Habit formation viewed as structural change in the behavioral network.

AUTHORS:

Kota Yamada[1,2], Koji Toda[1]

AFFILIATIONS:

1 Department of Psychology, Keio University, Tokyo, JAPAN

2 Japan Society for Promotion of Science, Tokyo, JAPAN

CORRESPONDENCE:

Kota Yamada

Department of Psychology, Keio University

Mita 2-15-45, Minato-ku, Tokyo, JAPAN

Email: haroldthebarrel.yk@gmail.com

Abstract

Habit formation is a process in which an action becomes involuntary. While goal-directed behavior is driven by its consequences, habits are elicited by a situation rather than its consequences. Existing theories have proposed that actions are controlled by these two distinct systems. Although canonical theories based on such distinctions are starting to be challenged, there is no theoretical framework that implements goal-directed behavior and habits within a single system. Here, we propose a novel theoretical framework by hypothesizing that behavior is a network composed of several responses. With this framework, we have shown that the transition of goal-directed actions to habits is caused by a change in a single network structure. Furthermore, we confirmed that the proposed network model behaves in a manner consistent with the existing experimental results reported in animal behavioral studies. Our results revealed that habit could be formed under the control of a single system rather than two distinct systems. By capturing habit formation as a single network change, this framework can help study habit formation for experimental and theoretical research.

Author summary

To obtain the desired consequences, organisms need to respond based on the knowledge of the consequences obtained by the response and the change in the environment caused by it. Such a process is called goal-directed behavior, which is flexible, but requires high computational cost. Once the same response is repeatedly performed under the same environment, the response becomes automatic, and transforms into a habit. In the canonical views, such a change from goal-directed response to habit was explained by the associative structures between the corresponding systems, goal-directed, and habit systems. However, the dichotomy between goal-directed responses and habits has recently been challenged. Here, we show that, instead of assuming two explicitly distinguished systems as in the canonical views, behavior is regarded as a network consisting of multiple responses, and that changes in the structure of the network cause two behavioral features, goal-directed behavior and habit. The transition from goal-directed behavior to habit has been operationally defined by sensitivity to the reward obtained by the response. We replicate such an experimental paradigm in the simulation and show that the behavioral network model can reproduce the empirical results on habit formation obtained from animal experiments. Our results demonstrate that habit formation can be explained in terms of changes in the network structure of behavior without assuming explicitly distinct systems and thus, provide a new theoretical framework for habit formation.

Introduction

To behave flexibly in a given environment, organisms need to choose their actions based on the consequences of the actions. This type of behavior is called goal-directed behavior. As we keep repeating the same action under a certain situation, the action is elicited by the situation rather than its consequences. This type of behavior is called a habit. Goal-directed behavior requires high computational resources because organisms must process the information about their external environment and how their actions affect it. In contrast, habit shows a more stereotyped and less flexible behavior, requiring less computation. In this sense, habit formation can be viewed as the optimization process of energy consumption by the organism.

Many existing theories about habit formation are based on evidence from experimental or theoretical research in psychology and neuroscience. Although almost all of them have different views, they have a common basic assumption that behavior is controlled by two distinctive systems: goal-directed and habit systems (Perez and Dickinson, 2020; Daw et al., 2011; Bogacz, 2020; Dezfouli and Balleine, 2012). They are distinct by two axes, based on their assumptions. One axis is whether rewards are necessary for the acquisition of a stimulus-response association. Researchers who assume that rewards are necessary for acquisition of stimulus-response (S-R) association proposed that rewards strengthen or stamp-in the S-R association, suggesting that such a process is similar to Thordikes's "law of effect" (Perez and Dickinson, 2020; Daw, et al., 2011). In contrast, other researchers proposed that rewards are not necessary for acquisition of S-R association, but only a response is required, likely similar to Guthrie’s view of learning (Guthrie, 1930). They proposed that S-R association is acquired only through history of responses, such as how many times the subject emitted a certain response under a given situation (Miller, Shenhav, and Ludvig, 2019; Bogacz, 2020). Another axis is whether the goal-directed and habit systems exert flat or hierarchical control over behavior. The former viewpoint, flat control, proposed that goal-directed and habit systems control behavior with different weights, and the difference in the weights determines whether the behavior is goal-directed or habit (Perez and Dickinson, 2020; Daw, et al., 2011; Bogacz, 2020). The latter viewpoint, hierarchical control, proposes that a response is selected in a goal-directed way and the responses up to the goal are generated as a sequence based on the response history (Dezfouli and Balleine, 2012). However, such views on habit formation have started to change. Recently, some researchers reviewed existing studies on habit formation and cast doubt on the canonical framework of habit formation by showing the possibilities that habits are also controlled by their consequences (De Houwer, 2019; Kruglanski and Szumowska, 2020).

Although there are many empirical and theoretical studies on habit formation, no one

asked the question "where should we attribute the cause of habit formation?" ever. All models introduced above attribute the cause of habit formation to the two systems. Instead of assuming the two systems, we decided to attribute the cause to the behavior structure. We raise a novel question, not to ask how the two systems control the responses, but how the structure of the behavior could be organized by the responses, and what characteristics in individual responses can arise from the structure.

Behavioral network

There are two methodological approaches for studying animal behavior. One stream is an in-laboratory psychological approach that studies the behavior of animals, including humans, under experimentally controlled environments. Here, investigators measure only experimentally defined responses of subjects (lever presses, key pecks, nose pokes, etc.) or put them into rigidly controlled situations where they can only engage in the responses. Another stream is an ethological approach that studies animal behavior under more natural and ecologically valid environments (Timbergen, 1951). Although these two approaches seem to conflict with each other, both are complementary for understanding behavior and its biological substrates. Recent advances in machine learning have allowed us to objectively measure the detailed structure of behaviors (Wiltschko, et al. 2015; Mathis, et al., 2018, Markowitz, et al. 2018). They are expected to reveal the relationship between behavior and its underlying mechanism in a way that integrates the different disciplines of psychology, neuroscience, and ethology (Datta, et al., 2019). However, conventional views on behavior in psychology and neuroscience are based on empirical results obtained from the approaches before the appearance of such a new framework. Here, we present a new theoretical novel framework that focuses on how behavior is organized and how its structure brings specific characteristics to behavior.

Existing studies measured only specific pre-defined responses of animals, including humans, and ignored various responses that the animals actually engaged in. However, there is considerable evidence that animals engage in various responses which affect the learned responses. For example, animals engage in a specific response immediately after the reward presentation (Falk, 1966; Hymowitz, 1971; Gentry, 1968; Levitsky, 1968), engage in responses irrelevant to an experiment (Skinner, 1948), show a specific response sequence between reward presentations (Staddon and Simmelhag, 1971), or show a specific response that counteracts learned responses (Breland and Breland, 1961). Theoretically, some characteristics of operant responses are explained by assuming the existence of other responses (Guthrie, 1930; Herrstein, 1970; Killeen, 1988; Yamada and Kanemura, 2020). These experimental facts and theoretical assumptions indicate that animal responses do not

exist in isolation but are associated with other responses. We assume such relationships between responses as a network in which responses and transitions between them are considered nodes and edges, respectively.

