

Inferring source of learning by chimpanzees in cognitive tasks using reinforcement learning theory

Satoshi Hirata¹, Yutaka Sakai²

¹ Wildlife Research Center, Kyoto University, Kyoto, Japan

² Brain Science Institute, Tamagawa University, Tokyo, Japan

Abstract

Reinforcement learning is a mathematical framework for learning better choices through trial-and-error. Recent studies revealed that reinforcement learning is applicable to animal behavior and cognition. However, applying reinforcement learning to animal behavior sometimes encounters difficulties because the information sources utilized by animals to make choices are often unknown, whereas this is identified as the “state” in the reinforcement learning framework. We sought to identify possible state settings including non-standard formulations suitable for explaining data from past chimpanzee studies. Although chimpanzees’ performance in a serial learning task was inconsistent with standard reinforcement learning formulations, we found that the combination of state-independent choice making and state-dependent evaluation produced consistent results. Exploration of state settings in reinforcement learning may shed new light on animal learning processes.

Keywords Reinforcement learning · Chimpanzee · Serial learning · Actor-Critic

1. Introduction

Reinforcement learning theory is a mathematical framework that can enable a computer or artificial intelligence to learn better choices to achieve goals through trial and error [1]. Previous studies have indicated that reinforcement learning theory also provides an effective approach for understanding animal learning processes [2,3,4,5,6,7]. In particular, temporal difference (TD) learning, an established class of reinforcement learning algorithms, has been applied to investigate the mechanisms of learning in the human and nonhuman animal brain [4, 8,9,10,11,12]. In such cases, reinforcement learning theory offers a constructive approach for making inferences about the way in which information is processed in the brain by comparing the performance of a mathematical model with actual behaviors.

Despite several studies using this approach successfully, the application of reinforcement learning to animal behavior sometimes encounters difficulty because the potential information sources utilized by animals for making choices and learning are often unknown. In the reinforcement learning framework, the source is considered as the “state,” and it is generally determined by a human programmer [1]. However, many different potential information sources are available in real-life

behavioral situations, which means that determining a unique state can be difficult. It is important to identify various possible state settings when addressing actual animal behavior, which may involve deviations from the standard form of the reinforcement learning algorithm.

In this paper, we attempt to provide a new perspective by examining possible state settings for explaining outcomes from past chimpanzee studies on learning. The data and observations described in the following sections were obtained by SH in the mid-late 2000s.

2. Background information

2.1 *Serial learning of Arabic numerals*

Some animals are capable of serial learning (i.e., learning the order of series of items), including nonhuman primates: chimpanzees are capable of learning how to touch Arabic numerals in ascending order (e.g., 1, 2, 3, 4) in a computer-controlled touch-screen task [13, 14]. On the basis of these studies, some researchers have claimed that chimpanzees have an understanding of the ordinal relationship of numerals [15].

At Hayashibara Great Ape Research Institute, Japan, SH trained several chimpanzees to use a computer-controlled touch-screen to learn a sequence of Arabic numerals (see [16] for an overview of research at this institute; the data for this particular task are unpublished). After some basic training to learn how to touch visual stimuli on the monitor, the chimpanzees were then trained to touch “1” first, followed by “2.” Once they had learned the order of “1–2”, the task proceeded to “1–2–3”, followed by “1–2–3–4” and so forth.

Figure 1 illustrates the flow of the task with the example sequence of “1–2–3–4.”; the methods followed previous studies on chimpanzees’ serial learning of Arabic numerals [13, 14]. The numerals “1,” “2,” “3,” and “4” are presented on a screen at random locations, and the subject is required to touch them in the order “1–2–3–4.” Each numeral touched immediately disappears from the screen and the task proceeds to the next step or ends, depending on the choice. For example, if the subject touches “1” first, then “1” disappears while “2,” “3,” and “4” remain on the screen. Then, if the subject touches “2”, “2” disappears while “3” and “4” remain on the screen, and so on until the only option the subject can choose is “4.” A food reward is delivered after successfully touching these numerical items on the screen. However, if the subject touches “1” first followed by “3”, for example, this constitutes an error, in which case all numerals on the screen disappear and no food reward is delivered.

