State-transition-free reinforcement learning in chimpanzees (*Pan troglodytes*)

Yutaro Sato, Yutaka Sakai, Satoshi Hirata

Abstract

The outcome of an action often occurs after a delay. One solution for learning appropriate actions from delayed outcomes is to rely on a chain of state transitions. Another solution, which does not rest on state transitions, is to use an eligibility trace (ET) that directly bridges a current outcome and multiple past actions via transient memories. Previous studies revealed that humans (*Homo sapiens*) learned appropriate actions in a behavioral task in which solutions based on the ET were effective but transition-based solutions were ineffective. This suggests that ET may be used in human learning systems. However, no studies have examined nonhuman animals with an equivalent behavioral task. We designed a task for nonhuman animals following a previous human study. In each trial, participants chose one of two stimuli that were randomly selected from three stimulus types: a stimulus associated with a food reward delivered immediately, a stimulus associated with a reward delivered after a few trials, and a stimulus associated with no reward. The presented stimuli did not vary according to the participants' choices. To maximize the total reward, participants had to learn the value of the stimulus associated with a delayed reward. Five chimpanzees (*Pan troglodytes*) performed the task using a touchscreen. Two chimpanzees were able to learn successfully, indicating that learning mechanisms that do not depend on state transitions were involved in the learning processes. The current study extends previous ET research by proposing a behavioral task and providing empirical data from chimpanzees.

Keywords

Chimpanzee; Credit assignment problem; Eligibility trace; Reinforcement learning

Introduction

When an agent learns to perform an action to obtain an outcome, the outcome usually appears immediately after the action, which reinforces the preceding action. However, learning becomes challenging when the outcome of an action is delayed and preceded by other actions because deciding how to distribute credit for the outcome to multiple actions is complex (i.e., the credit assignment problem) (Minsky, 1961). The theory of reinforcement learning (RL) (Sutton & Barto, 2018) provides several solutions for the credit assignment problem. It should be noted that the term *reinforcement learning* is used in three overlapping and inter-related disciplines: psychology, neuroscience, and computational science (Eckstein et al., 2021; Sutton & Barto, 2018; Yoo & Collins, 2022). In the current study, RL refers to the body of related knowledge and concepts

developed in computational science (Sutton & Barto, 2018). RL is formalized as follows (Sutton & Barto, 2018): in each time step, an agent selects an action in a certain environmental state. The agent's action elicits responses from the environment, such as a change of situation (i.e., a transition to the next state) and an outcome that has a specific immediate effect in relation to the agent (i.e., a reward). In the next time step, the agent takes an action in the state, which again leads to the state transition and reward feedback. The agent undertakes this process iteratively. In simple cases, it is assumed that the state transition and reward depend solely on the most recent state and action, and not on previous transitions and rewards. This feature is called the Markov property. The Markov process consisting of state, action, and reward is called the Markov decision process. The Markov property enables estimation of long-term rewards through a chain of step-by-step inference about state transition and reward (i.e., model of the environment). Learning strategies based on the step-by-step model of the environment are categorized into model-based RL, and strategies that are not based on such a model are categorized into model-free RL.

Standard solutions for the credit assignment problem involve computations resting on step-by-step state transitions. Solutions of this type are implemented in both model-based and model-free RLs. The model-based RL directly predicts delayed outcomes in the future through a chain of step-by-step inferences. In model-free RL, a canonical algorithm called temporal-difference (TD) learning (Sutton, 1988) (the algorithm we used is provided in the Online Supplementary Materials (OSM)) is widely used. TD learning estimates the value of delayed outcomes in the future through step-by-step experiences of state transitions. Although these transition-based solutions work well in Markov decision processes, they fail to bridge each action and the delayed outcome in non-Markov situations in which the probability of a delayed outcome is not determined by the current state.

Solutions that are independent of state transitions have also been proposed in model-free RL. One of these solutions is based on eligibility traces (ETs), which directly bridge a current outcome with multiple actions taken in recent experience (Singh & Sutton, 1996; Sutton & Barto, 2018). The ET of each action is a transient memory reflecting the recent frequency of the action, which determines the eligibility of an update by the current outcome. In other words, credits for a given reward are assigned to past actions depending on the recency of each action, in a manner such that a temporally closer action is assigned greater credit, which is accordingly reflected in the update of each action's value. Transient memories that are independent of state transitions enable bridging of the delayed outcome to the responsible actions, even in non-Markov situations. The ET method is usually combined with TD learning to assign prediction-based updates with recent actions (referred to as TD[λ] in Sutton & Barto, 2018), providing forward (TD) and backward (ET) views of learning processes.

Neurophysiological evidence to support TD learning is well established. Dopamine signals resemble the prediction error that can be used for TD learning (Schultz, 1997), and dopamine-dependent synaptic plasticity has been observed in cortico-striatal projections (Shen et al., 2008). In addition, recent studies have provided accumulating evidence that supports ET. For example, Yagishita et al. (2014) revealed the time window (typically around 0.6 s) in which the delayed dopamine signal affects dopamine-dependent synaptic plasticity. This suggests that ET might be implemented in the plasticity mechanisms at each synapse. Nonomura et al. (2018) reported that responses of striatal neurons to outcomes reflect the preceding action, which suggests the possibility that ET might be sustained by neuronal activity. This line of research has advanced our understanding regarding the proximate mechanisms underlying ET (see Gerstner et al., 2018, for a review of some key findings).

From a behavioral perspective, ET is suggested to play an important role, particularly in behavioral tasks involving delayed rewards that are delivered to participants after multiple actions (e.g., Bogacz et al., 2007; Gureckis & Love, 2009; Lehmann et al., 2019; Sakai et al., 2022; Tanaka et al., 2009; Tartaglia et al., 2017; Walsh & Anderson, 2011). For example, Tanaka et al. (2009) examined the role of serotonin in the learning of delayed reward and punishment in humans (*Homo sapiens*) using a behavioral task in conjunction with experimental manipulation of tryptophan intake. In each trial in this task, participants were presented with a pair of choice stimuli, selected one stimulus, and received visual feedback showing how much they earned or lost. Tanaka et al. (2009) used eight stimuli in total, each of which was associated with a monetary gain or loss. Importantly, the outcomes associated with half of the stimuli were not delivered within a trial but were delivered three trials later. Thus, upon receiving feedback at the end of a trial, participants needed to update the values of past actions (i.e., which stimulus they selected in previous trials) as well as the value of the action taken in that trial (i.e., which stimulus they selected in that trial). In addition, participants' choices did not affect which stimulus pair was presented in subsequent trials. That is, their actions were not related to state transitions. These features of the task negate the effects of transition-based solutions to the credit assignment problem (e.g., model-based or TD learning). Therefore, success in the task would support transition-free solutions, such as ET. Participants who underwent different tryptophan manipulations were found to achieve learning at different speeds. Computational modeling indicated that a learning algorithm with ET fitted to participants' behavior better than an algorithm without ET and another algorithm that held memory of the choices made in the past three trials. This result suggests that participants were unlikely to learn by processing multiple state-action pairs in parallel using working memory, which may be computationally too expensive. Furthermore, regarding learning of delayed punishment, a parameter in learning algorithms that determined how fast ET decayed (i.e., trace-decay parameter) was found to differ between conditions. Specifically, participants who consumed a large amount of tryptophan exhibited a larger trace-decay parameter compared with those who consumed no tryptophan. A higher trace-decay parameter indicated that the decay of ET was slower, and that past actions were eligible for value update for longer. Thus, this result suggested that tryptophan manipulation impacted the speed of ET decay. Taken together, these findings support the notion that participants can use their action history via ET, and highlighted the role of serotonin in learning processes involving delayed punishments. As argued in another study (Lehmann et al., 2019), behavioral tasks that are specifically designed for studying ET may be useful for closely focusing on ET while negating alternative learning methods, thereby revealing a behavioral signature of ET without relying on inferences from computational modeling.

Several studies of nonhuman animals (Nosarzewska et al., 2021; Smith et al., 2020; Zentall et al., 2022) have used apparently similar techniques to that of Tanaka et al. (2009). For example, in a study of rhesus macaques (*Macaca mulatta*), some individuals were found to learn discrimination even when the feedback for responses was delayed by one trial (Smith et al., 2020). Some pigeons were also found to perform well in similar tasks (Nosarzewska et al., 2021; Zentall et al., 2022). However, those studies focused on a different issue (i.e., explicit-declarative vs. implicit-procedural learnings) and did not discuss their findings in relation to ET. Thus, it remains unclear how nonhuman animals learn in a credit assignment problem situation, such as the task designed by Tanaka et al. (2009).

Furthermore, these studies did not analyze behavioral data using computational models of RL, which are useful for shedding light on a particular aspect of learning (Katahira, 2018; Scholl & Klein-

Flügge, 2018; Tartaglia et al., 2017). Tanaka et al. (2009) observed differences among conditions not only in tallied behavioral data (i.e., proportion of trials in which participants made optimal choices), but also in estimated parameters in models (i.e., trace-decay parameter). Additionally, a recent study highlighted the characteristics of learning in obsessive-compulsive disorder (OCD) patients using Tanaka's et al. (2009) task (Sakai et al., 2022). The OCD patients exhibited a poorer learning performance from delayed feedbacks compared to control participants. Moreover, computational modeling revealed that the OCD patients exhibited an imbalance in trace-decay parameters for positive and negative prediction errors. A similar approach would be useful in comparative psychology research to reveal similarities and differences in ETs across animal species (Redish et al., 2022), thereby providing insights into the timescale of learning abilities from an evolutionary perspective.

In the current study, we devised a behavioral task for nonhuman animals following the task designed by Tanaka et al. (2009). An essential step for promoting this line of research for nonhuman animals is to design a simple task according to the animals' level of abilities and motivation (Pike et al., 2021). For example, an experiment on mice (Akam et al., 2021) adapted an original task designed for humans (i.e., two-stage task: Daw et al., 2011) by modifying task features such as the number of action alternatives and reward probabilities, to encourage mice to engage in the task. We modified the original task designed by Tanaka et al. (2009) in several respects.