Here, we provide a computational formulation of the behavioral network and explain habit formation from the viewpoint of changes in the network structure. In simulation 1, we generated an arbitrary network and examined what kind of structure forms habit, and showed that habit formation occurs when edges are concentrated on a specific response. In Simulation 2, we examined whether the factors reported to promote or inhibit habit formation from existing behavioral studies have similar effects on the proposed model. There are three important factors on habit formation: 1) the amount of training (Adams, 1982; Dickinson, et al., 1995), 2) the schedule of rewards (Dickinson, Nicholas, and Adams, 1983), and 3) the presence or absence of choice (Colwill and Rescorla, 1985; Kosaki and Dickinson, 2010). The effects of these factors on the proposed model were consistent with the existing experimental results. These results imply that habit formation can be explained not by the control of the two systems, but by a single system constituting the change in the structure of the behavioral network. Furthermore, the results demonstrate that all responses are goal-directed, rather than the conventional dichotomy of goal-directed and habitual behaviors.

Results

We consider an agent as a network consisting of different categories of responses (e.g., lever pressing, grooming, stretching, etc.) and the transition between those responses. Fig. 1 illustrates the model scheme. Each response is assumed as a node, and the transition between responses (lever pressing after grooming) is assumed to be an edge. The probability of an edge between nodes is calculated according to the Q values of the transitions between the responses. As the agent obtains a reward from the transitions, the Q values increase.

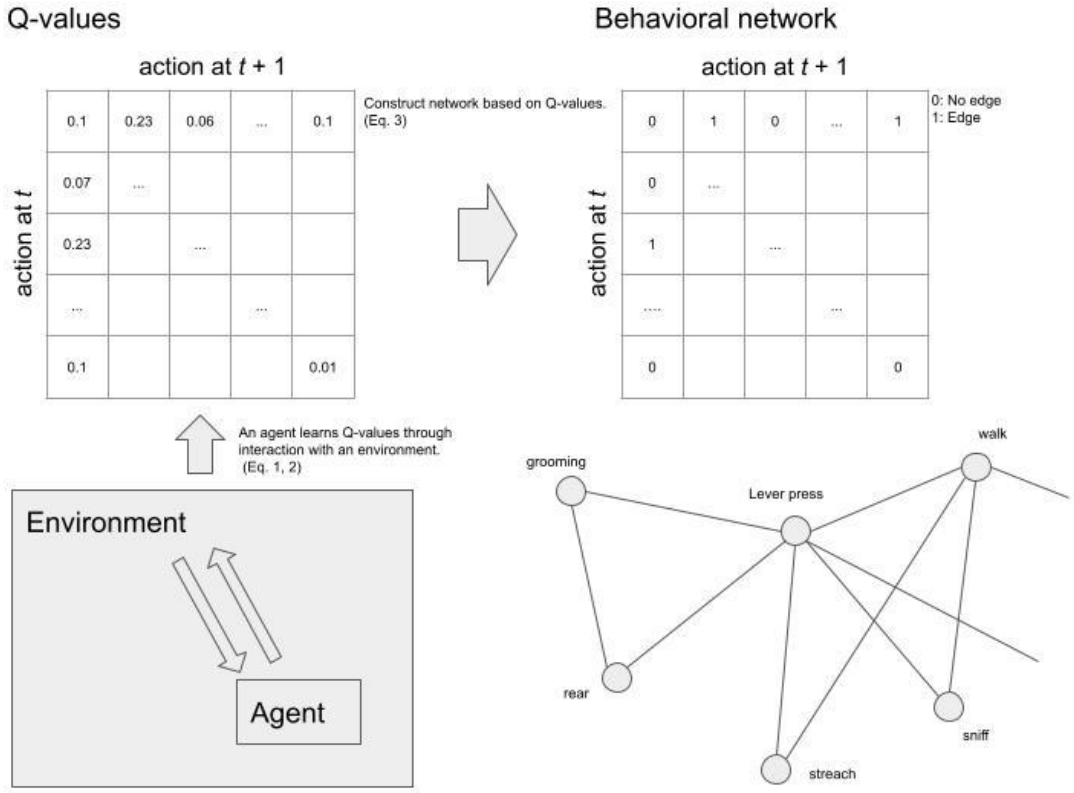


Figure 1. The scheme of the behavioral network.

Agents learn the Q-values from the interaction between the environment and generate a behavioral network based on these values. The behavioral network consists of multiple responses.

We employed Q-learning to learn the Q-values (Watkins and Dayan, 1992). The Q values are updated according to the following equation:

$$Q(a_{t-1}, a_t) \leftarrow Q(a_{t-1}, a_t) + \alpha \cdot \delta$$

In this equation, α denotes the learning rate, and δ is the reward prediction error (or temporal difference error). The reward prediction error was calculated as follows:

$$\delta = R(a_t) - Q(a_{t-1}, a_t)$$

In this equation, R_t denotes the reward obtained by a transition, and the reward functions are different between simulations, which have been explained in detail in the Materials and Methods section.

The probability that an edge is attached between any two nodes depends on the Q-value and is expressed in the form of proportional allocation. The probability was calculated using the following equation:

$$p_{i,j} = \frac{Q(i,j)}{\sum_{j \in N} Q(i,j)}$$

In this equation, N denotes the number of nodes in the network and all the responses that the agent can engage in. We also sampled two edges according to Equation 3, such that every

node had at least two edges. We used NetworkX (Hagberg, Swart, and Chult; 2008), a Python library for network analysis, to generate the network.

The algorithm for the agent to choose a response contains two steps: 1) choice of the response based on the value of the reward, and 2) searching the shortest path from the current engaging response to the chosen node. In the choice of the response based on the value of the reward, the probability of choosing a response is calculated by proportional allocation of the reward value. The shortest path search includes selecting the shortest path between the current response to the chosen response and the agent engaging in the responses containing the path in sequence.

The probability of response i was calculated according to the following equation:

$$p_i = \frac{r_i}{\sum_{iinN} r_i}$$

In this equation, r_i denotes the value of the reward obtained from response i. In our simulation, the value of the reward obtained from the operant response was 1.0, and the other response was 0.0028.

The shortest path search is used to find the shortest path between any two nodes in the network. We employed the Dijkstra method in all our simulations. If there were multiple shortest paths between any two nodes, we randomly choose one of them. We implemented the path search by using NetworkX (Hagberg, Swart, and Chult; 2008).