Figure 2 shows learning curves of the three chimpanzee subjects. In general, their performance declined steeply when a new numerical item was introduced. Although not represented in Fig. 2, transfer from the “1–2–3” to the “1–2–3–4” task was intermingled with the “3–4” task. That is, after the chimpanzees had learned the sequence “1–2–3”, they were presented with the task “3–4”, either before, after, or simultaneously with presentation of the task “1–2–3–4”. The experimenter (SH) did not have a predetermined criterion for proceeding to the next stage of learning but instead explored possible ways to enable the chimpanzees to learn the sequence by checking their individual motivation levels and performance; thus the process varied across individuals.

When “4” is newly introduced after the learning of “1–2–3,” a naïve prediction is that chimpanzees would touch “4” either first or last, preserving the already learned sequence “1–2–3”. However, the introduction of the new number “4” in fact disrupted the chimpanzees’ performance. They tended to touch “1” first, and “3” last, which meant that they tended to touch “4” at some point in between.

When the correct sequence is “1–2–3–4”, touching either “3” or “4” first or second results in failure; therefore, no bias for touching “3” or “4” would be expected to occur if the subjects were learning from scratch. However, when the proportion of the chimpanzees’ erroneous first and second touches was analyzed, touching “4” initially tended to be more frequent than touching “3” (circles in Fig. 1c), which decreased as learning progressed. The proportion of touching “4” was negatively correlated with the correct touch rate (Fig. 1d). These results suggest that the chimpanzees might be associating some kind of property with each numeral (“1,” “2,” “3,” and “4”) shared across the order of touches (first touch, second touch, and so forth). Additionally, successful performance might result from dissolving the error bias to “4” against “3” in the first and second touches.

2.2 TD learning and chimpanzee data

In TD learning, the process of learning is divided into two parts performed by an “actor” and a “critic,” respectively. The actor stochastically makes a choice on the basis of priorities of available actions. The action priorities typically depend on the “state” that indicates a given situation for action choice. The critic estimates how much of the reward would be given after the state, which is called the state value. If the current state changes to a higher-valued state, then the change is expected to be desirable, and vice versa. Thus, the temporal change in the state value leads to self-evaluation for the outcome of the chosen action. By using the evaluation of the “critic,” the “actor” updates the action priorities. The interplay between actor and critic enables learning of appropriate action choices in given situations.

Our consideration of chimpanzees’ initial error patterns, described in Sect. 2.1, suggests that action priorities might be common across the order of touches, i.e., the state setting for the actor might be independent of the touch order. In the standard formulation of TD learning, the actor and critic utilize a common state setting, which might naturally lead to the state setting of the latter also being independent of the touch order. If so, only the action priority of the last touch would be reinforced, because the reward directly follows only the last touch, and the critic provides no order-dependent evaluation. As a result, the actor would have learned to touch the last numeral first. A corollary of our reasoning in the TD learning framework is that the agent can never learn the serial learning task.

3. Proposed approach

3.1 Outline

The standard TD framework is not suitable for modelling a serial learning task, as described above. However, if we remove the constraint of the standard formulation in which actor and critic share a common state setting, it becomes relatively simple to see how the agent could learn the task despite the actor’s independence of the touch order. We assumed the actor’s independence of any state, and

the critic's dependence on it, to reflect the touch order. To test whether the state-independent actor and the touch-order-dependent critic were consistent with chimpanzees' behavior, we used the Actor-Critic algorithm, one of the standard TD learning algorithms used in areas such as robotics (e.g., enabling a robot to learn an optimal movement) and financial markets (e.g., enabling an artificial intelligence to predict stock price fluctuations). In the Actor-Critic algorithm, the actor and critic are explicitly separated. Thus, this algorithm was suitable for our research purpose because we could naturally introduce different state settings for actor and critic.