First, we focused on the learning of delayed rewards and did not examine the learning of punishments because punishment is difficult to implement in the task. For humans, punishment can be implemented as a loss of money or fictitious points. In contrast, such symbolic rewards would complicate learning for nonhuman animals. Instead, we used pieces of food (i.e., pellets) as rewards, which were consumed immediately. Hence, it was difficult to represent punishment. Second, we reduced the number of visual stimuli and associated each stimulus with the presence or absence of a food reward (i.e., one pellet or zero pellets), rather than a different amount of reward, because using different amounts of reward would require us to verify that animals were able to distinguish the different amounts of food. Third, the number of trials over which a reward was carried forward was not fixed, but was designed to vary trial-to-trial, which prevented participants from easily noticing the task rule. Nonetheless, the length of the delay in the outcome, i.e., the number of trials that had passed before the reward, was not determined randomly (unlike Jocham et al., 2016, Experiment 2). Rather, we designed the delay rule so that the number of possible rewards within a trial was limited to one or zero to avoid any burden for animals associated with distinguishing different amounts of food reward (see the Method section, Task design).

In our task, the delayed outcome created a non-Markov situation in which the probability of a delayed outcome was not determined by the current state. Computations resting on a chain of state transitions alone, either via model-based learning or TD learning without ETs, were insufficient to learn successfully in the task. Thus, this task is useful for examining learning mechanisms such as ETs that are not largely dependent on a chain of state transitions. Using this task, we examined the extent to which chimpanzees (*Pan troglodytes*) could solve a credit assignment problem by utilizing their action history via transition-free solutions such as ET. We also created a preliminary report on the performance of human participants in a similar task, which can be found in the OSM.

Method

Participants

Five chimpanzees participated in the present study (one male and four females; 11－25 years old; Table 1). Participation in the test was voluntary. Another chimpanzee who participated in the pilot experiments was not included in the current study because she chose not to undergo the test, possibly because of a lack of motivation. The chimpanzees lived in a social group, which consisted of those six individuals. They usually spent the daytime in outdoor enclosures furnished with platforms, ropes, or trees, creating a three-dimensional complex environment wherein they could comfortably exercise (total area: 294 m^2). The chimpanzees also used indoor enclosures at times, such as when they received meals. They received meals three times a day, consisting of fresh vegetables, fruits, nuts, and monkey chow. Water was available ad libitum from taps in the enclosures. Additional enrichment items were also provided to facilitate active foraging activities (e.g., small packages of food items) or improve comfort (e.g., pieces of burlap bags). The chimpanzees had previously experienced various cognitive experiments including touchscreen experiments (Sato et al., 2020).

Apparatus

Six 17-in. LCD touchscreen monitors (Touch Panel Systems K.K., Kanagawa, Japan, the model number was not recorded) were installed in a row in one of the outdoor enclosures (Fig. 1A). The resolution was set at 1,024 × 768 pixels. In this setting, chimpanzees could engage in experiments in their home enclosure without separation from other groupmates. An experimental booth was attached to the outside of the enclosure, in which a human experimenter operated experiments while observing chimpanzees through transparent panels (Fig. 1B). Touchscreens were mounted in a custom-made polycarbonate box, which was fixed in a metallic frame. Chimpanzees could reach the touchscreen through a rectangle opening of the box. Chimpanzees received food rewards through small holes in the lateral sides of the box. This setting allowed chimpanzees to touch the screen and receive rewards while preventing them from banging the screen aggressively. Laptop computers (TravelMate P250, Acer, Taiwan) and pellet dispensers (ENV-203-190, Med Associates Inc., VT, USA) were installed inside the experimental booth (Fig. 1B). The pellet dispenser was connected to the computer via an I/O unit (DIO-8/8 [USB] GY, Contec, Osaka, Japan) and a switching power supply (S82J-0124D, Omron, Kyoto, Japan), and the touchscreen was connected directly to the computer. The pellet dispensers released food rewards (190-mg banana flavor pellets: Dustless Precision Pellets, Primate, Purified F0035, Bio-Serv, NJ, USA) to the chimpanzee through pipes. Control of the experiment, recordings of responses, and regulation of pellet dispensers were carried out by the computer via an experimental program written in Microsoft Visual Studio Community 2017 v.15.6.1 (Microsoft Corporation, WA, USA).

Stimuli

We used a set of capital letters as visual stimuli, which were placed on a white square with a black edge (240 × 240 pixels; Fig. 2A). Prior to this experiment, we conducted several days of pilot experiments using a different set of letters to ensure that the chimpanzees could readily learn to discriminate alphabetic characters (OSM). Chimpanzees appeared to learn discrimination between pairs of letters, particularly when they were perceptually dissimilar. On the basis of this result, we selected three letters that were perceptually dissimilar (i.e., “F,” “J,” and “Q”). We used the same allocation of letters to represent three types of choice options (see Task design below) for all

participants. Note that these stimuli were presented on a black background, and thus the black edge of the white squares could not be distinguished from the background. From the participants' perspective, the stimuli appeared as a white square on which a letter was placed. This design was used to accommodate errors in touch detection when chimpanzees touched a peripheral part of the stimulus.

Task design

There were three types of choice options, two of which were randomly presented in each trial (hereafter, stimuli (a)–(c); Fig. 2). Participants selected one of two presented options by touching it on the screen and received a food reward depending on the choice. The stimuli (a) and (c) followed simple contingencies, whereas stimulus (b) followed a unique contingency, which was a key element of the task design. First, when participants selected stimulus (a) (represented by an image of a "J") in a trial, they received a food pellet at the end of the trial. Second, when participants selected stimulus (b) (represented by an image of an "F") in a trial, this choice led to a pellet at the end of another trial, several trials later. The delayed reward followed a specific formula, as detailed in the next paragraph. Finally, when participants selected stimulus (c) (represented by an image of a "Q") in a trial, this choice itself did not lead to a pellet. Note that the allocation of images to each stimulus was identical across individuals.

When participants selected stimulus (b) in a trial (say, trial t), the reward associated with this choice was not delivered at the end of that trial (i.e., trial t), but was delivered in a later trial (say, trial $[t + n]$) where stimuli (b) and (c) were presented again for the first time after trial t. Importantly, the delayed reward was delivered regardless of whether stimuli (b) or (c) were selected in the trial $(t + n)$. The selection of stimulus (b) or (c) in trial $(t + n)$ was immediately followed by a reward because of the stimulus (b) selected in trial t, but not because of the choice made in trial $(t + n)$. Because the presented stimulus combination was selected randomly, so was the distribution of the reward delay (as shown by histograms in Fig. 3). Even if stimulus (b) had been selected in multiple trials before a trial in which stimuli (b) and (c) were presented, only one pellet was delivered, and the redundant choices of stimulus (b) were ignored, which restricted the possible number of pellets to one or zero in each trial.

Therefore, a successful learner would be expected (1) to select stimulus (a) over stimulus (b), (2) to select stimulus (a) over stimulus (c), and (3) to select stimulus (b) over stimulus (c). Learning of (1) and (2) did not require ET because participants needed to simply select actions that were rewarded immediately (i.e., selecting stimulus (a)). In contrast, learning (3) did require ET because both the optimal and suboptimal actions (i.e., select stimulus (b) and stimulus (c), respectively) were not rewarded immediately, but only the optimal actions yielded rewards eventually, between which other actions intervened (i.e., choices in other trials).

Procedure

Experiments took place between 9 a.m. and 12 p.m. after chimpanzees received a morning meal at approximately 8 a.m. When a chimpanzee sat in front of a touchscreen, an experimenter started the experiment. In each trial, participants first touched a start button (i.e., green rectangle, 100 × 100 pixels) that appeared at the bottom of the screen. Subsequently, two of the three stimuli were presented on the screen at two of three possible locations (i.e., upper center, lower left or lower right; Fig. 2B). The two presented stimuli were determined randomly, and the locations of two stimuli were

determined pseudo-randomly: we created six patterns of location assignment by assigning three stimuli to three possible locations. Each of the six patterns appeared once in every six trials, while one of the three stimuli was hidden to present two stimuli. Participants were allowed to select either stimulus for 30 s. After they selected a stimulus by touching it or 30 s elapsed without response, a 3-s inter-trial interval followed. During the inter-trial interval, a blank black screen was presented, and chimpanzees could receive a pellet depending on the choice made in the current or past trials. Subsequently, participants proceeded to the next trial, and the start button appeared again. Upon touching the start button and stimuli, a click sound was played. Delivery of a food reward was coupled with a chime sound, and a bomb sound was played in the trials in which no reward was delivered. Chimpanzees were familiar with these sounds because they experienced similar sounds in previous touchscreen experiments (e.g., Sato et al., 2020). These sounds could potentially serve to inform chimpanzees of the presence or absence of reward, even when they failed to catch a pellet ejected from the hole. Note that in some trials at the beginning of the first day for one participant (Iroha), pellets were not delivered properly because they became stuck in a joint in the pipe. We did not exclude those trials from the analysis.

Chimpanzees underwent trials as long as they continued, up to 450 trials each day. When chimpanzees left the touchscreen, they were able to resume the experiment if they came back. The experiment lasted until the day each chimpanzee completed a minimum of cumulative 1,350 trials (range 1,350 − 1,499 trials). This took 3 − 8 days, in which some days of participation were in succession, whereas others were several days apart. On each day, a new session was administered. Carry-over rewards owing to a choice of stimulus (b) were not retained across different days. While chimpanzees were not performing the experimental task, they were allowed to perform an irrelevant, familiar task (i.e., touching a flickering dot among non-flickering dots), which prevented them from interfering with others who were undergoing the experiment.

Data analysis

Overall discriminative performance

Five trials were omitted from analysis: three trials were omitted because of time-over (one for Iroha, Misaki, and Zamba) and two were omitted because of unexpected errors (Natsuki; data sheet contained errors for unknown reasons: the sheet was blank for one trial and the delivery of two pellets was recorded for the other trial). First, we examined whether participants learned to make optimal choices. We divided trials depending on which two stimuli were presented and subsequently separated trials into bins comprising 20 trials. For each 20-trial bin, we calculated the proportion of trials in which participants made optimal choices (i.e., selecting stimulus (a) when stimulus (a) vs. stimulus (b) was presented or when stimulus (a) vs. stimulus (c) was presented; and selecting stimulus (b) when stimulus (b) vs. stimulus (c) was presented). When participants made optimal choices in ≥ 15 trials, performance was significantly higher than the chance level (binomial test, $p < .05$). Note that the last trial bin may have comprised fewer than 20 trials because we did not fix the total number of trials. Additionally, some of the other bins contained one missing data point because of time-over (three cases) or for the error for unknown reasons (one case).