Simulation 1: Network structure and habit formation

In simulation 1, we generated arbitrary networks and examined whether or not an agent showed habits. In particular, we systematically operated the degree of the edge concentration on the operant response. To examine the degree of habit, we used the reward devaluation procedure used in free-operant experimental situations. The earliest demonstrations of habit formation by Dickinson et al. used the reward devaluation procedure. In this procedure, the investigators train the animals to press the lever with a reward. After the animal learned lever pressings to obtain the reward, the value of reward was reduced by poisoning it with lithium chloride. Then, the investigators examined whether the animal pressed a lever to obtain the reward. If the animal presses the lever, although the reward is poisoned, the responses are considered to be habit. In this study, to reproduce the procedure, we set up the baseline and test phases where the values of reward are 1 and 0, respectively. We calculated the proportion of the operant response to the total number of responses to assess whether the operant response is habit or not.

Fig 2 (a) shows the devaluation ratio, which indicates a decrease in the operant response in the test phase. $Q_{operant}$ denotes a parameter that determines the cohesion of

edges to the operant response, and as the parameter increases, the number of edges connected to the operant response increases. As $Q_{operant}$ increased, the devaluation ratio increased. Fig 2 (b) shows the changes with respect to $Q_{operant}$ in the proportion of the number of edges the operant response acquired to the total number of edges in the network. As $Q_{operant}$ increased, the operant response acquired more edges. Fig 2 (c) shows the relationship between the devaluation ratios and the cohesion of the edges in the operant response. As the cohesion of the edges increased, the devaluation ratio increased. Fig 2 (d) shows the networks generated under the given $Q_{operant}$ and as $Q_{operant}$ increased, edges from other responses (black nodes) were concentrated to the operant response (the red node). These results show that when the operant response acquired most of the edges from other responses, habit formation occurred.

Simulation result

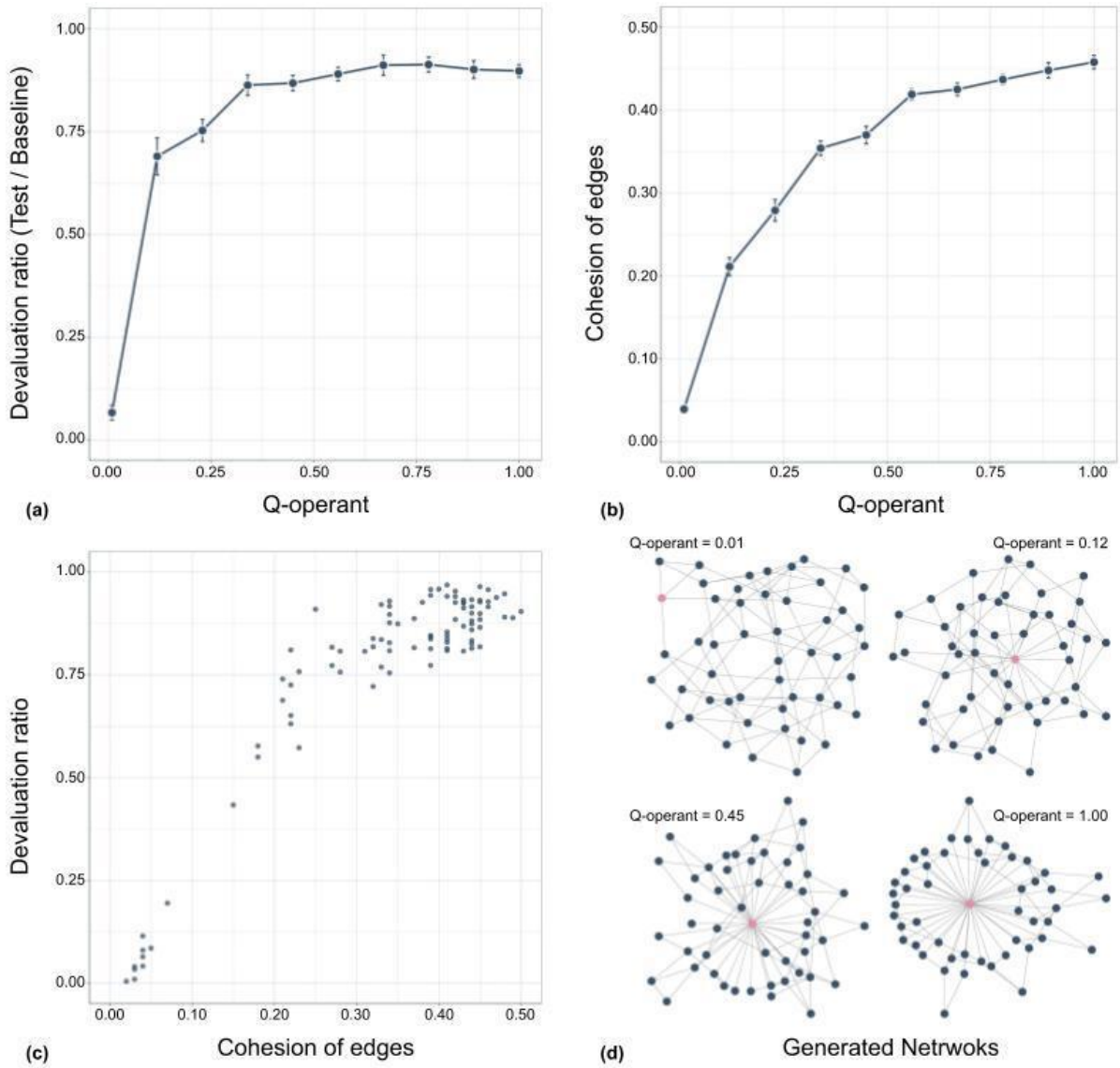


Figure 2. Results of Simulation 1

(a) The line plot shows a decrease in operant response due to reward devaluation per $Q_{operant}$. Devaluation ratio is the ratio of the proportion of operant response to the total number of responses in the test phase to the baseline phase. (b) It shows the change in the cohesion of edges in the operant response to changes in $Q_{operant}$. The cohesion of edges is the proportion of the number of edges connected to the operant response to the total number of edges in the network. (c) Plot of the devaluation ratio to the cohesion of the edges in the operant response. (d) Example of a network generated under an arbitrary $Q_{operant}$, where a red node and black nodes denote the operant and other responses respectively.

Interim Discussion

In simulation 1, we examined the structure of the network and habit formation under arbitrary Q-values and showed that habit formation occurs when edges from other responses are concentrated in the operant response. Fig 1 (a) and (b) show that as $Q_{operant}$ increased, the differences between the baseline and test phases decreased, which implies that habit formation occurs under high $Q_{operant}$. Fig. 1 (c) and (d) show that as $Q_{operant}$ increased, the number of edges acquired by the operant response increased. These results support the idea that habit formation can be explained by the concentration of the edges in the operant response.

As the number of edges acquired by the operant response increases, the operant response becomes a habit because the operant response is likely to be contained in the shortest path between any two responses. In the baseline phase, the value of the reward obtained from the operant response was high and the operant response was chosen frequently, so the proportion of the operant response to the total number of responses was high. After the reward devaluation, the operant response was not chosen because its reward value was 0.0. However, when the operant response acquired many edges from other responses, although it was not chosen, it was likely to be contained in the shortest path between any two nodes. Thus, the effect of reward devaluation varied depending on the number of edges acquired by the operant response.