3.2 Actor-critic algorithm with a state-independent actor in a serial learning task

We assumed a state-independent actor, namely with action priorities (or “policy-parameter” in the terminology of Actor-Critic) that depend only on available actions: touching numerals “1,” “2,” “3,” and “4”: q_1 , q_2 , q_3 , and q_4 . The probability of choosing action a is determined as

$$P(a) = e^{\beta q_a} / \sum_{a'} e^{\beta q_{a'}}$$

where the summation is performed among the available actions with given items. The parameter β controls randomness of choices ($\beta = 2$ in Fig. 4a). The critic estimates the value of each state. The value of current state $s(t)$ and the priority of chosen action $a(t)$ at time step t are updated as

$$\begin{aligned} V_s(t) &\leftarrow V_s(t) + \alpha (r(t+1) + V_s(t+1) - V_s(t)), \\ q_a(t) &\leftarrow q_a(t) + \alpha (r(t+1) + V_s(t+1) - V_s(t)), \end{aligned}$$

where $\leftarrow$ indicates substitution for the updated variable. Variable $r(t+1)$ is the amount of reward as the outcome of action $a(t)$. The discrete time steps are assigned to the first, second, third and fourth touches if they exist, and the trial end. At the end of the trial, no available actions exist, and hence V_s and q_a are not updated. In the case of a successful trial end, the reward $r(t) = 1$, and otherwise $r(t) = 0$. Parameter α is the learning rate ($\alpha = 0.0025$ in Fig. 4).

3.3 State settings for the critic

Several different state settings for the critic are plausible in the serial learning task. The first option is to consider the order of touch itself: $s =$ “1st touch,” “2nd touch,” “3rd touch,” “4th touch,” and “trial end” (Fig. 3a). The second option is to consider the pattern of remaining items which discriminates all combinations of available items, $\{1,2,3,4\}$, $\{2,3,4\}$, $\{3,4\}$, $\{4\}$, $\{1,2,3\}$, $\{2,3\}$, $\{3\}$, $\{4\}$, $\{1,2\}$, $\{2\}$ and trial end $\{\}$ (Fig. 3b). The third option is to consider the number of remaining numerical items (Fig. 3c). In this third option, the initial state depends on task conditions, for example, in “1–2–3–4” task, $s =$ “4 items remaining,” “3 items remaining,” “2 items remaining,” “1 item remaining,” and “no items remaining (trial end)”; in the “3–4” task, “2 items remaining,” “1 item remaining,” and “no items remaining (trial end).” For the notation of state values in the third option, we used the number of “o” symbols in the subscripts, i.e., V_{oooo} , V_{ooo} , V_{oo} , V_o .

4. Results and discussion

Figure 4 shows the result of the aforementioned third option where the number of remaining numeral items serves as the state setting for the critic. The modified Actor-Critic algorithm was able to learn the sequential touch task (“1–2,” “1–2–3,” and “1–2–3–4”; top in Fig. 4). This shows that learning can be achieved even if the state setting of actor and critic are different. Correct rates increased as the learning of state values in the critic proceeded (middle in Fig. 4a).

In this model, the action priorities (q_1 , q_2 , q_3 , and q_4) depended on only touching numerals “1,” “2,” “3,” and “4” (thus did not depend on the touch order), and the order of priorities was appropriately learned (bottom in Fig. 4a) owing to evaluation by the state values. Note that the priority of “3” (q_3) decreased from the initial value at the last phase of the “1–2–3” task. This led to error bias toward “4” against “3” in the initial phase of the “1–2–3–4” task where the decreased priority of “3” was carried over from the previous stage of learning “1–2–3”. We examined fractions of “3” and “4” errors in the first and second touches (i.e., the number of errors caused by touching “4” divided by the numbers of errors caused by touching “3” and “4,” in the first and second touches) in the same way as shown in Fig. 2c (circles of the top right of Fig. 4a under “1–2–3–4”). Similar to chimpanzees’ actual errors, the correct rate increased as the error bias dissolved in our Actor-Critic model.

The actual training sequences of the chimpanzees were not systematic and were intermingled with “3–4” tasks as described earlier. Therefore, we trained the modified Actor-Critic algorithm based upon the actual behavioral sequences of three individuals. As the error bias is determined by the difference of the action priorities, $q_4 - q_3$ in Actor-Critic learning, we calculated correlations between the action priority difference and the correct rate of the actual behavioral data (Fig. 4b–d). We confirmed negative correlations between the action priority difference and the correct response rate, when the Actor-Critic algorithm learned from the actual behavioral sequences.

