-to-get option is observed in non-human species.

4.3 Other psychological factors related to repetitive choice behaviors

In this thesis, I addressed asymmetric value updating and choice perseverance as source of repetitive choice behaviors. However, there are many psychological factors related to repetitive choice behaviors: sunk-cost effect (Arkes & Blumer, 1985; Haller & Schwabe, 2014; Olivola, 2018), rarity (Hertwig et al., 2004; Kahneman & Tversky, 1979; Williams et al., 2016; Worchel et al., 1975), curiosity (Kidd & Hayden, 2015; Rigoli et al., 2019) and gambler's fallacy (Jarvik, 1951; Jessup et al., 2011).

To understand how these factors induce the repetitive choices, let you consider the situation that you have bought many sweets to collect a secret free gift (e.g. snacks with a baseball card and "chocolate eggs"). If you buy extra sweets because a lot of money has already spent, this repetition is resulted from a sunk-cost effect. Sunk-cost effect refers to the pursuit of the option which significant, unrecoverable resources are invested previously. If you buy extra sweets because you want to know what is a secret gift, the cause of repetition is a curiosity which is a special form of information-seeking (Loewenstein, 1994; Oudeyer & Kaplan, 2009). If you believe that will definitely get a secret toy next time since you bought many times it but not get, your behavior is governed by a gambler's fallacy. Individuals adhering to the gambler's fallacy appear to assume non-independence between sequential outcomes (Tversky & Kahneman, 1971). Although these three factors (i.e., sunk-cost effect, curiosity, and gambler's fallacy) make you to continue previous actions, these factors are disappeared once you get a secret gift. However, in our experiment showed in Chapter 3, subjects that pursue the hard-to-get avatar continued to choose it after the avatar responded in a positive manner.

On the other hand, if you buy extra sweets because a secret gift is rare, a rarity is an important drive for the repetitive choice. When attempting to determine the value of unknown items, we may utilize the assumption that rare items are inherently more valuable than abundant items simply because they are rare (the scarcity heuristics; Williams et al., 2016). In our experiment, it is possible that the hard-to-get avatar is recognized as more valuable due to the fact that the avatar rarely returns a positive response. However, it is unknown whether the rarity influences the pursuing behaviors, and how subjects incorporate the rarity into the action selection if it modulates pursuing behaviors. To answer these questions, future studies should manipulate the rarity of avatars and investigate the effect of this manipulation on pursuing behaviors. Specifically, as Shin & Ariely, (2004) conducted, if subjects are instructed that the paired avatar is changed after a few trials, does information about rarity make the subjects that pursue the hard-to-get-avatar to shift their choice from the avatar that we have pursued until now to the other avatar? It is necessary to develop the cognitive computation models incorporating these additional factors and to clarify the interaction between psychological factors on the pursuing behaviors.

4.4 Usefulness of computational modeling

As I showed in Chapter 2, cognitive computation models allow us to dissociate complicated information processing underlying behaviors that are seemingly the same. By using cognitive computation models, I demonstrated that the pursuit of the hard-to-get option is mainly produced from the choice-dependent and outcome-independent information processing. Moreover, this choice-dependent processing strongly affects subjective value. In parallel, computational approach highlights the quantitative difference of information processing underlying qualitatively different behaviors. I showed in Chapter 3, computational models can clarify the difference of implicit information processing leading the individual difference in behaviors in the specific situation. As I mentioned in Chapter 1, computational modeling approach is tightly linked to neuroscience. Because computational models formulate implicit processing, the estimated variables provide the interpretation of the neural activity measured from behaving animals. According to these advantages, computational approach is essential to understand the mechanisms of animal behaviors.

On the other hand, computational modeling is not a panacea. As I investigated in Chapter 2 as well as Katahira (2018), an insufficient cognitive computation model results in incorrect interpretations for our behaviors. To avoid such model misspecification, careful simulations and well-designed psychological experiments are necessary (Wilson et al., 2019). In

these precautions are taken, cognitive computation modeling is an extremely useful tool, and significantly contribute to a comprehensive understanding of psychological processes.