Simulation 2: Devaluation and its effect on behavior under free-operant situation

In simulation 2, we let an agent learn Q-values under arbitrary experimental environments and examine whether habit formation occurs. Under free-operant situations, there are three factors that lead to an operant response to habit. The first is the amount of training, where one response is rewarded repeatedly under one situation, and the response becomes habit. The second schedule of reinforcement is the rule of reading responses. The variable interval (VI) schedule determines whether the reward is presented depending on the elapsed time, and the variable ratio (VR) schedule depends on the number of responses. Under the former schedule, the operant response tends to be a habit. The third factor was the presence of alternatives. If two alternatives are available under a situation and different rewards are obtained from them (e.g., left lever → food, right lever → water), the operant response does not become a habit. Here, we reproduce the above experimental settings and examine whether our model becomes a habit under these environments.

The only difference between simulations 1 and 2 is whether the agent learns the Q-values. Here, the agent experienced the training phase preceding the baseline phase, where the agent learned Q-values through interaction with a given environment and constructed a

network based on them. After the training phase, the agent experienced the baseline and test phases in the same way as in Simulation 1.

Simulation result

Fig 3 (a) shows the devaluation ratios under the VI and VR schedules as a function of the amount of training. As the amount of training increased, the devaluation ratios increased under both schedules. The devaluation ratios under the VI schedule were larger than those under the VR schedule. Fig 3 (b) shows the cohesion of edges in the operant response, and it increases as the amount of training increases under both schedules. The operant response acquired more edges under the VI schedule than under the VR schedule. Fig 3(c) shows the relationship between the devaluation ratios and cohesion of the edges in the operant response. As the cohesion of the edges increased, the devaluation ratio increased.

Fig 3 (d) shows the devaluation ratios under environments where the explicit alternative is (Choice) and is not (No choice). Under the choice environment, the devaluation ratios are larger than those of the no-choice environment. Fig 3 (e) shows the networks constructed after the training phase under the choice and no-choice environments. The operant response and the alternative response acquired most of the edges on the network in the choice environment. In contrast, in the no-choice environment, the operant response acquired more edges than the other responses, but others acquired a few edges.

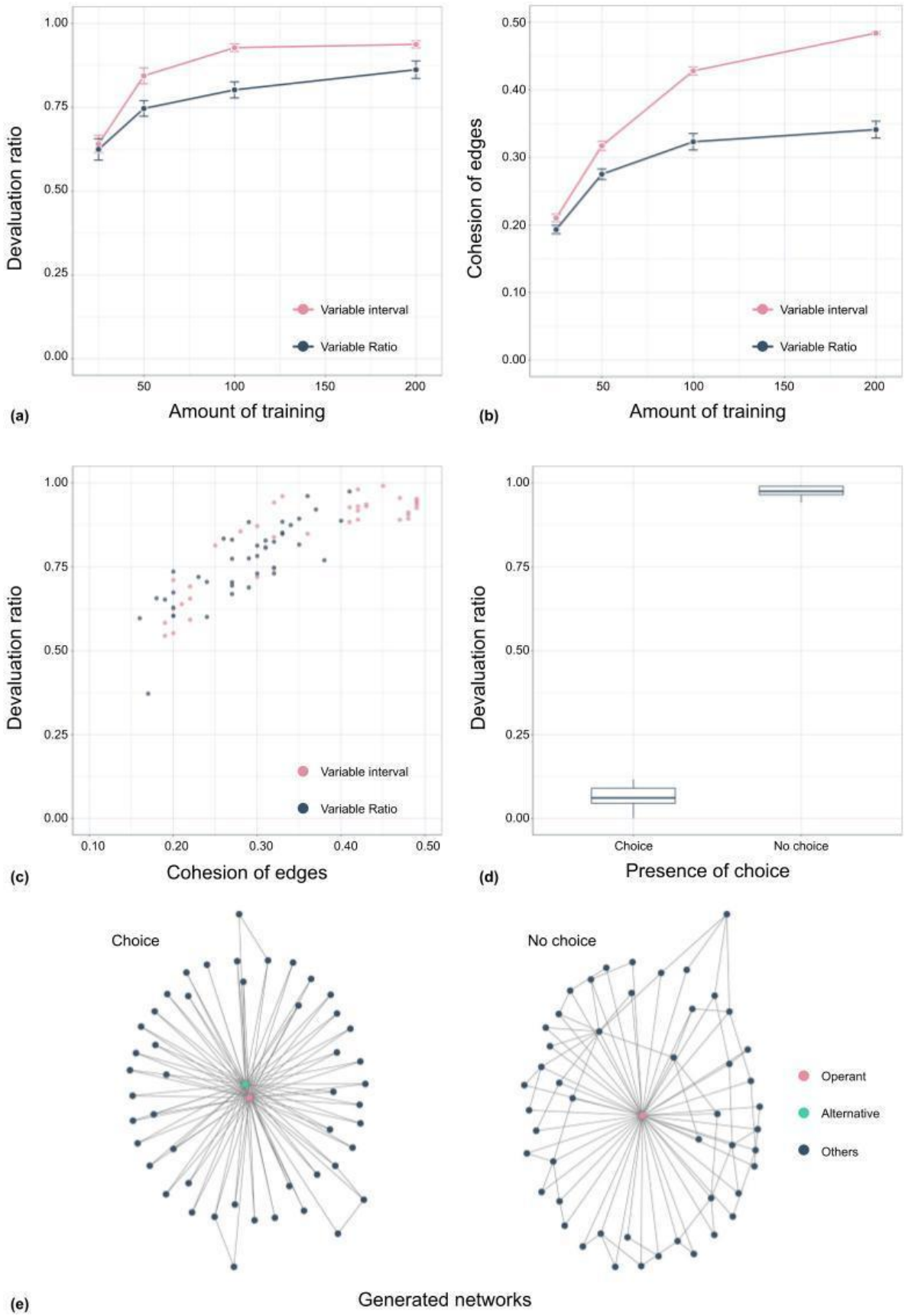


Figure. 3 The results of simulations under VI, VR and presence of choice or not.

(a) shows the changes in devaluation ratio with increasing amount of training.

The red and blue lines denote the data obtained from the simulation under VI and VR schedules respectively. (b) shows the changes in the cohesion of edges in the operant response with increasing amount of training. (c) Scatter plot of the devaluation ratio to the cohesion of the edges in the operant response. (d) Boxplot of devaluation ratio under the choice (left) and no choice (right) situations. (e) Generated networks under the choice (left) and no choice (right) situations.