We also confirmed negative correlations in the other possible state settings for the critic: state as touch order [= aforementioned first option]: $\rho = -0.96$, $n = 30$, $p = 1.7 \times 10^{-16}$ in Mizuki, $\rho = -0.77$, $n = 79$, $p = 1.2 \times 10^{-16}$ in Loi, $\rho = -0.86$, $n = 143$, $p = 9.4 \times 10^{-43}$ in Zamba; state as pattern of remaining items [= aforementioned second option]: $\rho = -0.95$, $n = 30$, $p = 1.7 \times 10^{-16}$ in Mizuki, $\rho = -0.77$, $n = 79$, $p = 1.6 \times 10^{-16}$ in Loi, $\rho = -0.86$, $n = 143$, $p = 1.1 \times 10^{-42}$ in Zamba; Spearman’s correlation. Even so, we consider the third option (i.e., state as the number of remaining items) as more parsimonious than the other two options, for two reasons. First, many different states are produced if we use the pattern of remaining items as the state setting (the second option described in 3–3: Fig. 3b) whereas there are fewer states if we use either touch order (the first option described in 3–3: Fig. 3a) or number of remaining items (the third option described in 3–3: Fig. 3c) as the state setting. Second, if an agent employs touch order as the state setting, then it needs an additional process to store information about the touch order, because the order is invisible when the task is being performed. In contrast, the number of remaining items is always visible on the display during the task. Relying on invisible information compared to visible information would increase cognitive load.

What the state-independent actor in our model learned was not the ordinal relationship of the numerals, but rather the priority value of each (i.e., “1,” “2,” “3,” and “4”). In other words, the actor does not use ordinal information that “3” comes after “2” and before “4”, whereas when humans

learn a numerical sequence, we do use ordinal information. Our result raises the possibility that chimpanzees were not learning the ordinal relationship of numerals, but instead were learning the priority value of each individual numeral one by one and updating the values when a new item was introduced, as in our Actor-Critic model.

This interpretation does not refute the claim that chimpanzees are capable of learning ordinal relationships. Perhaps they initially employ a strategy similar to the state-independent actor, proceeding to an understanding of ordinal relationships after more experience. It is also possible that an agent might combine both strategies to learn such a task. Importantly, the fact that chimpanzees are able to learn to touch numerals in an ascending order does not prove that they have humanlike understanding of the ordinal relationship of numerical characters.

The reason why the state-independent actor could learn correct serial touching is because the critic provides evaluation signals with the actor step by step. State values can be estimated by the critic after sufficient training even if the item sequence is initially long (e.g., “1–2–3–4”). Therefore, the final performance after sufficient training does not depend on the gradual order of tasks (i.e., “1–2”, “1–2–3” and “1–2–3–4”).

The task order, however, affects learning speed after a new item is added. In cases where the state-setting is the number of remaining items (Fig. 3c), the state values are common across different tasks. Therefore, the state value learned in “1–2” task is also effective in “1–2–3” and “1–2–3–4” tasks. This facilitates quick learning after a task change. In contrast, state values in the other state-settings (Fig. 3a,b) are not common across different tasks. Although the present work did not address possible variations in performance among different state-settings of the critic, a future direction is to evaluate the validity of state-setting by exploring chimpanzees’ performances in different task orders and by comparing them with simulation results of our Actor-Critic models.

5. Conclusion

We attempted to identify the information source for decision making and learning, to explain the behavior of chimpanzees in a cognitive task, in the framework of reinforcement learning. We introduced a non-standard formulation of an algorithm, and successfully applied it to the performance of chimpanzees in a serial learning task with the assumption of a “state-independent actor.”

Although some chimpanzees may learn to perform well in serial learning tasks, they may do so without understanding the ordinal relationships between Arabic numerals. It is possible to perform above chance level by learning the action priority value of each stimulus one by one and/or state by state.

Learning to solve a task can be achieved via different cognitive mechanisms. The use of reinforcement learning theory may help to offer an alternative perspective in the field of animal cognition. One of the benefits of reinforcement learning theory is that it enables analysis of trial-by-trial data to infer an underlying algorithm, not merely success and/or error rate [17, 18]. Recent advances in information technology, such as reinforcement learning and machine learning, will lead to new insights into how the brain works and how it evolved.