Reinforcement learning models

Second, we fitted an ET model combined with multiple standard TD learning algorithms (TD[λ] in Sutton & Barto, 2018) to the data from each participant separately and compared performance across algorithms. Note that we merged data from different days for the sake of simplicity, as if participants

had completed all trials in succession in a day. Five invalid trials, which were omitted from the analysis described in the previous section, were also omitted from model fitting, which may have altered the number of trials over which a carry-over reward was carried forward, although this constituted only a few cases among > 1,300 trials for each chimpanzee. A detailed description of algorithm and model fitting is provided in the OSM. Briefly, we considered three variants of TD learning (overviewed in Sutton & Barto, 2018): SARSA (Rummery & Niranjan, 1994), Q-learning (Watkins & Dayan, 1992), and Actor-Critic (e.g., Barto et al., 1983). The action to take at each trial is determined using those action values in conjunction with soft-max. In addition to the three algorithms, we fitted the respective algorithm without ET. These algorithms did not include ET, and, at each time step, the action value for a taken action (but not the values for actions not taken) was updated. We compared the performances of those algorithms for each participant separately based on the widely applicable information criterion (WAIC). Specifically, a model of the lowest WAIC value was favored. We used R software v.4.1.0 (R Core Team, 2021) and Stan software v.2.21.0 (Stan Development Team, 2019) via CmdStan v.2.27.0 in conjunction with cmdstanr v.0.4.0 R package (Gabry & Češnovar, 2021). For R and Stan codes, we referred to Katahira (2018) and the StatModeling Memorandum website (https://statmodeling.hatenablog.com/entry/calc-waic-wbic, accessed on 17 January 2021).

Post hoc analysis

Furthermore, we performed post hoc analyses to examine the effect of simpler learning mechanisms, namely the effect of immediate rewards and the effect of experienced delay length on participant performance. As described above (see Task design), the selection of either stimulus (b) or (c) in trial *t* could be immediately followed by a reward. This reward was linked to a past choice of stimulus (b), and was not influenced by the stimulus choice (stimulus (b) or (c)) in trial *t*. However, from the participants' perspective, the reward at trial *t* appeared to be caused by the choice of either stimulus (b) or (c) in trial *t*. Thus, it is possible that the participants chose a certain stimulus simply because that choice appeared to be more likely to be followed by an immediate reward.

To address this, we examined whether incidental immediate rewards occurred more frequently after selecting stimulus (b) versus stimulus (c). We compared the proportion of rewarded trials in which participants selected stimulus (b) to that for trials in which they selected stimulus (c) using Fisher's exact probability tests via the exact2x2 package in R (Fay, 2010). Moreover, for individuals who learned successfully, we ran a similar analysis focusing on trials completed before they achieved high performance. We anticipated that after they had learned successfully, they would have preferentially selected stimulus (b) over stimulus (c) in trials in which those two stimuli were presented. Thus, the choice of stimulus (b) would have been more likely to have been followed by an immediate reward.

Next, to examine whether the two successful individuals (see the Results section) accidentally experienced shorter delays compared with the three unsuccessful individuals, we examined the effect of experienced delay length on participant performance. Delay length was computed as the number of trials between the selection of stimulus (b) and the delivery of a delayed reward for this choice. When participants had selected stimulus (b) multiple times before a delayed reward, the delay length was computed from the most recent choice of stimulus (b) and the other choices of stimulus (b) were ignored. We compared the delay length among five chimpanzees using the asymptotic Kruskal-Wallis test, followed by comparisons between individuals using exact Wilcoxon-Mann-Whitney tests

via the coin package in R (Hothorn et al., 2006). Cliff's d was calculated using the effsize package in R (Torchiano, 2020). All statistical tests were two-tailed (α = .05).

Results

Overall discriminative performance

Figure 4 shows the proportion of trials in which each participant performed optimal choices. All five chimpanzees came to select stimulus (a) over the other stimuli in later 20-trial bins (binomial tests, $p < .05$). Furthermore, two chimpanzees (Mizuki and Natsuki) selected stimulus (b) over stimulus (c) in later trial bins (binomial tests, $p < .05$), whereas the other three did not.

Reinforcement learning models

Table 2 shows the WAIC values for each model. Overall, an algorithm with ET (i.e., SARSA[λ]) was favored for all of the participants, regardless of their performance. Table 3 shows the summary of sampling from posterior distributions for the parameters of the favored model for each participant.

Post hoc analysis

Figure 5 shows the proportion of trials in which the participants received a reward, shown separately according to stimulus. Regarding the two individuals who learned to select stimulus (b) instead of stimulus (c) (i.e., Mizuki and Natsuki), the proportion of rewarded trials significantly differed between stimuli (b) and (c). Mizuki selected stimulus (a), (b), and (c) in 816, 472, and 62 trials, respectively. The proportion of rewarded trials was significantly different between stimulus (b) and (c) (rewarded trials/non-rewarded trials: 407/65 and 21/41, respectively, *OR* [95% confidence interval (CI)] = 12.13 [6.72, 22.01], $p < .001$). Natsuki selected stimulus (a), (b), and (c) in 919, 565, and 13 trials, respectively. The proportion of rewarded trials was significantly different between stimulus (b) and (c) (477/88 and 2/11, respectively, *OR* = 29.54 [7.07, 189.04], $p < .001$).

Then, we looked at the trials completed before these participants appeared to acquire the ability to distinguish stimulus (b) from stimulus (c) (Fig. 6). We analyzed the first 90 trials completed by Mizuki because after trial 93, the task performance exceeded the chance level for trials containing stimuli (b) and (c) (i.e., first trial in the 20-trial bin of stimuli (b) vs. (c) for which the performance first exceeded the chance level was trial 93). In the first 90 trials, Mizuki selected stimuli (a), (b), and (c) in 29, 26, and 35 trials, respectively. The proportion of rewarded trials was not significantly different between stimuli (b) and (c) (8/18 and 10/25, *OR* = 1.11 [0.33, 3.44], $p = 1$). Regarding the other individual (Natsuki), it was difficult to define the pre-acquisition phase in a similar manner because she achieved above-chance performance quite early. Thus, we tentatively used the first 60 trials completed by Natsuki. In the first 60 trials, Natsuki selected stimuli (a), (b), and (c) in 26, 25, and nine trials, respectively. The proportion of rewarded trials was significantly different between stimulus (b) and (c) (16/9 and 0/9, *OR* = NA [3.1, NA], $p = .001$).

Among the three individuals who did not learn to select stimulus (b) over stimulus (c) (i.e., Iroha, Misaki, and Zamba), the proportion of rewarded trials was significantly different between stimulus (b) and (c) in one individual (Iroha). In total, Iroha selected stimuli (a), (b), and (c) in 649, 506, and 225 trials, respectively. In trials in which she selected stimulus (b) or (c), the proportion of rewarded

trials was significantly different between the two stimuli (263/243 and 73/152, respectively, *OR* = 2.25 [1.62, 3.15], $p < .001$). In contrast, the proportion of rewarded trials was not significantly different between stimuli (b) and (c) in the other two individuals (Misaki and Zamba). Misaki selected stimuli (a), (b), and (c) in 456, 346, and 601 trials, respectively. The proportion of rewarded trials was not significantly different between stimulus (b) and (c) (79/267 and 164/437, respectively, *OR* = 0.79 [0.57, 1.08], $p = .142$). Zamba selected stimuli (a), (b), and (c) in 856, 295, and 249 trials, respectively. The proportion of rewarded trials was not significantly different between stimulus (b) and (c) (133/162 and 131/118, respectively, *OR* = 0.74 [0.52, 1.04], $p = .086$).

We compared the experienced delay length and found a significant difference among the five chimpanzees (n = 336, 243, 428, 479, 264, respectively, $\chi^2[4] = 42.3$, $p < .001$). We then compared the experienced delay between each individual who learned successfully and the three individuals who did not learn successfully, for a total of six tests (i.e., Mizuki vs. Iroha, Misaki, and Zamba, separately; Natsuki vs. Iroha, Misaki, and Zamba, separately). The two successful individuals did not necessarily experience shorter delays compared with the three unsuccessful individuals. One successful individual (Mizuki) experienced significantly longer delays compared with two of the unsuccessful individuals (vs. Iroha: $Z = -4.30$, $p < .001$, Cliff's d = 0.17; vs. Misaki: $Z = -3.75$, $p < .001$, Cliff's d = 0.17), although these delays were not significantly different from those of the third unsuccessful individual (vs. Zamba: $Z = -1.40$, $p = .163$, Cliff's d = −0.06). The other successful individual (Natsuki) experienced significantly longer delays compared with two of the unsuccessful individuals (vs. Iroha: $Z = -3.58$, $p < .001$, Cliff's d = 0.14; vs. Misaki: $Z = -3.07$, $p = .002$, Cliff's d = 0.13), although she experienced significantly shorter delays when compared with the other unsuccessful individual (vs. Zamba: $Z = -2.08$, $p = .037$, Cliff's d = −0.09).

Discussion

We designed a behavioral task to examine solutions to the credit assignment problem (transition-based/transition-free) in nonhuman animals in a non-Markov paradigm, which has been less studied compared with Markov decision processes (Walsh & Anderson, 2014), following a previous study on humans (Tanaka et al., 2009). Using this task, two of five chimpanzees were able to learn to select the stimulus associated with rewards that were carried forward to a later trial and were intervened by other actions. As in Tanaka et al.'s (2009) study, keeping past actions and outcomes in working memory and processing them online seemed implausible because the number of intermediate trials between the choice and the reward varied randomly each time, and was often more than one (Fig. 3). Transition-based assignment was ineffective because state transition was independent of the participant's actions. Therefore, the current results suggest that the two chimpanzees appropriately learned the values of actions leading to rewards that were carried forward to a later trial via a process of transition-free assignment, such as ET.