Interim Discussion

In simulation 2, we examined whether our model shows similar behavior to real animals in environments that affect habit formation, and our model reproduced the similar results reported from the empirical studies. Fig 3 (a) shows the devaluation ratios under the VI and VR schedules with different amounts of training, and the devaluation ratios increased as the amount of training increased and were larger under the VI schedule than the VR schedule. These results imply that the VI schedule and a large amount of training promote habit formation in the operant response. Fig 3 (d) shows the devaluation ratios under the environments where the explicit alternative is or is not. The devaluation ratio in the choice environment is smaller than that in the no-choice environment, and the result implies the presence of alternative disturbed habit formation.

The amount of training affects the structure of the network (Fig 3b), and as the amount of training increases, the cohesion of edges in the operant response increases. The smaller the amount of training, the smaller the Q-values of the transition from other responses to the operant response. Consequently, the probability that an edge is attached to the operant response is smaller. As shown in simulation 1, habit formation occurs when the operant response acquires most of the edges in the network. Thus, the amount of training affects habit formation.

The devaluation ratio is larger under the VI schedule than under the VR schedule because the reward probability of the transition from other responses to the operant response is larger under VI schedule. Under the VI schedule, the reward probability increases as the elapsed time from the last response increases. In contrast, under the VR schedule, the response probability remained constant for all the operant responses. Thus, under the VI schedule, the reward probability of the operant response increases as the agent spends more time on the other responses. This characteristic of the VI schedules promotes habit formation.

Although the operant response acquired most of the edges on the network under the choice environment, the operant response did not become a habit. There are two reasons for this. First, the agent chooses its response based on the value of the reward obtained from the

response. In the test phase, the value of the reward obtained from the operant response was reduced, and that of the alternative response remained the same value as the baseline. Thus, the agent chose the alternative response more in the test phase than in the baseline phase. Second, if only the operant response acquired most edges, any shortest path may contain the operant response. However, the alternative response acquired most of the edges, so that any shortest path contained the alternative response. Thus, the operant response no longer has a greater chance of being engaged, and habit formation does not occur.

Discussion

In this paper, we explain habit formation as changes in network structure by assuming the behavior of organisms viewed as a network of responses. In simulation 1, we generated arbitrary networks and examined the underlying structure of habit formation. We revealed that habit formation occurs when a particular response acquires most of the edges from other responses. In Simulation 2, we simulated the environments that were reported to promote or inhibit habit formation from existing studies and examined whether the proposed model showed habit formation. The results were consistent with experimental results reported by many laboratories, suggesting that our results demonstrate habit formation as a structural change in the behavioral network.

Relationship to other theoretical models of habit formation

Although there are many models of habit formation, all of them are viewed as goal-directed behavior and habits as interactions between two distinctive associative structures. In this paper, we succeeded in providing a novel explanation by taking a more macro view of behavior. Specifically, the proposed model differs from existing models in two ways. First, the idea is not to regard behavior as a single element, but as a network of interconnected responses. Conventional views focus only on responses under highly constrained experimental situations, such as lever pressings or button pushings, and ignore the macro structure of behavior that the real organisms may have. Responses of organisms, including humans, are not independent of each other, but they are probabilistically conditioned by the preceding and succeeding responses. In the proposed model, the structures of such responses are represented as a network, and habit formation is explained as a change in the structure. Second, the proposed model does not explicitly assume two systems corresponding to goal-directed behavior and habits. We assumed that all behaviors are choices based on the value of rewards and searches for the shortest path in the network. In other words, all behaviors are goal-directed. However, the structure of the network constrains the available shortest path

between arbitrary responses, resulting in habit formation.

Conventional models can be classified by the manner in which the two systems control behavior: flat or hierarchical. Our model significantly differs from those that control behavior in a flat way. In these models, the two systems control behavior with different weights, and the relationship between the weights determines whether the behavior is goal-directed or habits. On the other hand, our model shows that the structure of the network causes habit formation. The hierarchical model assumes that habits are the formation of response chains. This view of habit formation is similar to that of our model as a change in the structure of a network is a change in the connection of individual responses, which can be interpreted as a response chain in a short time frame. However, in our model, response chains are not formed, but only specific chains are available in a given network structure. Therefore, the treatment of habit formation in our model is similar to that of Dezfouli and Balleine (2012), but with a more macro view. In our model, habits are considered as constraints on the goal-directed choice process of the shortest path search imposed by the macro structure of the behavioral network.

Our model provides a different explanation for habit formation in free-operant situations from the conventional theory (Dickinson, 1985; Perez and Dickinson, 2020), which explains habit formation by response-outcome correlation. In the VR schedule, the number of rewards increased as the number of responses increased. Thus, a positive correlation between responses and rewards was obtained. On the other hand, in the VI schedule, the correlation between responses and rewards is smaller because the reward presentation follows time. The higher the response-outcome correlation, the harder the habit formation. Our model explained habit formation by the transitions between responses that were rewarded rather than response-outcome correlation. Operant responses are likely to be rewarded from other responses owing to the time-dependent property of the VI schedule. On the other hand, the VR schedule does not have such property, and all transitions, including self-transition, are rewarded with the same probability. Thus, in the VI schedule, the operant response acquires many edges from other responses, and habit formation occurs. In other words, our model provides an explanation in terms of changes in the microscopic reward probability of the schedule, rather than in terms of the macroscopic time scale of the response-reward correlation.

Neural substrates of behavioral network

The corticostriatal network is involved in habit formation, and generates response patterns (Graybiel, 1998; 2008). Especially, dorsolateral striatum (DLS) is known to be important in transition from goal-directed behavior to habits (Yin, Knowlton, and Balleine, 2004). For example, DLS activity changes as proceedings of training and responses become habits

(O'Hare, et al., 2016; Tang, et al. 2007), and lesion of DLS turns habits into goal-directed behavior after extended training (Yin, Knowlton, and Balleine, 2006). DLS also carries forming response sequences (Yin, 2010) and motor routines (Jurado-Parras, et al., 2020). In addition to learned behavior, DLS also encodes innate response sequences (Aldridge and Berridge, 1998). These facts imply that habit formation and learning or the formation of response sequences have common neural substrates.

A recent study reported that DLS encodes not only information about response sequences but also more divergent information about behavior, which are topographically categorized responses and transitions between them (Markowitz, et al., 2018). They recorded the DLS activities of mice with fiber-photometry under an open-field situation, and reported neural activities that correlated with the behavior. The activities differed depending on the preceding and succeeding responses, and DLS encoded a transition between the responses. Moreover, the behavior of the mice with DLS lesions showed random transitions of the responses compared to the sham-lesion group. These results imply that the information encoded in DLS is the transition of the structure of behavior. Thus, the function of the DLS might be well understood by considering the habit and goal-directed behavior from the viewpoint of the behavioral network.