Acknowledgements

This study was supported financially by the Ministry of Education, Culture, Sports, Science, Japan Society for the Promotion of Science to SH (23H00494). We thank Kristin Havercamp and James Anderson for editing a draft of this manuscript. We also thank Benjamin Knight, MSc., from Edanz (<https://jp.edanz.com/ac>) for English editing.

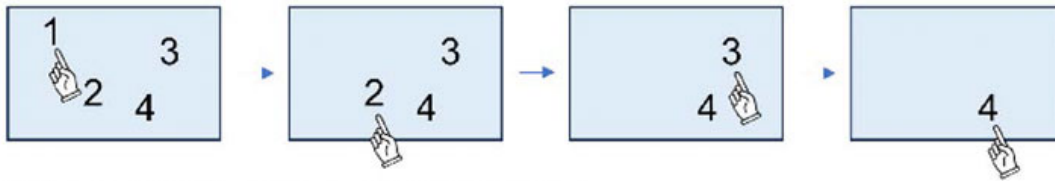
References

- Sutton RS, Barto AG (2018) Reinforcement learning: an introduction, 2nd edn. The MIT Press, Cambridge
- Lee D, McGreevy BP, Barraclough DJ (2005) Learning and decision making in monkeys during a rock-paper-scissors game. *Cogn Brain Res* 25:416–430
- Lee D, Conroy ML, McGreevy BP, Barraclough DJ (2004) Reinforcement learning and decision making in monkeys during a competitive game. *Cogn Brain Res* 22:45–58
- Law CT, Gold JI (2009) Reinforcement learning can account for associative and perceptual learning on a visual-decision task. *Nat Neurosci* 12:655–663
- Bari BA, Moerke MJ, Jedema HP, Effinger DP, Cohen JY, Bradberry CW (2022) Reinforcement learning modeling reveals a reward-history-dependent strategy underlying reversal learning in squirrel monkeys. *Behav Neurosci* 136:46–60
- Brea J, Clayton NS, Gerstner W (2023) Computational models of episodic-like memory in food-caching birds. *Nat Commun* 14
- Sato Y, Sakai Y, Hirata S (2023) State-transition-free reinforcement learning in chimpanzees (Pan troglodytes). *Learn Behav* 51:413–427
- Schultz W (1997) Dopamine neurons and their role in reward mechanisms. *Curr Opin Neurobiol* 7:191–197
- Kravitz AV, Tye LD, Kreitzer AC, Francisco S, Francisco S (2012) Distinct roles for direct and indirect pathway striatal neurons in reinforcement. *Nat Neurosci* 15:816–818
- Yagishita S, Hayashi-Takagi A, Ellis-Davies GCR, Urakubo H, Ishii S, Kasai H (2014) A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science* (1979) 345:1616–1620
- Averbeck BB, Costa VD (2017) Motivational neural circuits underlying reinforcement learning. *Nat Neurosci* 20:505–512
- Nonomura S, Nishizawa K, Sakai Y, Kawaguchi Y, Kato S, Uchigashima M, Watanabe M, Yamanaka K, Enomoto K, Chiken S, Sano H, Soma S, Yoshida J, Samejima K, Ogawa M, Kobayashi K, Nambu A, Isomura Y, Kimura M (2018) Monitoring and updating of action selection for goal-directed behavior through the striatal direct and indirect pathways. *Neuron* 99:1302–1314.e5
- Kawai N, Matsuzawa T (2000) Numerical memory span in a chimpanzee. *Nature* 403:39–40
- Inoue S, Matsuzawa T (2007) Working memory of numerals in chimpanzees. *Curr Biol* 17:1004–1005

- Biro D, Matsuzawa T (2001) Chimpanzee numerical competence: cardinal and ordinal skills. In: Matsuzawa T (ed) Primate origins of human cognition and behavior. Springer, Tokyo, pp 199–225
- Idani G, Hirata S (2007) Studies at the Great Ape Research Institute, Hayashibara. In: Washburn DA (ed) Primate perspectives on behavior and cognition. American Psychological Association, Washington DC, pp 29–36
- Doya K (2007) Reinforcement learning: computational theory and biological mechanisms. HFSP J 1:30
- Wilson RC, Collins AGE (2019) Ten simple rules for the computational modeling of behavioral data. Elife 8:1–33