4.5 Concluding remarks

Through this chapter, I discussed the mechanisms of the choice perseverance behind pursuing behaviors. People persist in their actions even when the consequences of one's own choices are not desirable. In Chapter 3, the pursuit of the hard-to-get avatar showed in Pursuit group seems to be associated with one aspect of stalking behaviors. Stalkers relentlessly approach their target even if they don't get any favors. This behavior is socially problematic, developing criminal behavior and psychiatric disorders. Furthermore, I found that the computational process which is independent of chosen outcome leads to repetitively choose the hard-to-get target, consequently increasing subjective attractiveness for the target despite seldom positive reactions. This finding highlights the computational processing of intrinsic information for the understanding of maladaptive choice behaviors. How to modify these problematic behaviors is the general interest. Psychotherapist often attempts to modify problematic behaviors by altering action-outcome associations. However, our present findings imply that controlling environmental factors (i.e., manipulating the reward probability in the surrounding environment) is insufficient to solve the pursuing behaviors. Future study should investigate how to modify the outcome-independent process underlying the pursuing behavior, and might contributes to develop the novel approach of behavior modification.

Like a coin, there are always two sides to every event in this world. Pursuing behaviors are not always problematic. As I mentioned in Chapter 1, scientists passionately pursue their interest even if they do not get desirable results. Old painters that were never admired in their lifetime continued to paint in their passions. Although pursuing behaviors are looked like irrational for observers, such behavior is essential for achieving great works. The only way for getting the chance of success is never give up. Choice perseverance seems to be associated with the tolerance of negative outcomes. My working hypothesis is that the expected outcomes from the current choice mediates the choice perseverance. As I mentioned in Chapter 1, positivity and confirmation biases refer to the preferential attention to own desirable outcomes, resulting in the reducing the effect of undesirable outcomes on future decisions. If my hypothesis is true, optimistic imaginations underlying choice perseverance also prevent the impact of undesirable outcomes. However, the choice perseverance is significantly differed with positivity and confirmation biases because the optimistic imagination depends on the expected outcomes but

not on the obtained outcomes. Thus, the optimistic imagination might maintain one's belief even in difficult situations.

In summary, choice perseverance not only develops the problematic behaviors observed in psychiatric disorders and criminal behaviors, but also contributes to maintain our passion and motivation. Studying pursuing behaviors might provide valuable information not only to prevent pathological behaviors but also to improve our passion and motivation in our daily life.

References

Akaishi, R., Umeda, K., Nagase, A., & Sakai, K. (2014). Autonomous Mechanism of Internal Choice Estimate Underlies Decision Inertia. *Neuron*, *81*(1), 195–206. https://doi.org/10.1016/j.neuron.2013.10.018

Alós-Ferrer, C., Hügelschäfer, S., & Li, J. (2016). Inertia and Decision Making. *Frontiers in Psychology*, *7*(February), 1–9. https://doi.org/10.3389/fpsyg.2016.00169

Alós-Ferrer, C., & Shi, F. (2015). Choice-induced preference change and the free-choice paradigm: A clarification. *Judgment and Decision Making*, *10*(1), 34–49. https://doi.org/10.2139/ssrn.2062507

Ariely, D., & Norton, M. I. (2008). How actions create - not just reveal - preferences. *Trends in Cognitive Sciences*, *12*(1), 13–16. https://doi.org/10.1016/j.tics.2007.10.008

Arkes, H. R., & Blumer, C. (1985). The psychology of sunk cost. *Organizational Behavior and Human Decision Processes*, *35*(1), 124–140. https://doi.org/10.1016/0749-5978(85)90049-4

Barto, A. G. (1997). Neural Systems for Control. In O. M. Omidvar & D. L. Elliott (Eds.), *Reinforcement learning* (Issue 1997, pp. 7–27).

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., & Fox, J. (2019). *Linear Mixed-Effects Models using "Eigen" and S4*. https://cran.r-project.org/web/packages/lme4/lme4.pdf

Bertelson, P. (1965). Serial choice reaction-time as a function of respomse versus signal-and-response repetition. *Nature*, *206*, 217–218. https://doi.org/http://dx.doi.org/10.1038/2051060a0

Bonaiuto, J. J., De Berker, A., & Bestmann, S. (2016). Response repetition biases in human perceptual decisions are explained by activity decay in competitive attractor models. *ELife*, *5*(DECEMBER2016), 1–28. https://doi.org/10.7554/eLife.20047

Boorman, E. D., Behrens, T. E., & Rushworth, M. F. (2011). Counterfactual choice and learning in a Neural Network centered on human lateral frontopolar cortex. *PLoS Biology*, *9*(6). https://doi.org/10.1371/journal.pbio.1001093

Brehm, J. W. (1956). Postdecision changes in the desirability of alternatives. *Journal of Abnormal and Social Psychology*, *52*(3), 384–389. https://doi.org/10.1037/h0041006

Cockburn, J., Collins, A. G. E., & Frank, M. J. (2014). A Reinforcement Learning Mechanism Responsible for the Valuation of Free Choice. *Neuron*, *83*(3), 551–557. https://doi.org/10.1016/j.neuron.2014.06.035