Corticostriatal circuits, the associative network, which consists of the prefrontal cortex, dorsomedial, or ventral striatum, plays a role in goal-directed behavior (Balleine and O'Doherty, 2010). The dorsomedial striatum (DMS) is known to be involved in the acquisition of goal-directed behavior, maintaining sensitivity to outcomes, and expressing goal-directed behavior (Ostlund and Balleine 2005, Yin et al., 2005). The DMS receives excitatory inputs from the prefrontal cortex, whereas the DLS receives inputs from the sensorimotor and premotor cortices (Yin et al., 2005). In the canonical dichotomous view of habit formation, goal-directed behavior is replaced by habit after extensive training. After habit formation, the contribution of DLS becomes more important than that of DMS (Yin et al., 2004; Yin et al., 2005). However, even after extensive training, many brain areas such as the prefrontal cortex, anterior cingulate cortex, ventral striatum, and DMS are modulated by anticipated rewards (Niki and Watanabe, 1979; Schultz et al., 1992; Shidara and Richmond, 2002; Watanabe, 1996; Yin et al, 2004; Yin et al., 2005). In our model, any response emitted by an agent is considered goal-directed. Regardless of the training stage, our agents choose their responses based on the value of the rewards. Therefore, the fact that regions involving goal-directed behavior are modulated by anticipated rewards even after extensive training, our assumptions do not contradict each other. Combined with the fact that DLS is more responsible for sequential responding than DMS (Turner et al., 2021), the transition from DMS to DLS during habit formation might reflect the corresponding behavioral sequence induced by changes in the behavioral network.

Relationship to other behavioral phenomena

Animals engage in specific responses, such as orienting, approaching, and consummatory behavior, just after the presentation of the reward. Specific action sequences are observed during experiments, and learning is sometimes disrupted by innate responses. These experimental and observational facts lead us to assume that behavior is a network constructed from responses.

In our model, the structure of a network depends only on past experiences under a given situation. In other words, our model does not consider the connections between specific responses that real organisms may have. Thus, we could not reproduce this phenomenon. However, our model can be further extended and modified to include this phenomenon.

Schedule-induced behavior, observed under intermittent schedules of reinforcement, is a behavioral phenomenon in which animals show aggression or water intake just after reward presentation. This phenomenon can be attributed to the innate connections between reward consumption and schedule-induced behavior. Because of these connections, animals tend to engage in aggression or water intake immediately after reward presentation. Similarly, terminal behavior, which occurs as approaches reward presentations, can be explained by assuming an innate connection, which may explain the fact that animals show a specific sequence of responses during the experiment.

To deal with such phenomena, we believe that it is possible to express the innate susceptibility of edges as a prior distribution and impose constraints on the probabilities of edges attached by learning. Furthermore, we can systematically treat phenomena such as misbehavior and biological constraints on learning by examining differences in prior distributions among species and environments. Thus, we can extend our model to a comprehensive framework of behavior that incorporates the innate behavior of organisms under natural settings.

Limitations and future directions

Our model has three major limitations. First, as we discussed in the previous section, our model does not consider innate constraints that real organisms have, and we believe that we can solve the problem by expressing the innate constraints as a prior distribution. Second, it could not treat the self-transition of each response. Third, it can only deal with experiments on habit formation under free-operant situations.

Our model cannot treat the self-transition of responses because we employed the shortest path search algorithm to generate response sequences. Any self-transition makes paths between any two responses longer, and paths containing self-transitions must not be the shortest paths. However, animals show a particular response pattern, which is called bout-

and-pause and characterized by phasic bursts of one response and pauses following them. Such response patterns imply that the responses have self-transitions. To solve this problem, it is necessary to employ a different algorithm to generate response sequences that allow self-transition.

All of our simulations deal with experiments in free-operant situations, and not with recent experiments with the two-stage Markov decision task. This is not a specific problem for our model; other existing models treat either of them. Although many experiments have been conducted in both experimental tasks, the differences and identities of the procedures and results among them have not been systematically examined. To obtain a more unified understanding of habit formation, we should conduct a systematic analysis of the procedures and results employed and obtained from existing studies. Therefore, the validation of our model is limited to habit formation in free-operant situations.

Recent advances in machine learning allow us to measure animal behavior more objectively and precisely than ever before. By utilizing these techniques, it is possible to understand behavior on a macroscale rather than capture the behavior in highly constrained experimental settings. Our model provides a novel perspective on how behavior can be viewed on macroscale behavioral phenomena, which would further help us understand the functioning of the brain in behavioral changes.

Conclusion

In this paper, we provide a novel perspective on habit formation by assuming behavior as a network. In existing models, goal-directed behavior and habits are controlled by two distinct systems. On the other hand, our model shows that although all responses are goal-directed, both goal-directed and habits result from the structure of the network. It proposes that habit formation is not caused by a change in the control of the two systems, but rather by a continuous change in a single system. Furthermore, the most important feature of our model, which differentiates it from other models, is that behavior is a network constructed from responses. With this view, we have succeeded in providing a novel explanation for habit formation. This implies that the possible algorithms can be changed depending on how one views the behavior of organisms. Our study also suggests that changing the method of capturing behavior could be a fundamental step in understanding the biological structure of the behavior.

Materials and Methods

Simulation environments

Our simulation consisted of three phases. The first phase is the learning phase, where an agent learns Q-values in a given environment through interaction with it. The second is the baseline phase, where the agent engages various responses by choices of responses based on the value of rewards and shortest path searches on a network that is constructed from Q-values learned during the learning phase. The third and last phase is the test phase in which the agent engages various responses in the same way as the baseline phase, but the reward obtained from an operant response is devalued.

Learning Phase

In learning, an agent engages various responses in a given environment and learns Q-values through interaction with it. The learning rules are defined by Eqs. 1 and 2. In this phase, because the agent has not learned Q-values and does not construct a network, it chooses a response randomly and transitions to the directory from the current engaging response. After making the transition, the environment gives a reward to the agent, and the agent updates Q-values following Eq. 1 and 2. After the agent completes the learning phase, it constructs a network based on the learned Q-values during the learning phase.

Reward devaluation procedure

The reward devaluation is a procedure to examine whether an operant response is goal-directed or habit under free-operant situations. First, an animal learns that he or she can obtain a reward, food, or sucrose solution by pressing the lever. Learning lever pressings, the animal was placed in an experimental environment and trained to the operant response. After the training, reward devaluation was done by poisoning it with lithium chloride and a brief period was added where the animal can access the reward freely. Then, the animal was put into the experimental environment again and examined whether the number of operant responses decreased. If the number of responses does not change, it implies that the response is no longer controlled by its consequence, and the response becomes a habit. In contrast, if the number of responses decreases, the response is controlled by its consequences, such as goal-directed behavior. In our simulation, to reproduce the procedure, we reduced the value of the reward obtained from the operant response after the baseline phase.

Baseline

In the baseline phase, an agent travels on a network by choosing a response, a goal, following Eq. 4 and searching for the shortest path between a currently engaging response and the goal.

The simulation contains three steps: 1) choice of response based on reward values, 2) searching for the shortest path between the current response and the goal, and 3) engaging responses successively contained in the path. We calculated the proportion of an operant response to the total number of responses after some loops of the above 3 three steps. Algorithm 1 shows the pseudocode of the simulation in the baseline phase.