In this task, the selection of stimulus (b) or (c) could be immediately followed by a reward. However, it is unlikely that the successful individuals chose stimulus (b) over (c) solely because of incidental immediate rewards. For instance, for one successful learner (Mizuki), the proportion of trials followed by a reward was not different between trials in which she selected stimulus (b) versus those in which she selected stimulus (c) in the early part of the session (i.e., her first 90 trials). When we considered all of the trials (i.e., when the dataset included the trials completed after the two

successful individuals learned to constantly select stimulus (b) over stimulus (c)), the proportion of trials followed by a reward was different between trials in which they selected stimulus (b) versus those in which they selected stimulus (c). This difference is likely the result of successful learning by the participants to select stimulus (b). This finding suggests that this individual (Mizuki) was unlikely to learn the task solely via the uncertainty of the reward associated with each stimulus. Rather, especially when focusing on the early part of the session, we can interpret this to mean that stimuli (b) and (c) were not differentially uncertain with regard to the possibility of immediate rewards.

Three chimpanzees were not able to learn to select the stimulus associated with the carry-over reward. It is unlikely that the failure of those chimpanzees was simply caused by the basic features of the task, such as difficulty distinguishing the visual stimuli (i.e., letters) or a lack of motivation for the food reward (i.e., pellets) because those chimpanzees learned to select the stimulus associated with an immediate reward. It is also unlikely that the unsuccessful individuals experienced longer delays compared with the two successful individuals. In fact, we only found this difference when comparing a pair of individuals (Natsuki vs. Zamba), and we found a difference in the opposite direction in four comparisons, although those differences appeared to be subtle. Daily observations of these chimpanzees (see also Sato et al., 2020) revealed that the two successful individuals (Natsuki & Mizuki) usually performed better in touch-panel cognitive tasks compared with the remaining individuals (Iroha, Misaki, & Zamba). Thus, the individual difference we observed was likely to have been caused by a difference in a general learning capacity expressed in such an experimental setting, rather than being specific to this task. In future studies, it would be interesting to examine the individual differences in relation to biological characteristics (e.g., serotonin: Tanaka et al., 2009) and temperament (e.g., compulsive tendency: Sakai et al., 2022), which were found to affect credit assignment in previous studies. Despite these individual differences in chimpanzees, the current study suggests that the human ability to learn using state-transition-free mechanisms (such as the ET) (e.g., Tanaka et al., 2009) is shared with chimpanzees to some extent.

A deeper inquiry is warranted as to why the three chimpanzees failed to acquire a preference for stimulus (b) over stimulus (c), with reference to more general, well-studied learning phenomena. First, the cues present at the time of delayed reward delivery might have overshadowed the contributions of ETs. In particular, stimuli (b) and (c) were presented immediately before the delayed reward delivery, and these stimuli were not directly related to that reward. Relatedly, working memory might have been involved, for example, in feeding information into RL processes (reviewed in Yoo & Collins, 2022). A recent study (Ben-Artzi et al., 2022) suggested that human participants with a higher working memory capacity are less likely to assign credits for rewards to an irrelevant aspect of their actions. The researchers speculated that lower working memory capacity enables credit to simply be assigned to the representation that is activated when selection is made, regardless of its relevance to the outcome. The issue appears to be somewhat inevitable when researchers attempt to design a task in which a reward is delivered in a later, irrelevant trial, as we did in the current study. Nonetheless, the issue might be mitigated by introducing a new pair of stimuli (say stimuli (d) and (e)). Subsequently, the delayed rewards resulting from the selection of stimulus (b) would be delivered in a later trial in which stimuli (d) and (e) were presented, instead of stimuli (b) and (c). This design might reduce the difficulty in discriminating between stimuli (b) and (c) by preventing the credit for the delayed reward from being assigned to the selection of stimulus (c), which is in practice in competition with stimulus (b). This possibility would be worth investigating;

however, this design may introduce an additional cognitive load of remembering and processing a larger number of visual stimuli (i.e., six). In conjunction with this, it might be ideal for a future study to involve a larger number of participants and counterbalance the allocation of the role of the visual stimulus (in our case, capital letters) across participants, which would allow for an analysis of the effect of stimulus counterbalancing.

Second, we omitted rewards resulting from the redundant choices of stimulus (b) to limit the number of available rewards in each trial to one or zero. By implementing this reward-exclusion criterion when stimulus (b) had been selected on multiple trials before stimuli (b) and (c) were presented, the relative value of stimulus (b) might be degraded due to the reduced contingency with a reward. That is, in a session, the number of rewards deriving from selections of stimulus (b) was smaller than the number of the selection of stimulus (b) −naively similar to "partial reinforcement" − unlike the selection of stimulus (a), which led to rewards of equal number to the selection of stimulus (a) (i.e., continuous reinforcement). This reduced reinforcement in stimulus (b) would have lowered the relative value of this stimulus. In addition, the selection of stimulus (c) in a trial of stimulus (b) versus stimulus (c) could immediately be followed by a reward, owing to the past selection of stimulus (b), which could spuriously increase the value of stimulus (c). These combined factors might obscure the difference in values between stimuli (b) and (c), thereby making it difficult to successfully learn to select stimulus (b) over stimulus (c). A possible task design to solve this issue would be to allow multiple delayed rewards to be delivered in a single trial after multiple selections of stimulus (b). Nonetheless, it should be noted that the rule for stimulus (b) and task difficulty may produce a tradeoff. That is, to enable multiple delayed rewards to be delivered in one trial, researchers need to devote additional effort to ensure that participants discriminate different numbers of rewards. In the current study, we placed greater importance on designing an easier task for nonhuman animals. The task should be modified in future studies depending on participants' characteristics and research purposes. If researchers prioritize increasing the relative value of stimulus (b) over reducing task difficulty, it would be worth trying to change the task rule such that multiple delayed rewards could be delivered in a single trial.

Computational modeling also suggested that ET was important for successful learning in this task. An algorithm with ET (i.e., SARSA[λ]) was favored for all of the participants. However, somewhat unexpectedly, this result held true for unsuccessful learners, irrespective of their performance. It is possible that the computational modeling results partially depended on the details of algorithms (Nassar & Frank, 2016). For example, although we eventually specified the initial policy parameters at zero in the models, we also attempted to include those initial values as parameters to be estimated because some participants exhibited a pre-existing bias for particular stimuli from the beginning (e.g., a female chimpanzee (Natsuki) selected stimulus (c) in the first six trials, although this choice did not yield rewards). A previous study (Lehmann et al., 2019) also reported initial biases in participant actions, and thus specified the initial action values at certain fixed values. When there is such a pre-existing bias, initialization of policy parameters at zero has been suggested to yield spurious changes (i.e., "pseudo-learning") (Katahira et al., 2017). Although, in the course of model specification, we decided not to include the initial values as parameters (which did not improve model convergence drastically) for the sake of simplicity, those initial attempts (not reported) implied that the modeling results could be affected by algorithmic details. This was also true for specification of prior distributions for some parameters (i.e., α and β).

Generally speaking, model fitting can produce only relative results (Palminteri et al., 2017; Wilson & Collins, 2019). Countless other possible algorithms could comprise the ET. Thus, we did not conclude that SARSA(λ) is the appropriate model for behavior in our task. We concluded only that a model that considers the ET would be relatively better than models without the ET. Moreover, an advantage of our approach was that performance in the task per se indicated the involvement of ET (as in Lehmann et al., 2019).

We designed a simple task to examine learning in nonhuman animals using state-transition-free mechanisms. However, some other cognitive processes might have been involved in learning. Specifically, the current study did not address concurrent memory processes in detail because we aimed to examine ET as a putative RL mechanism. Specifically, it remains unclear whether the two successful individuals associated the selection of stimulus (b) with a specific reward delivered in a later trial, or if they associated it with an overall increase in reward density. Although ETs can be recruited in either type of learning, further examination of this issue may be helpful for clarifying the cognitive processes involved in this task. To this end, it might be useful to change the rule of the delayed reward after the participant's preference for stimulus (b) over stimulus (c) had been established, and to observe the participant's behavioral responses to this sudden rule change. For instance, a future study could slightly delay the timing of the delayed reward delivery from the original timing. A recent touchscreen-task study (Tomonaga et al., 2023) delayed the task feedback to examine the reaction of chimpanzees. They found that chimpanzees looked up at a food dispenser after making a response more often when they made a correct choice compared with when they made a wrong choice. This behavior indicated their confidence that they would receive a food reward when they had made a correct choice. Beran et al. (2015) examined chimpanzees' behaviors in computerized tasks when the reward delivery was delayed and occurred in another enclosure. After making a correct response, chimpanzees moved to that enclosure to obtain a food reward, which the authors interpreted as reflecting chimpanzees' confidence about their correctness. It is possible that other subtler behavioral differences might also emerge, such as changes in reaction times and anticipatory-consummatory oral movements (i.e., licking for liquid reward) as observed in monkeys (Watanabe et al., 2001). Also, it would be useful try to extinguish the stimulus (b) contingency to observe whether chimpanzees would change their behaviors (cf. frustrative nonreward, Amsel, 1958; see Papini et al., 2022, for a recent review) such as reducing their selection of stimulus (b). This observation could indicate whether the delayed reward induced the chimpanzees' selection of stimulus (b) over stimulus (c). Chimpanzees might exhibit other behavioral responses that hint at their anticipation of a food reward. For example, in a previous study (Beran, 2001), chimpanzees requested a food reward from the human experimenter by exhibiting gestures, touching on the mesh, and pursing their lips more often in trials in which they made a correct response but the reward was omitted compared with trials in which they made an incorrect response. The author speculated that these types of behaviors reflet chimpanzees' expectation about upcoming food rewards, owing to their correct responses. Chimpanzees might exhibit behavioral signs of negative affect in response to the disappearance of a contingent reward. For example, there is a preliminary report that a chimpanzee appeared to scratch herself more often when no feedback was given for her response (although the author did not distinguish correct vs. wrong responses) compared with when auditory feedback was given for a correct response (Itakura, 1993). In a risky choice situation (Rosati & Hare, 2013), chimpanzees scratched themselves, banged the bar or mesh, and emitted vocalizations more often when confronted with the bad outcome compared with the good outcome. Observations of

these type of behaviors could provide insight regarding the extent to which chimpanzees expect a delayed reward to be delivered in a specific trial.