Sequence of “1-2-3-4”

1) Correct response



2) An example of incorrect response

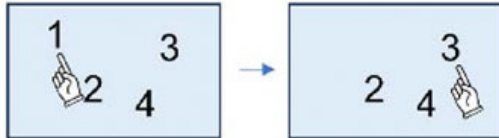


Fig. 1 The flow of the “1-2-3-4” serial learning task

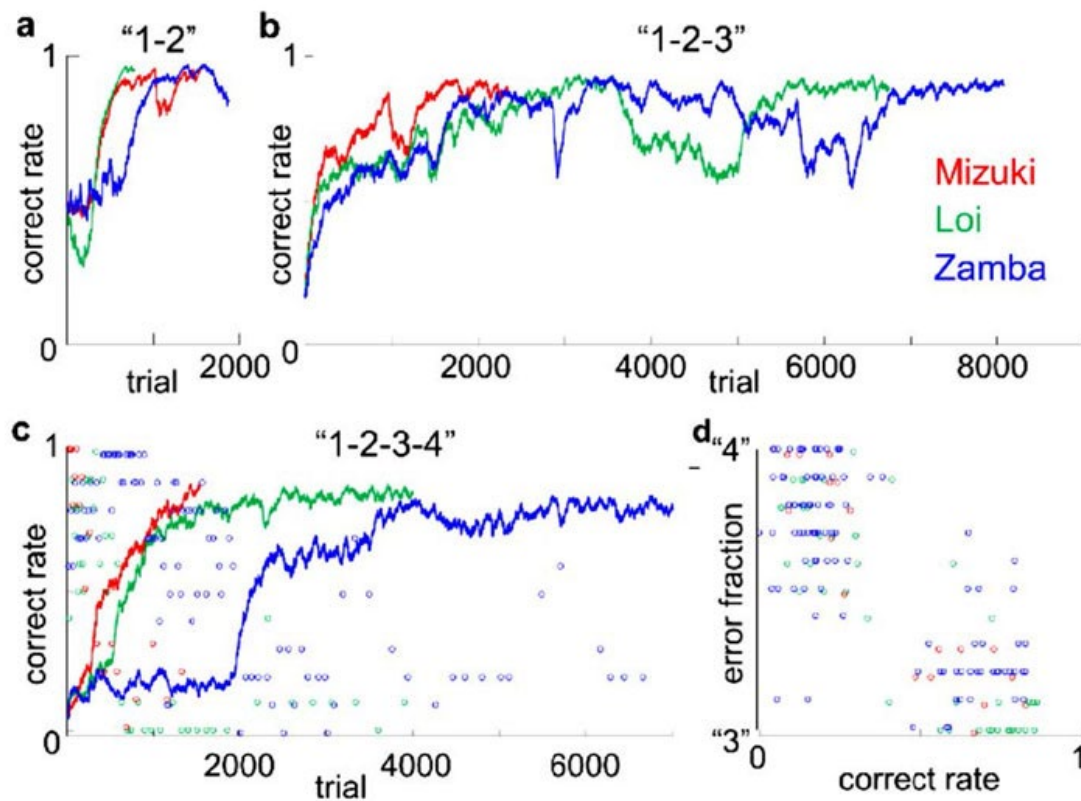
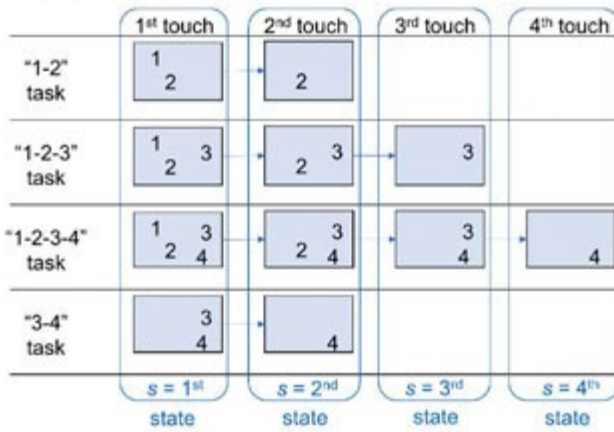