Algorithm 1 Pseudocode of the baseline phase

INPUT: *N*: number of nodes, *loops*: number of loops per simulation

OUTPUT: *prop*: proportion of the operant response to overall responses

```
Initialize the reward value r obtained by the operant response to 1 and the other to 0.14
Initialize current_response with random integer ranged from 0 to N
Initialize number_of_all_responses, number_of_operant with 0

for i in range(loops):
    goal_response = choose response according to Eq. 4
    shortest_path = find a shortest path from the current_response to the
    goal_response.
    number_of_response += number of responses included in the shortes_path
    if “operant” in shortest_path:
        number_of_{operant} += 1

prop = number_of_operant / number_of_all_responses
return prop
```

Devaluation

In the test phase, the agent behaved in the same way as in the baseline phase. The difference between the test and baseline phases is only the value of the reward obtained from the operant response. In the baseline phase, we set the value to 1, and in the test phase, we set it to 0. At the baseline phase, we calculated the proportion of the number of operant responses to the total number of responses.

Simulation 1

In Simulation 1, we examined the relationship between the structure of the network and habit formation. To construct an arbitrary network structure, we systematically operate Q-values instead of when the agent learns them. We only operated the Q-value of the operant response to examine the effect of the degree of edge concentration on the operant response on habit formation, and other Q-values were the same across each condition. We defined Q-values using the following equation:

$$Q(i,j) = q \times q$$

q denotes a vector where q_0 denotes the Q-value of the operant response, other elements denote other responses, and Q is defined as its outer matrix. We changed q_0 from 0.01 to 0.1, across conditions, and $q_n (n \neq 0)$ was fixed at 0.0028. We constructed a network using Eqs. 3, and run simulations under the baseline and test phases.

Simulation 2

In simulation 2, the agent learns Q-values in an experimental environment and constructs a network based on the Q-values in the learning phase. We conducted simulations with variable interval (VI), variable ratio (VR), concurrent VI VI, and VI with non-contingent rewards. Moreover, we changed the number of rewards in the learning phase to examine the effect of training on habit formation. Each simulation has been further described in the following sections.

VR VI comparison and amount of training

The VR schedule presents rewards depending on the operant response of the agent. At each response, whether the reward is presented at a probability, which is the same as in the simulations. This means that reward presentation follows the Bernoulli process, and the number of responses required to obtain rewards follows the geometric distribution. We generate pseudo-random numbers following the distribution in order for the numbers to converge to the distribution in all simulations. More specifically, we divided the interval ranging from 0 to 1 into equal divisions according to the number of rewards, and the percentile points of the distribution were calculated for each point. Algorithm 2 shows how to generate the required number of responses that follow the geometric distribution.

Algorithm 2 Definition of variable ratio schedule; *generate_VR(p, n)*

INPUT: *p*: parameter for geometric distribution, *n*: number of reward per

simulation

OUTPUT: *rr*: list of required responses per trial

points = Divide the range 0 - 1 into *n* equal parts

rr = Calculate percentile point of geometric distribution for each *points*

return *rr* sorted randomly

The VI schedule presents rewards depending on the time lapse. However, the agent must emit responses to obtain rewards. Reward availability is determined at each time step according to a probability, and once the reward becomes available, it remains available until the agent takes the response. Reward availability follows the Poisson process, and the intervals between each reward follow an exponential distribution. Pseudo-random numbers are generated following the distribution in the same manner as the VR schedule. Algorithm 3 shows how to generate inter-reward intervals that follow an exponential distribution.

Algorithm 3 Definition of variable interval schedule; *generate_VI(p, n)*

INPUT: *lambda*: parameter of exponential distribution, *n*: number of reward per simulation

OUTPUT: *iri*: list of inter reward intervals per trial

points = Divide the range 0 - 1 into *n* equal parts

iri = Calculate percentile point of exponential distribution for each *points*

return *iri* sorted randomly

In the VI schedule, the reward probability increases as the time from the last response increases, and the reward probability in the simulations depends on the duration required by each response. Herein, we assume that each response requires a different duration. However, the same response requires the same duration. We set the available response category to 50, so we determined 50 durations for responses. We generate durations from the exponential distribution where the expected duration is 2.5, and the minimal duration is 0.1; we employ the same method as the VI schedule to generate them. We applied a minimal duration (0.1) as the duration to the operant response and remaining durations to other responses.

Algorithms 4 and 5 show the flow of the simulation in the learning phase with VI and VR. In both schedules, the flow is the same, excluding reward presentation depending on the

response or time. First, response durations, required numbers of response or inter-reward intervals, and Q-values were initialized. Then, the agent chooses a response, and the environment determines whether the reward is presented. If the agent chooses the operant response, the reward presentation is determined by the schedule, and if it chooses another response, the reward is always presented. Then, the agent updates the Q-values using eq. 1 and 2. The simulation is terminated when the agent obtains a predetermined number of rewards and constructs a network using Eq. 3. The number of rewards was 25, 50, 100, and 200. After the learning phase, we conducted the simulation with the baseline and test phases in the same way as in Simulation 1.

Algorithm 4 Learning phase under VI schedule

INPUT: *iri*: mean inter reward interval, *nr*: number of rewards, *nn*: number of nodes

OUTPUT: *net*: generated network under a given environment

```
response_duration = generate_VI(1 / 2.5, nn)
IRI = generate_VI(1 / iri, nr)
Initialize the reward value R obtained by the operant response to 1 and the other to 0.14
Initialize Q_values with nn * nn matrix where all elements are zero
Initialize current_response with random integer ranged from 0 to nn

for current_IRI in IRI:
    while True:
        next_response = choose a response randomly
        count down the current_IRI for the time it takes to emit next_response
        set r to the value of the reward obtained by the next_response
        if current_IRI > 0. and next_response is "operant":
            set r to 0
        calculate temporal difference error according to Equation 2
        update Q value according to Equation 1
        set current_response to next_response
        if current_IRI <= 0 and current_response is "operant":
            break
```

generate a network according to Equation 3 and return it

Algorithm 5 Learning phase under VR schedule

INPUT: *rr*: mean required response to obtain rewards, *nr*: number of rewards, *nn*: number of nodes

OUTPUT: *net*: generated network under a given environment

```
response_duration = generate_VI(1 / 2.5, nn)
RR = generate_VR(1 / rr, nr)
Initialize the reward value R obtained by the operant response to 1 and the other to 0.14
Initialize Q_values with nn * nn matrix where all elements are zero
Initialize current_response with random integer ranged from 0 to nn

for req in RR:
    while True:
        next_response = choose a response randomly
        if next_response is "operant" count down req
        set r to the value of the reward obtained by the next_response
        if req > 0 and next_response is "operant":
            set r to 0
        calculate temporal difference error according to Equation 2
        update Q value according to Equation 1
        set current_response to next_response
        if req <= 0 and current_response is "operant":
            break
generate a network according to Equation 3 and return it
```

Comparison between choice and single schedule

To examine the degree of habit formation when an explicit alternative is given, we used

an environment that mimics the experiment conducted by Kosaki and Dickinson (2010), where the effect of the presence or absence of the alternative on habit formation. Here, the agent can engage in two operant responses, and different rewards are assigned to each response. For example, two levers were inserted into the apparatus and pressing the left lever produced food, and the right levers produced a sucrose solution. In addition, as a control condition, we used an environment in which the agent can engage only one operant response, but the reward unavailable from the operant response is presented independent of the agent responses.