In their daily lives, animals often need to take multiple actions before they receive benefits, and they do not always have access to the information that is necessary to guide optimal behaviors (Seo & Lee, 2010). Solutions equipped with ETs are adaptive in situations involving delayed rewards and violating the Markov property (Sutton & Barto, 2018). In conclusion, we designed a behavioral task to investigate how nonhuman animal learning systems solve the credit assignment problem using ETs. We believe that this task is applicable to a wide range of animal species. Future comparative psychology research taking this approach may provide valuable insight into the timescale of learning abilities from an evolutionary perspective. Moreover, this task could be a common currency task (Pike et al., 2021) to test various species of animals and better understand serotonergic system deficits (Tanaka et al., 2009) and OCD (Sakai et al., 2022). The current study provided data from chimpanzees, one of our evolutionarily closest relatives, taking the first step toward elucidating those issues.

Acknowledgements

We thank the staff and researchers at Kumamoto Sanctuary for their help with the study, particularly Dr. N. Morimura and Dr. F. Kano. We thank Benjamin Knight, MSc., from Edanz (https://jp.edanz.com/ac) for editing a draft of this manuscript.

References

- Akam, T., Rodrigues-Vaz, I., Marcelo, I., Zhang, X., Pereira, M., Oliveira, R. F., Dayan, P., & Costa, R. M. (2021). The anterior cingulate cortex predicts future states to mediate model-based action selection. Neuron, 109(1), 149-163.e7. https://doi. org/10.1016/j.neuron.2020.10.013
- Amsel, A. (1958). The role of frustrative nonreward in noncontinuous reward situations. Psychological Bulletin, 55(2), 102−119. https:// doi. org/10. 1037/h0043 125
- Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Transactions on Systems, Man and Cybernetics, SMC- 13(5), 834−846. https://doi. org/10. 1109/TSMC. 1983. 63130 77
- Ben-Artzi, I., Luria, R., & Shahar, N. (2022). Working memory capacity estimates moderate value learning for outcome-irrelevant features. Scientific Reports, 12, 19677. https://doi. org/10. 1038/ s41598- 022- 21832-x
- Beran, M. J. (2001). Do chimpanzees have expectations about reward presentation following correct performance on computerized cognitive testing? The Psychological Record, 51(2), 173−183. https:// doi. org/10. 1007/BF033 95393
- Beran, M. J., Perdue, B. M., Futch, S. E., Smith, J. D., Evans, T. A., & Parrish, A. E. (2015). Go when you know: Chimpanzees' confidence movements reflect their responses in a computerized memory task. Cognition, 142, 236−246. https://doi. org/10. 1016/j. cogni tion. 2015. 05. 023

- Bogacz, R., McClure, S. M., Li, J., Cohen, J. D., & Montague, P. R. (2007). Short-term memory traces for action bias in human reinforcement learning. Brain Research, 1153(1), 111−121. https:// doi. org/10. 1016/j. brain res. 2007. 03. 057
- 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
- Eckstein, M. K., Wilbrecht, L., & Collins, A. G. E. (2021). What do reinforcement learning models measure? Interpreting model parameters in cognition and neuroscience. Current Opinion in Behavioral Sciences, 41, 128−137. https://doi. org/10. 1016/j. cobeha. 2021. 06. 004
- Fay, M. (2010). Confidence intervals that match Fisher's exact or Blaker's exact tests. Biostatistics, 11(2), 373−374. https://doi. org/10. 1093/biost atist ics/kxp050
- Gabry, J., & Češnovar, R. (2021). cmdstanr: R Interface to "CmdStan." https://mc- stan. org/cmdst anr
- Gerstner, W., Lehmann, M., Liakoni, V., Corneil, D., & Brea, J. (2018). Eligibility traces and plasticity on behavioral time scales: Experimental support of neoHebbian three-factor learning rules. Frontiers in Neural Circuits, 12, 53. https://doi. org/10. 3389/fncir. 2018. 00053
- Gureckis, T. M., & Love, B. C. (2009). Short-term gains, long-term pains: How cues about state aid learning in dynamic environments. Cognition, 113(3), 293−313. https://doi. org/10. 1016/j. cogni tion. 2009. 03. 013
- Hothorn, T., Hornik, K., van de Wiel, M. A., & Zeileis, A. (2006). A Lego system for conditional inference. The American Statistician, 60(3), 257−263. https://doi. org/10. 1198/00031 3006X 118430 Itakura, S. (1993). Emotional behavior during the learning of a contingency task in a chimpanzee. Perceptual and Motor Skills, 76(2), 563−566. https://doi. org/10. 2466/pms. 1993. 76.2. 563
- Jocham, G., Brodersen, K. H. H., Constantinescu, A. O. O., Kahn, M. C. C., Ianni, A. M., Walton, M. E. E., Rushworth, M. F. F. S., & Behrens, T. E. E. J. (2016). Reward-guided learning with and without causal attribution. Neuron, 90(1), 177−190. https://doi. org/10. 1016/j. neuron. 2016. 02. 018
- Katahira, K. (2018). Kodo deta no keisanron moderingu—Kyoka gakusyu moderu wo rei toshite— [Computational Modeling of Behavioral Data]. Ohmsha.
- Katahira, K., Yu, B., & Nakao, T. (2017). Pseudo-learning effects in reinforcement learning model-based analysis: A problem of misspecification of initial preference. PsyArXiv. https://doi. org/ 10. 31234/osf. io/a6hzq
- Lehmann, M. P., Xu, H. A., Liakoni, V., Herzog, M. H., Gerstner, W., & Preuschoff, K. (2019). One-shot learning and behavioral eligibility traces in sequential decision making. eLife, 8, e47463. https://doi. org/10. 7554/eLife. 47463
- Minsky, M. (1961). Steps toward artificial intelligence. Proceedings of the IRE, 49(1), 8−30. https://doi. org/10. 1109/JRPROC. 1961. 287775
- Nassar, M. R., & Frank, M. J. (2016). Taming the beast: Extracting generalizable knowledge from computational models of cognition. Current Opinion in Behavioral Sciences, 11, 49−54. https://doi. org/10. 1016/j. cobeha. 2016. 04. 003

- Nonomura, S., Nishizawa, K., Sakai, Y., Kawaguchi, Y., Kato, S., Uchigashima, M., ⋯, Kimura, M. (2018). Monitoring and updating of action selection for goal-directed behavior through the striatal direct and indirect pathways. Neuron, 99(6), 1302− 1314.e5. https://doi. org/10. 1016/j. neuron. 2018. 08. 002
- Nosarzewska, A., Peng, D. N., & Zentall, T. R. (2021). Pigeons acquire the 1-back task: Implications for implicit versus explicit learning? Learning & Behavior. https://doi. org/10. 3758/ s13420- 021- 00468-3
- 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
- Papini, M. R., Guarino, S., Hagen, C., & Torres, C. (2022). Incentive disengagement and the adaptive significance of frustrative nonreward. Learning & Behavior, 50(3), 372−388. https://doi. org/10. 3758/s13420- 022- 00519-3
- Pike, A. C., Lowther, M., & Robinson, O. J. (2021). The importance of common currency tasks in translational psychiatry. Current Behavioral Neuroscience Reports, 8(1), 1−10. https://doi. org/ 10. 1007/s40473- 021- 00225-w
- R Core Team. (2021). R: A laguage and environment for statistical computing. R Foundation for Statistical Computing. https:// www.r- proje ct. org/
- Redish, A. D., Kepecs, A., Anderson, L. M., Calvin, O. L., Grissom, N. M., Haynos, A. F., ⋯, Zilverstand, A. (2022). Computational validity: Using computation to translate behaviours across species. Philosophical Transactions of the Royal Society B, 377(1844), 20200525. https://doi. org/10. 1098/rstb. 2020. 0525
- Rosati, A. G., & Hare, B. (2013). Chimpanzees and bonobos exhibit emotional responses to decision outcomes. PLoS ONE, 8(5), e63058. https://doi. org/10. 1371/journ al. pone. 00630 58
- Rummery, G. A., & Niranjan, M. (1994). On-line Q-learning using connectionist systems. CUED/F-INFENG/TR 166. Cambridge University Engineering Department.
- Sakai, Y., Sakai, Y., Abe, Y., Narumoto, J., & Tanaka, S. C. (2022). Memory trace imbalance in reinforcement and punishment systems can reinforce implicit choices leading to obsessivecompulsive behavior. Cell Reports, 40(9), 111275. https://doi. org/10. 1016/j. celrep. 2022. 111275
- Sato, Y., Sakai, Y., & Hirata, S. (2020). Computerized intertemporal choice task in chimpanzees (Pan troglodytes) with/without postreward delay. Journal of Comparative Psychology, 135(2), 185−195. https://doi. org/10. 1037/com00 00254
- Scholl, J., & Klein-Flügge, M. (2018). Understanding psychiatric disorder by capturing ecologically relevant features of learning and decision-making. Behavioural Brain Research, 355, 56−75. https://doi. org/10. 1016/j. bbr. 2017. 09. 050
- Schultz, W. (1997). Dopamine neurons and their role in reward mechanisms. Current Opinion in Neurobiology, 7(2), 191−197. https://doi. org/10. 1016/S0959- 4388(97) 80007-4
- Seo, H., & Lee, D. (2010). Orbitofrontal cortex assigns credit wisely. Neuron, 65(6), 736− 738. https://doi. org/10. 1016/j. neuron. 2010. 03. 016
- Shen, W., Flajolet, M., Greengard, P., & Surmeier, D. J. (2008). Dichotomous dopaminergic control of striatal synaptic plasticity. Science, 321(5890), 848−851. https://doi. org/10. 1126/scien ce. 11605 75