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. In M. R. Delgado, E. A. Phelps, & T. W. Robbins (Eds.), *Decision Making, Affect, and Learning: attention and performance XXIII* (1st ed., pp. 3–38). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199600434.003.0001

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

Daw, N. D., & Tobler, P. N. (2013). Value Learning through Reinforcement: The Basics of Dopamine and Reinforcement Learning. *Neuroeconomics: Decision Making and the Brain: Second Edition*, 283–298. https://doi.org/10.1016/B978-0-12-416008-8.00015-2

Doya, K. (2007). Reinforcement learning: Computational theory and biological mechanisms. *HFSP Journal*, *1*(1), 30. https://doi.org/10.2976/1.2732246

Egan, L. C., Santos, L. R., & Bloom, P. (2007). The origins of cognitive dissonance: Evidence from children and monkeys. *Psychological Science*, *18*(11), 978–983. https://doi.org/10.1111/j.1467-9280.2007.02012.x

Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2017). From Anomalies to Forecasts : Toward a Descriptive From Anomalies to Forecasts : Toward a Descriptive Model of Decisions. *Psychological Review*, *124*(4), 369–409. http://departments.agri.huji.ac.il/economics/teachers/ert_eyal/CPC2015.pdf

Erev, I., Haruvy, E., Kagel, J. H., & Roth, A. E. (2013). Learning and the Economics of Small Decisions. In J. H. K. and A. E. Roth. (Ed.), *The Handbook of Behavioral Economics*. Princeton Univ Press.

Eysenck, H. J. (1959). Learning Theory Therapy. *Journal of Mental Science*, *105*(438), 61–75.

Eysenck, H. J. (1987). Behavior Therapy. In *Theoretical Foundations of Behavior Therapy* (Issue 1970, pp. 3–35). Springer US. https://doi.org/10.1007/978-1-4899-0827-8_1

Fischer, A. G., & Ullsperger, M. (2013). Article Real and Fictive Outcomes Are Processed Differently but Converge on a Common Adaptive Mechanism. *Neuron*, *79*(6), 1243–1255. https://doi.org/10.1016/j.neuron.2013.07.006

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 of the United States of America*, *104*(41), 16311–16316. https://doi.org/10.1073/pnas.0706111104

Gershman, S. J., Pesaran, B., & Daw, N. D. (2009). Human Reinforcement Learning Subdivides Structured Action Spaces by Learning Effector-Specific Values. *Journal of Neuroscience*, *29*(43), 13524–13531. https://doi.org/10.1523/jneurosci.2469-09.2009

Gershman, Samuel J. (2015). Do learning rates adapt to the distribution of rewards? *Psychonomic Bulletin and Review*, *22*(5), 1320–1327. https://doi.org/10.3758/s13423-014-0790-3

Ghalanos, A., & Maintainer, S. T. (2015). *Rsolnp: general non-linear optimization using augmented Lagrange multiplier method*.

Ghalanos, A., & Theussl, S. (2015). *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method.* R Package Version 1.16. https://rdrr.io/cran/Rsolnp/

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

Glimcher, P. W. (2009). Chapter 32 - Choice: Towards a Standard Back-pocket Model. In P. W. Glimcher & E. Fehr (Eds.), *Neuroeconomics* (pp. 503–521). Elsevier. https://doi.org/10.1016/B978-0-12-374176-9.00032-4

Gold, J. I., Law, C. T., Connolly, P., & Bennur, S. (2008). The relative influences of priors and sensory evidence on an oculomotor decision variable during perceptual learning. *Journal of Neurophysiology*, *100*(5), 2653–2668. https://doi.org/10.1152/jn.90629.2008

Haller, A., & Schwabe, L. (2014). Sunk costs in the human brain. *NeuroImage*, *97*, 127–133. https://doi.org/10.1016/j.neuroimage.2014.04.036

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*(8), 534–539. https://doi.org/10.1111/j.0956-7976.2004.00715.x

Hornsby, A. N., & Love, B. C. (2020). How decisions and the desire for coherency shape subjective preferences over time. *Cognition*, *200*(February), 104244. https://doi.org/10.1016/j.cognition.2020.104244

Huys, Q. J. M., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R. J., & Dayan, P. (2011). Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. *PLoS Computational Biology*, *7*(4). https://doi.org/10.1371/journal.pcbi.1002028

Huys, Q. J. M., Moutoussis, M., & Williams, J. (2011). Are computational models of any use to psychiatry? *Neural Networks*, *24*(6), 544–551. https://doi.org/10.1016/j.neunet.2011.03.001