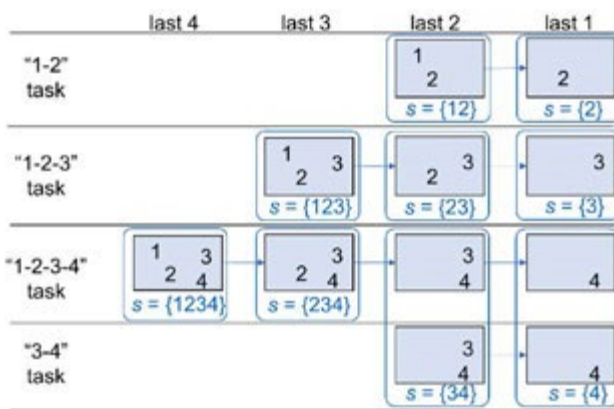
Fig. 2 Learning curves in respective tasks and error fractions in “1-2-3-4” task. a Correct rates in “1-2” task. Each color represents an individual. Correct rates are exponentially weighted averages with time scale of the most recent 100 trials. Initial values were set to chance level (0.5 in the “1-2” task). b The same as a but in the “1-2-3” task. c Solid lines are the same as a but in the “1-2-3-4” task. Circles indicate fractions of “3” and “4” errors at the first and second choices. Each circle was

plotted per 10 trials of “3” or “4” errors at the first and second choices. Hence, the horizontal density of circles indicates occurrence of “3” and “4” errors. d Relationship between “3”–“4” error fractions and correct rates. The vertical axis is common to circles in c. Each correct rate was calculated in the span along 10 trials of “3” and “4” errors for the corresponding error fraction. The error fraction and correct rate were negatively correlated (Spearman’s correlation, $\rho = -0.86$, $n = 18$, $p = 3.8 \times 10^{-6}$ in Mizuki, $\rho = -0.75$, $n = 44$, $p = 4.3 \times 10^{-9}$ in Loi, $\rho = -0.58$, $n = 103$, $p = 1.1 \times 10^{-10}$ in Zamba)