- Singh, S. P., & Sutton, R. S. (1996). Reinforcement learning with replacing eligibility traces. Machine Learning, 22(1−3), 123−158. https://doi. org/10. 1007/BF001 14726
- Smith, J. D., Jackson, B. N., & Church, B. A. (2020). Monkeys (Macaca mulatta) learn two-choice discriminations under displaced reinforcement. Journal of Comparative Psychology, 134(4), 423−434. https://doi. org/10. 1037/com00 00227
- Stan Development Team. (2019). Stan User's Guide 2.21.
- Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. Machine Learning, 3(1), 9−44. https://doi. org/10. 1023/A: 10226 33531 479
- Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). The MIT press. https://lccn. loc. gov/20180 23826
- Tanaka, S. C., Shishida, K., Schweighofer, N., Okamoto, Y., Yamawaki, S., & Doya, K. (2009). Serotonin affects association of aversive outcomes to past actions. Journal of Neuroscience, 29(50), 15669− 15674. https://doi. org/10. 1523/JNEUR OSCI. 2799- 09. 2009
- Tartaglia, E. M., Clarke, A. M., & Herzog, M. H. (2017). What to choose next? A paradigm for testing human sequential decision making. Frontiers in Psychology, 8, 312. https://doi. org/10. 3389/ fpsyg. 2017. 00312
- Tomonaga, M., Kurosawa, Y., Kawaguchi, Y., & Takiyama, H. (2023). Don't look back on failure: Spontaneous uncertainty monitoring in chimpanzees. Learning & Behavior. https://doi. org/10. 3758/ s13420- 023- 00581-5
- Torchiano, M. (2020). effsize: Efficient effect size computation. https://doi. org/10. 5281/zenodo. 14806 24, R package version 0.8.1 (https://CRAN.R- proje ct. org/packa ge= effsi ze).
- Walsh, M. M., & Anderson, J. R. (2011). Learning from delayed feedback: Neural responses in temporal credit assignment. Cognitive, Affective and Behavioral Neuroscience, 11(2), 131−143. https:// doi. org/10. 3758/s13415- 011- 0027-0
- Walsh, M. M., & Anderson, J. R. (2014). Navigating complex decision spaces: Problems and paradigms in sequential choice. Psychological Bulletin, 140(2), 466−486. https://doi. org/10. 1037/ a0033 455
- Watanabe, M., Cromwell, H. C., Tremblay, L., Hollerman, J. R., Hikosaka, K., & Schultz, W. (2001). Behavioral reactions reflecting differential reward expectations in monkeys. Experimental Brain Research, 140, 511−518. https://doi. org/10. 1007/s0022 10100 856
- Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. Machine Learning, 8(3−4), 279−292.
- Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. Elife, 8, e49547. https:// doi. org/10. 7554/eLife. 49547
- Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G. C. R., Urakubo, H., Ishii, S., & Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. Science, 345(6204), 1616−1620. https://doi. org/10. 1126/scien ce. 12555 14
- Yoo, A. H., & Collins, A. G. E. (2022). How working memory and reinforcement learning are intertwined: A cognitive, neural, and computational perspective. Journal of Cognitive Neuroscience, 34(4), 551−568. https://doi. org/10. 1162/jocn_a_ 01808
- Zentall, T. R., Peng, D. N., & Mueller, P. M. (2022). 1-Back reinforcement matching and mismatching by pigeons: Implicit or explicit learning? Behavioural Processes, 195, 104562. https://doi. org/10. 1016/j. beproc. 2021. 104562

Funding

This study was supported financially by the Ministry of Education, Culture, Sports, Science, Japan Society for the Promotion of Science to YSato (grant number 19J22889), to SH (grant numbers 26245069, 18H05524, 23H00494), and to TM (grant number 16H06283); a Program for Leading Graduate Schools to TM (U04); and the Great Ape Information Network.

Declarations

Ethics approval

Animal husbandry and research protocols complied with the Guide for Animal Research Ethics provided by the Wildlife Research Center, Kyoto University (No. WRC-2020-KS006A). For human participants (Online Supplementary Materials (OSM)), the research protocol was approved by the Ethics Committee of the Unit for Advanced Study of Mind at Kyoto University (2-P-16).

Consent to participate

Informed consent was obtained from all individual human participants included in the study (OSM).

Consent for publication

Human participants (OSM) signed informed consent that included publishing their data.

Table 1 Participant demographics

Name	**Sex**	**Age (years)**	**GAIN ID[a]**
Iroha	Female	11	0708
Misaki	Female	21	0593
Mizuki[b]	Female	23	0559
Natsuki	Female	15	0677
Zamba	Male	25	0543

In the “Age” column, participants’ ages as of the start of the study are shown

[a] Great Ape Information Network (GAIN: https://shigen.nig.ac.jp/gain/) provides detailed information for those chimpanzees

[b] This chimpanzee was reared by human caretakers while also having regular opportunities for interactions with conspecific peers

Table 2 Model comparison on the basis of widely applicable information criterion (WAIC)

Name	SARSA		Q-learning		Actor-Critic	
	w/ ET	**w/o ET**	**w/ ET**	**w/o ET**	**w/ ET**	**w/o ET**
Iroha	**0.394**	0.437	0.415	0.415	0.511	0.510
Misaki	**0.513**	0.651	0.587	0.644	0.588	0.652
Mizuki	**0.174**	0.212	0.218	0.218	0.209	0.225
Natsuki	**0.136**	0.171	0.173	0.174	0.177	0.280
Zamba	**0.302**	0.332	0.330	0.331	0.336	0.339

For each participant (shown in each row), the lowest WAIC is shown in bold. The names of participants who succeeded in learning are shown in bold

Table 3 Parameter estimates via the SARSA(λ) model for each participant

Name	Maximum a posteriori estimate [95% credible intervals]			
	α	**β**	**γ**	**λ**
Iroha	0.16 [0.11, 0.20]	0.23 [0.20, 0.32]	1.00 [0.99, 1.00]	0.94 [0.92, 0.96]
Misaki	0.11 [0.09, 0.13]	0.74 [0.61, 0.91]	0.99 [0.98, 0.99]	0.98 [0.97, 0.99]
Mizuki	0.23 [0.19, 0.30]	0.31 [0.23, 1.13]	0.99 [0.94, 0.99]	0.91 [0.88, 0.93]
Natsuki	0.50 [0.41, 0.60]	2.17 [1.80, 2.62]	0.87 [0.82, 0.90]	0.87 [0.79, 0.94]
Zamba	0.24 [0.18, 0.40]	0.46 [0.32, 0.74]	0.99 [0.98, 1.00]	0.90 [0.82, 0.94]

The names of participants who succeeded in learning are shown in bold

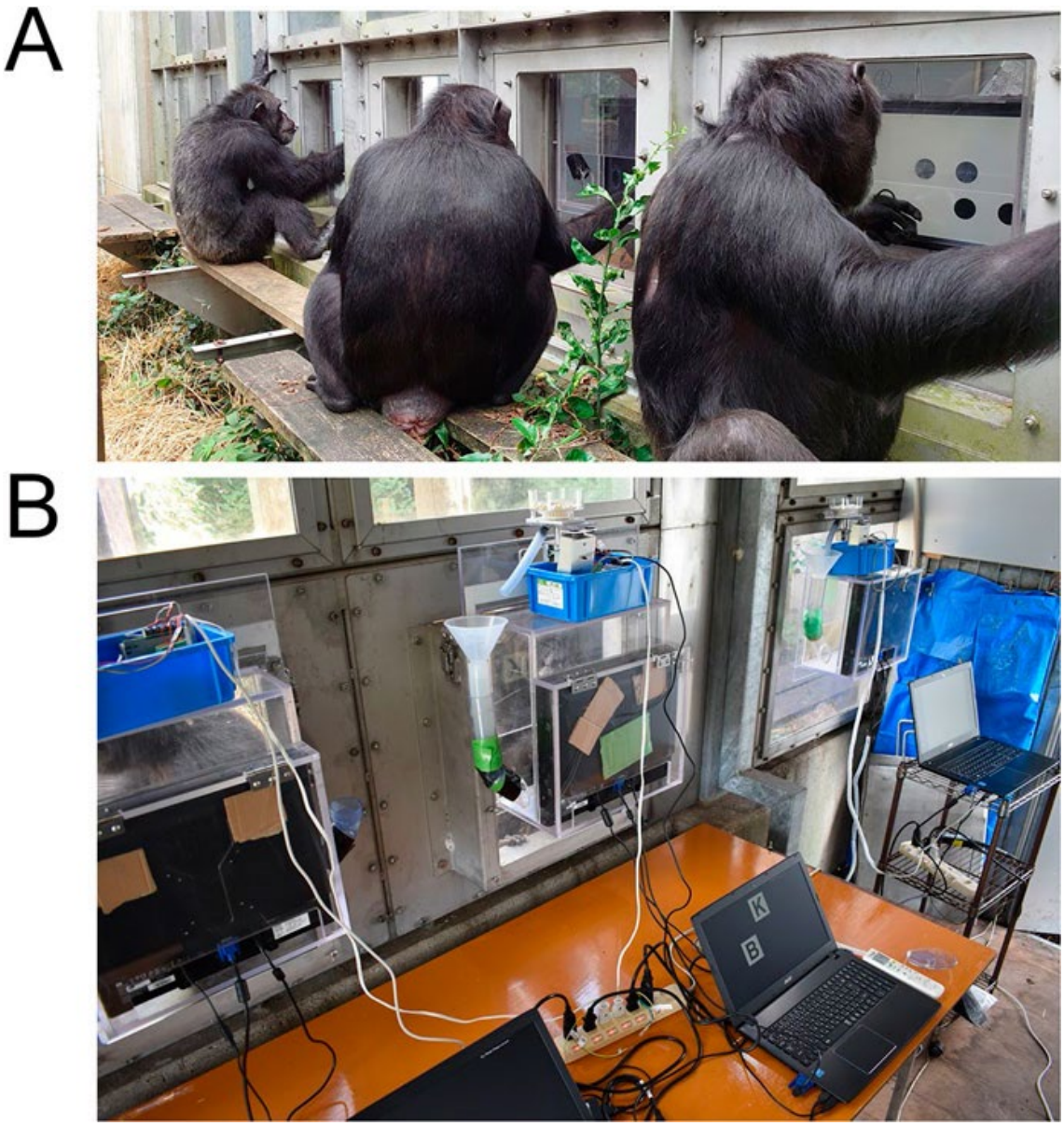

Fig. 1
Experimental setting A from the chimpanzee enclosure and B from inside the experimental booth.
Note. The chimpanzee shown at the right in A was engaging in an irrelevant task for distraction (see the Method section, Procedure)