Iigaya, K., Hauser, T. U., Kurth-Nelson, Z., O'Doherty, J. P., Dayan, P., & Dolan, R. J. (2020). The value of what's to come: Neural mechanisms coupling prediction error and the utility of anticipation. *Science Advances*, *6*(25). https://doi.org/10.1126/sciadv.aba3828

Izuma, K., & Murayama, K. (2013). Choice-Induced Preference Change in the Free-Choice Paradigm: A Critical Methodological Review. *Frontiers in Psychology*, *4*(February), 1–12. https://doi.org/10.3389/fpsyg.2013.00041

Jarvik, M. E. (1951). Probability learning and a negative recency effect in the serial anticipation of alternative symbols. *Journal of Experimental Psychology*, *41*(4), 291–297. https://doi.org/10.1037/h0056878

Jessup, R. K., Doherty, J. P. O., & Win, L. (2011). *Human Dorsal Striatal Activity during Choice Discriminates Reinforcement Learning Behavior from the Gambler ' s Fallacy*. *31*(17), 6296–6304. https://doi.org/10.1523/JNEUROSCI.6421-10.2011

Kahneman, B. Y. D., & Tversky, A. (1979). *Kahneman2013*. *47*(2), 263–291.

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

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. https://doi.org/10.1016/j.jmp.2015.03.006

Katahira, K. (2016). How hierarchical models improve point estimates of model parameters at the individual level. *Journal of Mathematical Psychology*, *73*, 37–58. https://doi.org/10.1016/j.jmp.2016.03.007

Katahira, K. (2018). The statistical structures of reinforcement learning with asymmetric value updates. *Journal of Mathematical Psychology*, *87*, 31–45. https://doi.org/10.1016/j.jmp.2018.09.002

Kidd, C., & Hayden, B. Y. (2015). Perspective The Psychology and Neuroscience of Curiosity. *Neuron*, *88*(3), 449–460. https://doi.org/10.1016/j.neuron.2015.09.010

Koster, R., Duzel, E., & Dolan, R. J. (2015). Action and valence modulate choice and choice-induced preference change. *PLoS ONE*, *10*(3), 1–10. https://doi.org/10.1371/journal.pone.0119682

Kuzmanovic, B., & Rigoux, L. (2017). Valence-dependent belief updating: Computational validation. *Frontiers in Psychology*, *8*(JUN), 1–11. https://doi.org/10.3389/fpsyg.2017.01087

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. https://doi.org/10.1901/jeab.2005.110-04

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, *1*(4), 0067. https://doi.org/10.1038/s41562-017-0067

Lieberman, M. D., Ochsner, K. N., Gilbert, D. T., & Schacter, D. L. (2001). Do Amnesics Exhibit Cognitive Dissonance Reduction? The Role of Explicit Memory and Attention in Attitude Change. *Psychological Science*, *12*(2), 135–140. https://doi.org/10.1111/1467-9280.00323

Loewenstein, G. (1987). Anticipation and the Valuation of Delayed Consumption. *The Economic Journal*, *97*(387), 666. https://doi.org/10.2307/2232929

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, *116(1)*, 75–98. https://doi.org/10.1037/0033-2909.116.1.75

Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019). Habits without values. *Psychological Review*, *126*(2), 292–311. https://doi.org/10.1037/rev0000120

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation Bias in a Simulated Research Environment: An Experimental Study of Scientific Inference. *Quarterly Journal of Experimental Psychology*, *29*(1), 85–95. https://doi.org/10.1080/00335557743000053

Nakao, T., Kanayama, N., Katahira, K., Odani, M., Ito, Y., Hirata, Y., Nasuno, R., Ozaki, H., Hiramoto, R., Miyatani, M., & Northoff, G. (2016). Post-response βγ power predicts the degree of choice-based learning in internally guided decision-making. *Scientific Reports*, *6*(April), 1–9. https://doi.org/10.1038/srep32477

Niv, Y. (2020). On the Primacy of Behavioral Research for Understanding the Brain. *Current Controversies in Philosophy of Cognitive Science*, 134–151. https://doi.org/10.4324/9781003026273-16

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, *32*(2), 551–562. https://doi.org/10.1523/JNEUROSCI.5498-10.2012

Olivola, C. Y. (2018). *The Interpersonal Sunk-Cost Effect*. https://doi.org/10.1177/0956797617752641

Oudeyer, P. Y., & Kaplan, F. (2009). What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics*, *3*(NOV), 1–14. https://doi.org/10.3389/neuro.12.006.2007

Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature Communications*, *6*, 1–14. https://doi.org/10.1038/ncomms9096

Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S.-J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLOS Computational Biology*, *13*(8), e1005684. https://doi.org/10.1371/journal.pcbi.1005684

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433. https://doi.org/10.1016/j.tics.2017.03.011

Peeters, G. (1971). The positive-negative asymmetry: On cognitive consistency and positivity bias. *European Journal of Social Psychology*, *1*(4), 455–474. https://doi.org/10.1002/ejsp.2420010405

Perez, O. D., & Dickinson, A. (2020). A Theory of Actions and Habits: The Interaction of Rate Correlation and Contiguity Systems in Free-Operant Behavior. *Psychological Review*. https://doi.org/10.1037/rev0000201

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Redish, A. D., & Johnson, A. (2008). A unified framework for addiction: vulnerabilities in the decision process. *Behavioral Brain Science*, *31*(4), 415–487. https://doi.org/10.1017/S0140525X0800472X

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: current research and theory* (1st ed., pp. 64–99). Appleton-Century-Crofts.

Rigoli, F., Martinelli, C., & Shergill, S. (2019). The role of expecting feedback during decision-making under risk. *NeuroImage*, *202*(April), 116079. https://doi.org/10.1016/j.neuroimage.2019.116079

Schonberg, T., Bakkour, A., Hover, A. M., Mumford, J. A., Nagar, L., Perez, J., & Poldrack, R. A. (2014). Changing value through cued approach: An automatic mechanism of behavior change. *Nature Neuroscience*, *17*(4), 625–630. https://doi.org/10.1038/nn.3673

Schönberg, T., Daw, N. D., Joel, D., & O'Doherty, J. P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *Journal of Neuroscience*, *27*(47), 12860–12867. https://doi.org/10.1523/JNEUROSCI.2496-07.2007

Sears, D. O. (1983). The person-positivity bias. *Journal of Personality and Social Psychology*, *44*(2), 233–250. https://doi.org/10.1037/0022-3514.44.2.233

Sharot, T., Martino, B. De, & Dolan, R. J. (2009). How choice reveals and shapes expected hedonic outcome. *Journal of Neuroscience*, *29*(12), 3760–3765. https://doi.org/10.1523/JNEUROSCI.4972-08.2009

Shin, J., & Ariely, D. (2004). Keeping doors open: The effect of unavailability on incentives to keep options viable. *Management Science*, *50*(5), 575–586. https://doi.org/10.1287/mnsc.1030.0148

Shteingart, H., Neiman, T., & Loewenstein, Y. (2013). The role of first impression in operant learning. *Journal of Experimental Psychology: General*, *142*(2), 476–488. https://doi.org/10.1037/a0029550

Skinner, B. F. (1938). *The behavior of organisms: an experimental analysis.* Appleton-Century.

Sugawara, M & Katahira, K. (2019). Cognitive biases and perseverance in reinforcement learning: Doesyourcurrentchoicebehaviordependonpast“ choiceoutcome”or“ choiceperse”? *The Japanese Journal of Psychonomic Science*, *38*, 1–17.

Sutton, R. S., & Barto., A. G. (1998). *Reinforcement Learning: An Introduction.* MIT Press.

Thorndike, E. L. (1911). *Animal intelligence; experimental studies,*. The Macmillan Company,. https://doi.org/10.5962/bhl.title.55072

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110. https://doi.org/10.1037/h0031322

Urai, A. E., Braun, A., & Donner, T. H. (2017). Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature Communications*, *8*. https://doi.org/10.1038/ncomms14637

Urai, A. E., de Gee, J. W., Tsetsos, K., & Donner, T. H. (2019). Choice history biases subsequent evidence accumulation. *ELife*, *8*, 1–34. https://doi.org/10.7554/eLife.46331

von Neumann, J., & Morgenstern, O. (1944). Theory of Games and Economic Behavior. In *Princeton University Press*.

Watoson, J. B. (1930). *Behaviorism.* Norton.

Williams, C. C., Saffer, B. Y., McCulloch, R. B., & Krigolson, O. E. (2016). The scarcity heuristic impacts reward processing within the medial-frontal cortex. *NeuroReport*, *27*(7), 522–526. https://doi.org/10.1097/WNR.0000000000000575

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *ELife*, *8*, 1–35. https://doi.org/10.7554/eLife.49547

Worchel, S., Lee, J., & Adewole, A. (1975). Effects of supply and demand on ratings of object value. *Journal of Personality and Social Psychology*, *32*(5), 906–914. https://doi.org/10.1037//0022-3514.32.5.906