a State setting: touch order



b State setting: pattern of remaining items



c State setting: number of remaining items

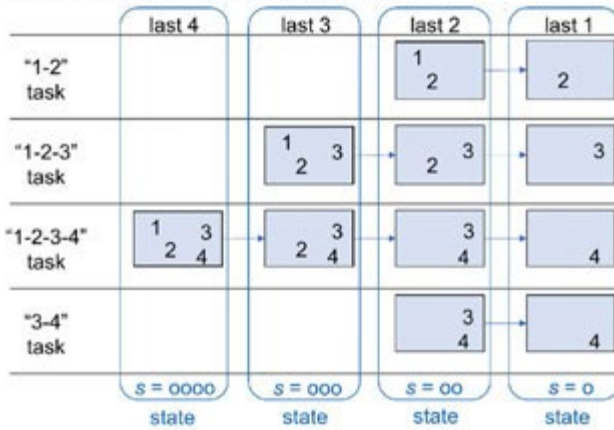


Fig. 3 Plausible state settings

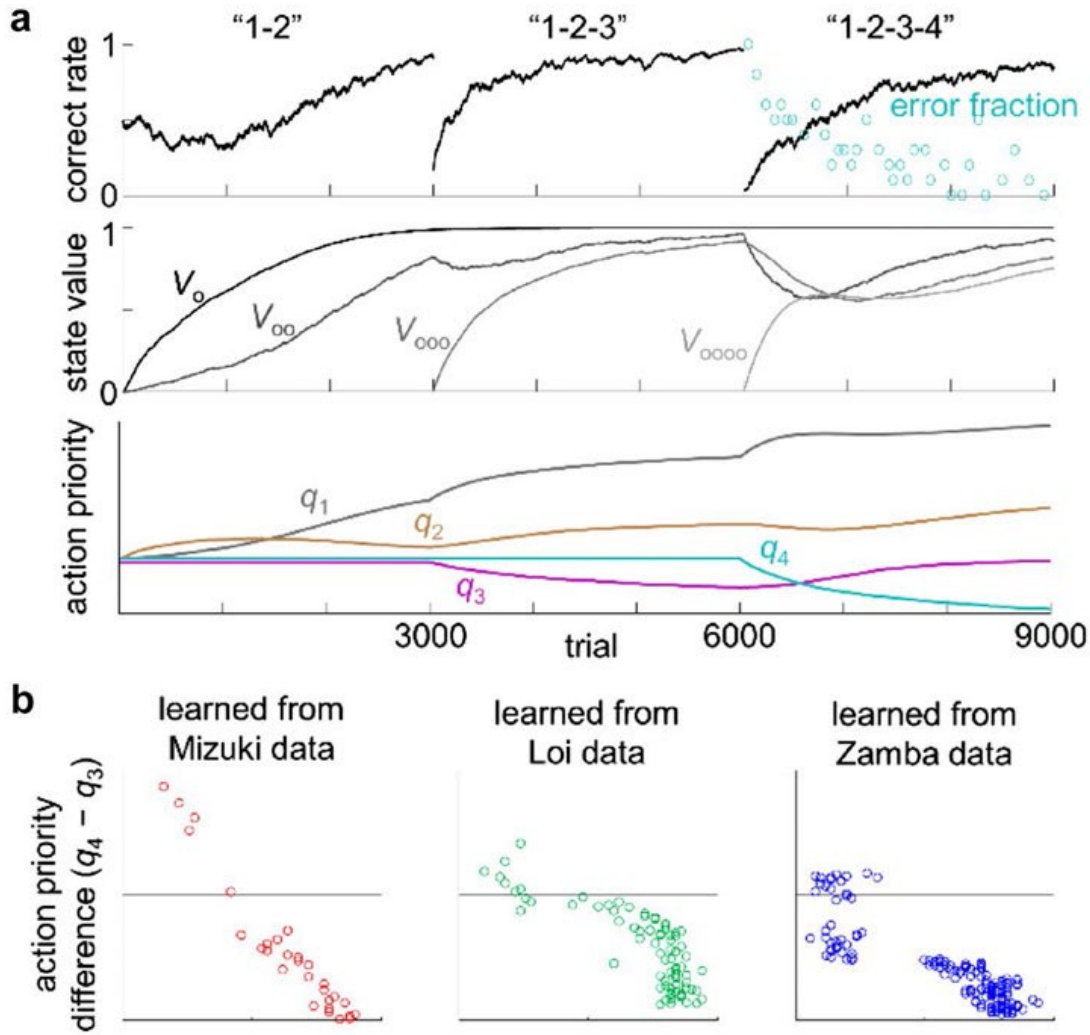


Fig. 4 Results of Actor-Critic algorithm with state-independent actor and state-dependent critic. a Simulation of Actor-Critic learning in "1-2," "1-2-3," and "1-2-3-4" tasks. The top panel shows correct rates calculated as in Fig. 1. The middle panel shows the state values of the critic depending on the number of remaining items (indicated as the number of "o" symbols in the subscripts of V). The bottom panel shows state-independent action priorities of four items, "1," "2," "3," and "4" ($= q_1, q_2, q_3$, and q_4). The axis scale is not shown because the action priorities of the Actor-Critic algorithm are dimensionless values, and only differences among priorities are meaningful. b-d) The results of the Actor-Critic algorithm that learned actual behavioral sequences of chimpanzees Mizuki (b), Loi (c) and Zamba (d). The plots show the correlations of difference between action priorities for "4" and "3" (i.e., $q_4 - q_3$) in the actor with actual correct rates of behavioral data in the "1-2-3-4" task. Each circle was plotted per 50 trials of the "1-2-3-4" task. The action priority difference and actual correct rate were negatively correlated (Spearman's correlation, $\rho = -0.06$, $n = 30$, $p = 1.5 \times 10^{-16}$ in Mizuki, $\rho = -0.77$, $n = 79$, $p = 7.8 \times 10^{-17}$ in Loi, $\rho = -0.86$, $n = 143$, $p = 1.2 \times 10^{-42}$ in Zamba)