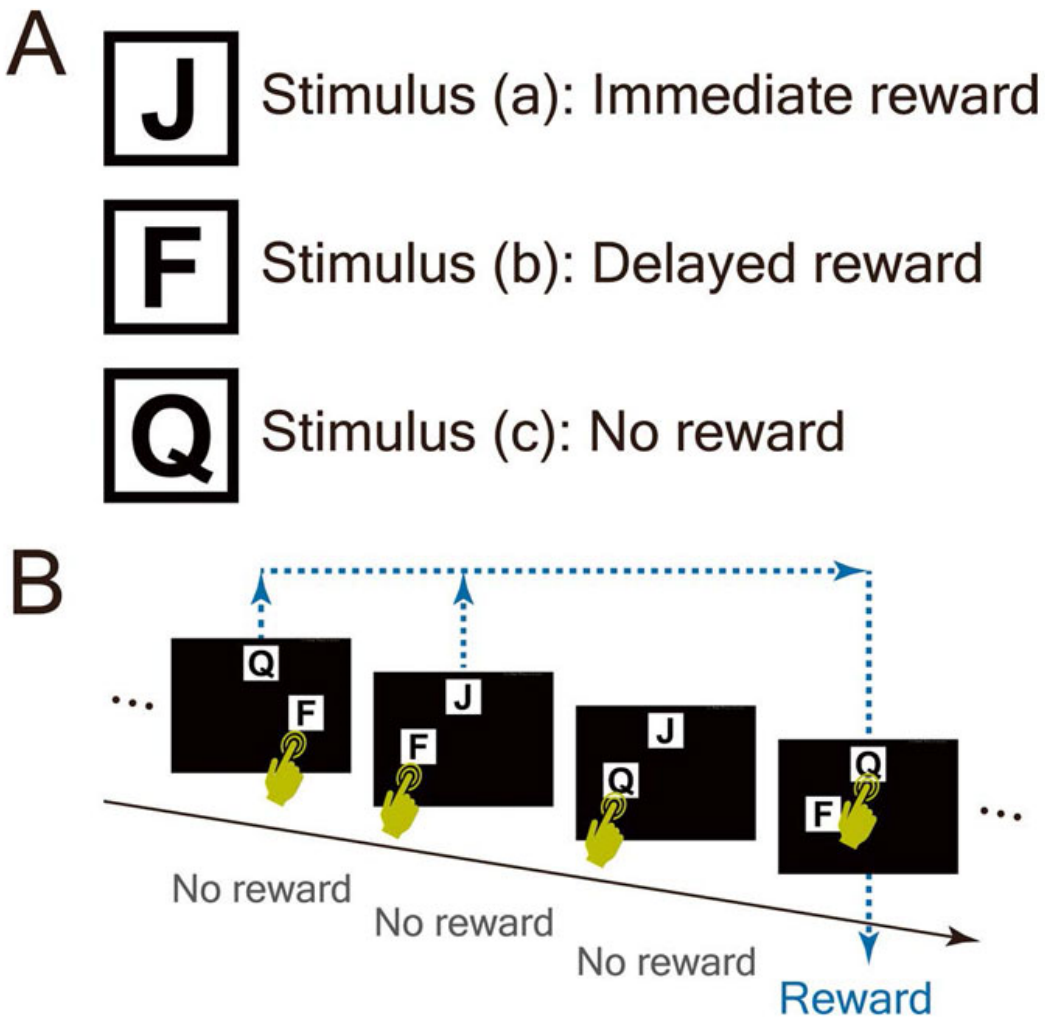

Fig. 2
A Visual stimuli used to represent stimuli (a), (b), and (c). **B** Schematic illustration of the rule of the delayed rewards. *Note.* The hand icon (taken from Microsoft PowerPoint) represents the choices of a hypothetical participant. In this example, stimulus (b) was selected in the leftmost trial (trial *t*). In trial (*t* + 1), stimulus (b) is again selected. In trial (*t* + 2), stimulus (c) (represented by capital "Q") is selected, which does not lead to any reward. Finally, in trial (*t* + 3), stimulus (c) is selected, which does not lead to any reward by itself. However, stimuli (b) and (c) are presented in that trial. Thus, a delayed reward owing to a previous choice of stimulus (b) was delivered. Note that only one reward is delivered in trial (*t* + 3), although stimulus (b) has been selected twice before that trial. Also, note that stimuli (b) and (c) are presented in trial *t* too, but for simplicity, we suppose that stimulus (b) has not been chosen before that trial

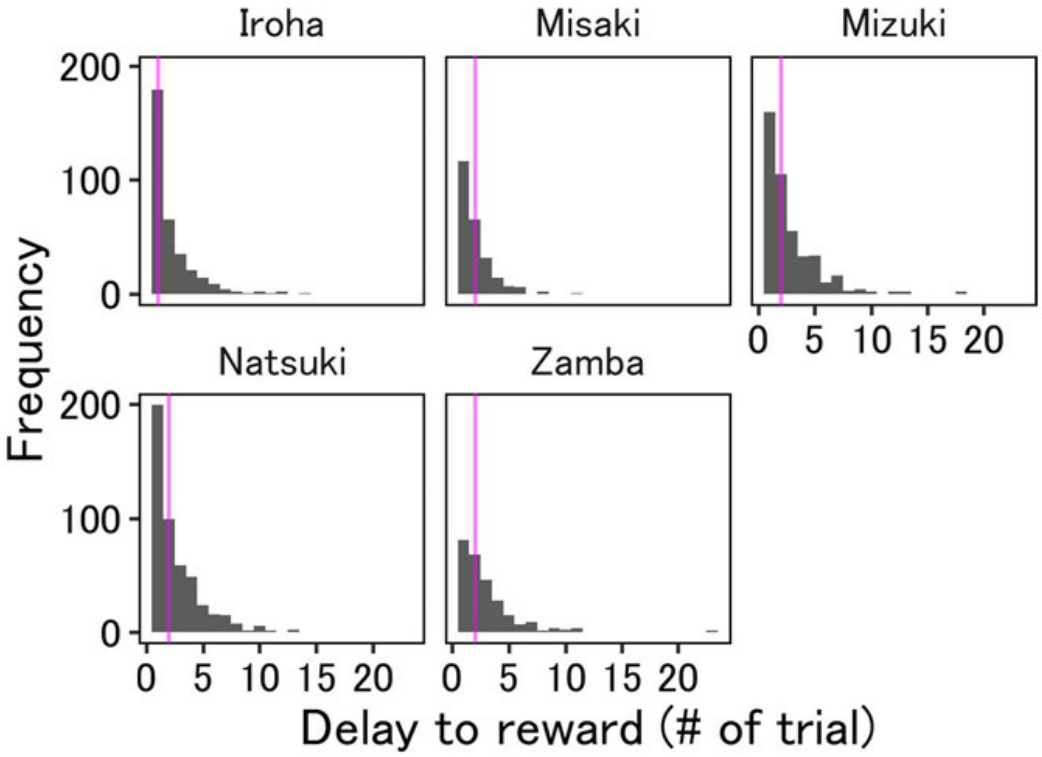

Fig. 3
Histograms showing the actual delay (number of trials) that chimpanzee participants experienced between a choice of stimulus (b) and the delivery of its reward. *Note.* Magenta (light gray) vertical line shows the median for each participant. Note that when stimulus (b) is selected in multiple trials before a delayed-reward delivery, the delay is counted only from the latest choice of stimulus (b). Also, note that the invalid five trials (see the Method section, Data Analysis) were considered in the calculation of those delays, whereas those data were omitted from analysis

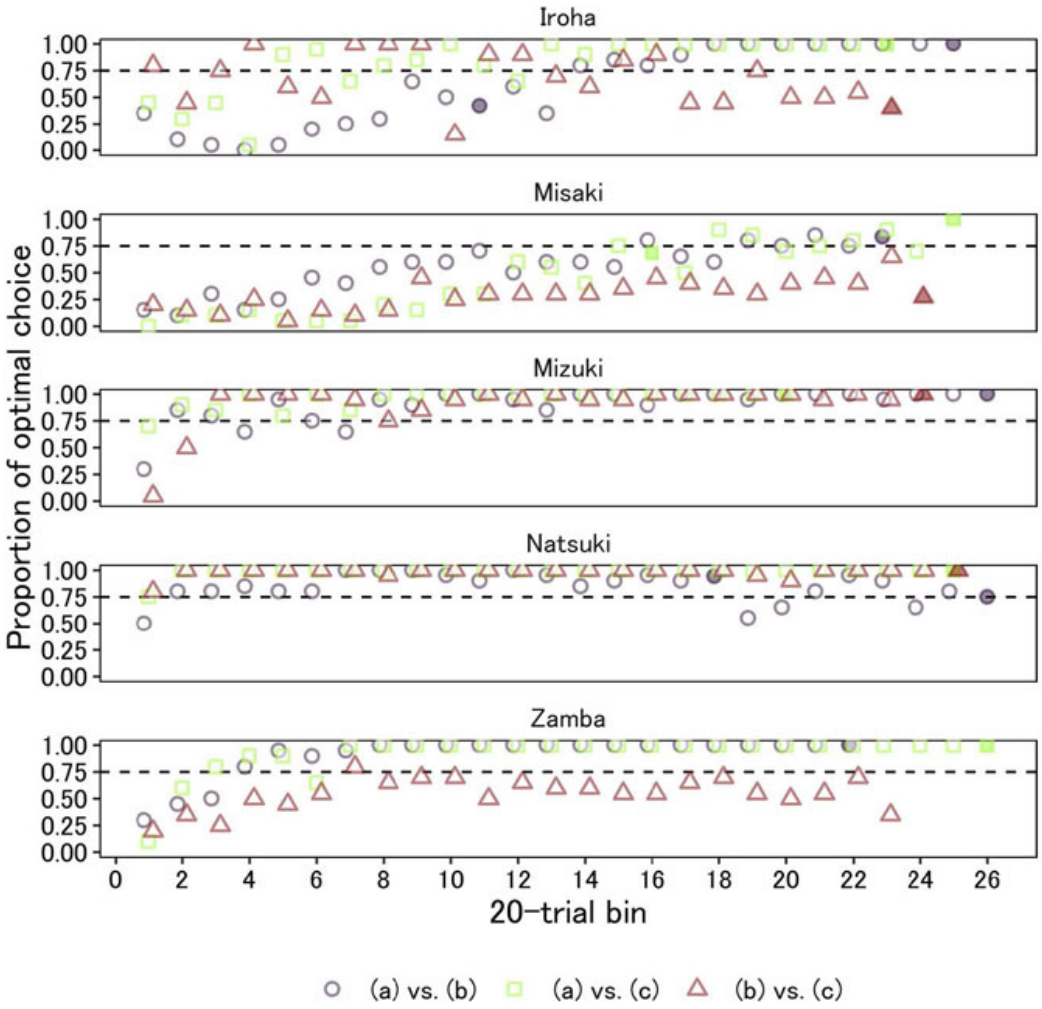

Fig. 4
Proportion of correct choices in 20-trial bins for chimpanzee participants. Purple (dark gray) circles, green (light gray) squares, and red (dark gray) triangles show data from trials in which stimuli (a) vs. (b), (a) vs. (c), or (b) vs. (c) were presented, respectively. Each panel shows one participant's data. Dotted horizontal lines show the border of significance by binomial test (15/20 = 0.75). *Note.* The optimal choices are to select stimulus (a) in trials in which stimulus (a) is presented and to select stimulus (b) in trials in which stimuli (b) and (c) are presented. The dots on or above the dotted line (0.75) indicate significantly higher performance than that predicted by chance (0.5) for bins comprising 20 trials (shown in open dots) but this is not necessarily so for bins comprising fewer trials (shown as filled dots)

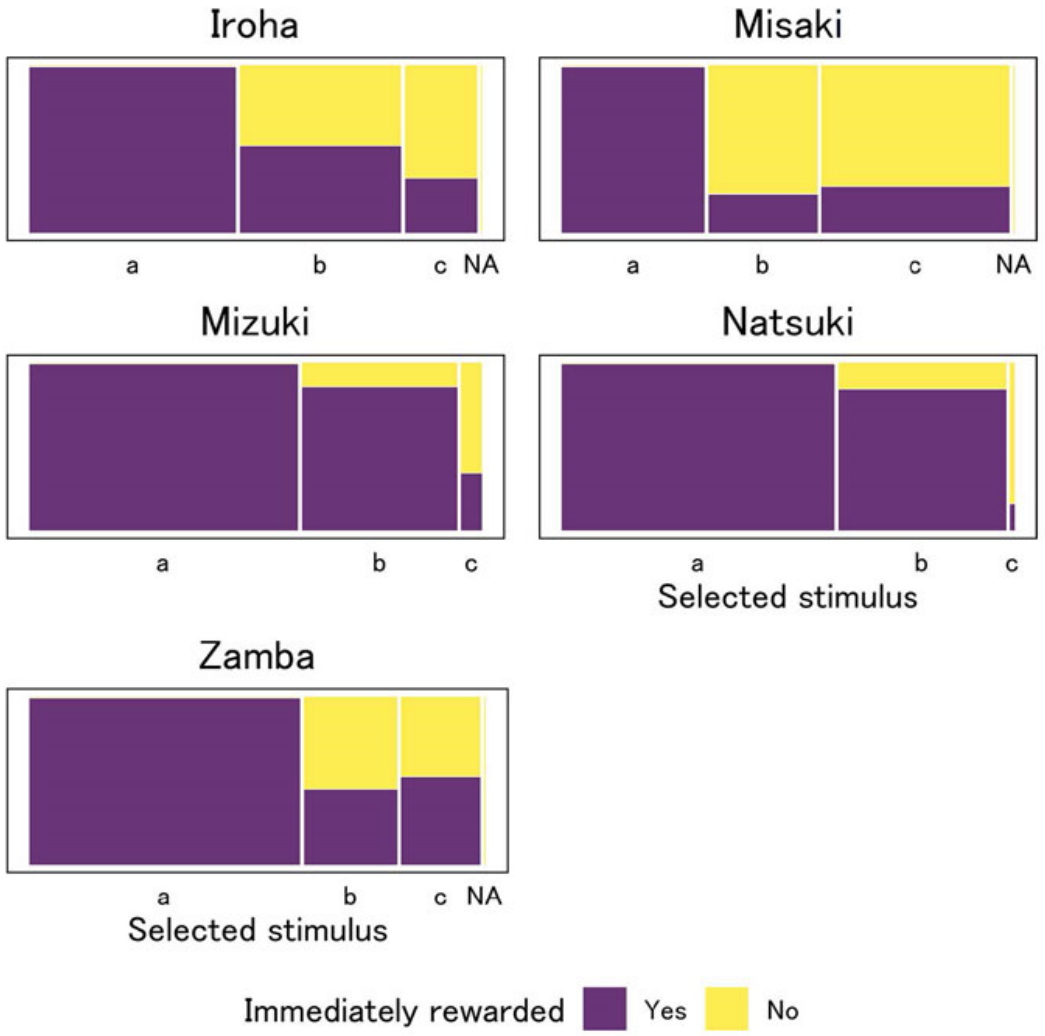

Fig. 5
Mosaic plot showing the proportion of trials in which the choice of each stimulus was immediately followed by a reward. *Note*. When participants selected stimulus (a) in a trial, this choice always led to a reward at the end of that trial. When participants selected stimulus (b) or (c) in a trial, this choice itself did not lead to a reward at the end of that trial, but this choice could be followed by a reward at the end of that trial owing to a past choice of stimulus (b). The width of bars in each panel varied in relation to the proportion of trials in which each stimulus was selected. NA refers to trials in which participants did not make a choice for 30 s

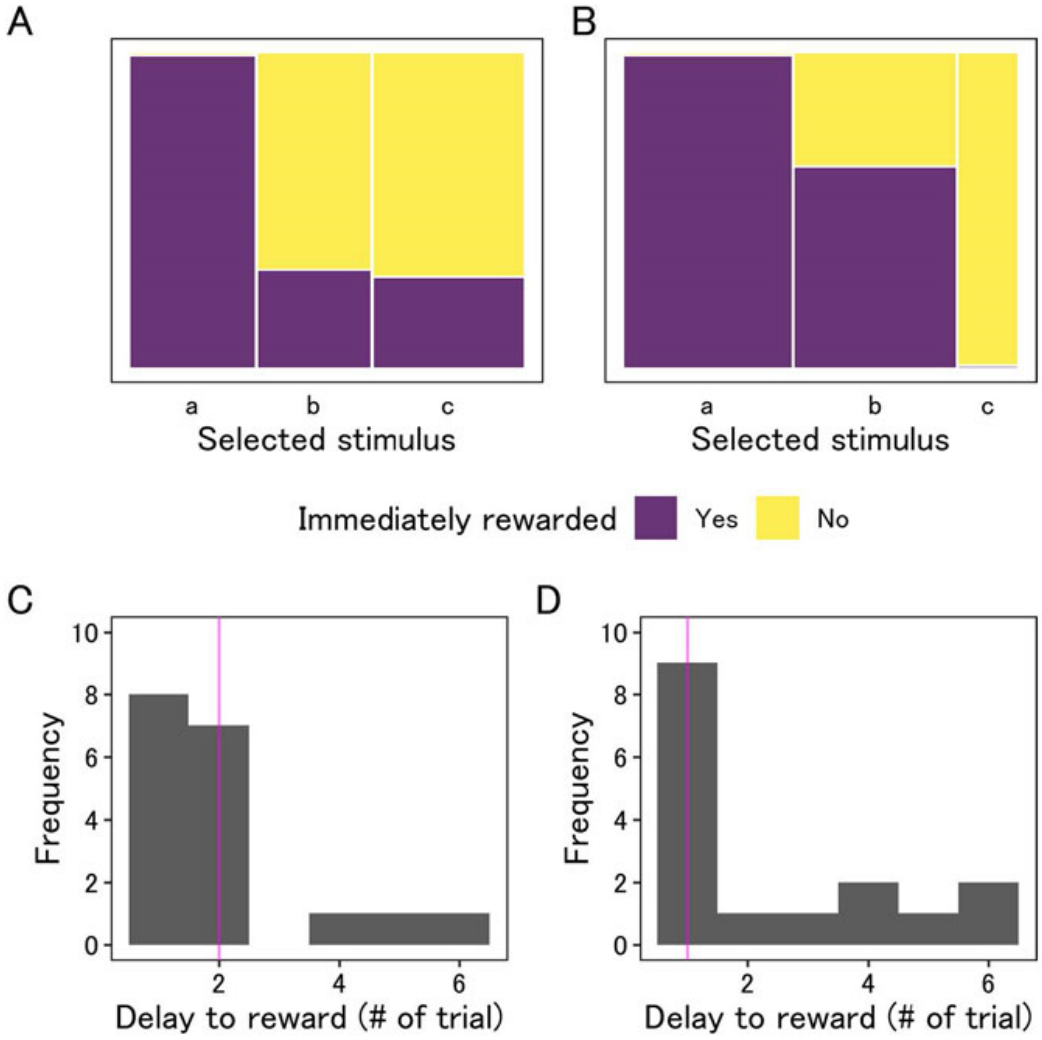

Fig. 6

Mosaic plot showing the proportion of trials in which the choice of each stimulus was immediately followed by a reward in Mizuki (**A**) and Natsuki (**B**) before they achieved high performance. Histograms showing the actual delay (number of trials) that Mizuki (**C**) and Natsuki (**D**) experienced when selecting stimulus (b) before they achieved high performance. *Note.* In mosaic plots (**A** and **B**), the width of bars in each panel varied in relation to the proportion of trials in which each stimulus was selected. In histograms (**C** and **D**), magenta (light gray) vertical lines show the median for each participant.