Algorithm 6 shows the flow of the simulation in the learning phase, where an explicit alternative is. The flow is the same as that of the single schedule (Algorithm 4), but there are two differences. First, since there are two operant responses, we assigned two VI schedules with the same value. Second, although the rewards obtained by each response are different, we assumed that the values of both rewards are the same, and it is 1. As in the case of a single schedule, the agent chooses a response randomly, learns Q-values, and constructs a network in the learning phase.

Algorithm 6 Training under the situation with choice between two different operant responses

INPUT: *iri*: mean inter reward interval, *nr*: number of rewards, *nn*: number of nodes

OUTPUT: *net*: generated network under a given environment

response_duration = generate_VI(1 / 2.5, nn)

IRI_left = generate_VI(1 / iri, nr)

IRI_right = generate_VI(1 / iri, nr)

Initialize the reward value *R* obtained by the operant response to 1 and the other to 0.14

Initialize *Q_values* with *nn* * *nn* matrix where all elements are zero

Initialize *current_response* with random integer ranged from 0 to *nn*

set *reward_left* to 0

set *reward_right* to 0

current_IRI_left = IRI_left[reward_left]

current_IRI_right = IRI_left[reward_right]

```
while reward_left + reward_right < nr:
    while True:
        next_response = choose a response randomly
        count down the current_IRI_right and current_IRI_left for the time it takes to emit
        next_response
        set r to the value of the reward obtained by the next_response (R[next_response])
        if current_IRI_left > 0. and next_response is “operant-left”:
            set r to 0
        if current_IRI_right > 0. and next_response is “operant-right”:
            set r to 0
        calculate temporal difference error according to Equation 2
        update Q value according to Equation 1
        set current_response to next_response

        if current_interval_left <= 0. and next_response is “operant-left”:
            reward_left += 1
            current_interval_left = IRI_left[reward_left]
        if current_interval_right > 0. and next_response is “operant-right”:
            reward_right += 1
            current_interval_right = IRI_left[reward_right]

generate a network according to Equation 3 and return it
```

Algorithm 7 shows the flow of the simulation in the learning phase, where only one operant response is available. The value of the response-independent reward is assumed to be 1, and when it is presented, the value is added to the reward obtained by the response the agent actually does. Inter-reward intervals of response-independent rewards are set to be the same as those of the rewards obtained from the operant response so that the total number of rewards obtained in the simulation is not different from that in the simulation with an explicit alternative. The rest of the flow was the same as in the condition with an explicit alternative.

Algorithm 7 Training under no choice situation

INPUT: *iri*: mean inter reward interval, *nr*: number of rewards, *nn*: number of nodes

OUTPUT: *net*: generated network under a given environment

```
response_duration = generate_VI(1 / 2.5, nn)
IRI_op = generate_VI(1 / iri, nr)
IRI_nc = generate_VI(1 / iri, nr)
Initialize the reward value R obtained by the operant response to 1 and the other to 0.14
Set non contingent reward value r_nc to 1
Initialize Q_values with nn * nn matrix where all elements are zero
Initialize current_response with random integer ranged from 0 to nn

set reward_op to 0
set reward_nc to 0

current_IRI_op = IRI_op[reward_op]
current_IRI_nc = IRI_nc[reward_nc]

while reward_op + reward_nc < nr:
    while True:
        next_response = choose a response randomly
        count down the current_IRI_op and current_IRI_nc for the time it takes to emit
        next_response
        set r to the value of the reward obtained by the next_response
        if current_IRI_op > 0. and next_response is “operant”:
            set r to 0
        if current_IRI_nc <= 0.:
            r += r_nc
        calculate temporal difference error according to Equation 2
        update Q value according to Equation 1
```

```
        set current_response to next_response

        if current_IRI_op <= 0. and next_response is "operant":
            reward_op += 1
            current_IRI_op = IRI_op[reward_op]
        if current_IRI_nc <= 0.:
            reward_nc += 1
            current_IRI_nc = IRI_nc[reward_nc]

    generate a network according to Equation 3 and return it
```

After the learning phase, we conducted the baseline and test phases, but the treatment was different from the single schedule because there were two rewards. In the baseline phase of the choice condition, we set the second element of the reward vector to 1, and the rest of the elements were the same as in Algorithm 1. In the test phase, we reduced only the first element of the reward to 0. In the no-choice condition, response-independent rewards are presented randomly, so the value of the reward is distributed among all responses. Specifically, the reward value of 1. divided by the total number of responses (0.02) was added to all elements of the reward vector. In the test phase, only the reward obtained by the operant response is devalued, but since there is also a component of response-independent reward, we subtracted the value of the reward obtained by the operant response from the baseline reward value and considered the value of the reward obtained by the operant response of 0.02, in the test phase.

Acknowledgement

This study was supported in part by Grant-in-Aid for JSPS Fellows (20J21568) to KY and Grants-in-Aid for Scientific Research (18KK0070, 19K03385, 19H01769, 19H05316) to KT from the Japan Society for the Promotion of Science (http://www.jsps.go.jp/english/e-grants). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Reference

Adams, C. D. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology Section B*, *34*(2b), 77-98.

Aldridge, J. W., & Berridge, K. C. (1998). Coding of serial order by neostriatal neurons: a “natural action” approach to movement sequence. *Journal of Neuroscience*, *18*(7), 2777-2787.

Balleine, B. W., & O'doherty, J. P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, *35*(1), 48-69.

Bogacz, R. (2020). Dopamine role in learning and action inference. *Elife*, *9*, e53262.

Breland, K., & Breland, M. (1961). The misbehavior of organisms. *American Psychologist*, *16*(11), 681.

Colwill, R. M., & Rescorla, R. A. (1985). Instrumental responding remains sensitive to reinforcer devaluation after extensive training. *Journal of Experimental Psychology: Animal Behavior Processes*, *11*(4), 520.

Datta, S. R., Anderson, D. J., Branson, K., Perona, P., & Leifer, A. (2019). Computational neuroethology: a call to action. *Neuron*, 104(1), 11-24.

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.

De Houwer, J. (2019). On how definitions of habits can complicate habit research. *Frontiers in Psychology*, *10*, 2642.

Dezfouli, A., & Balleine, B. W. (2012). Habits, action sequences and reinforcement learning. *European Journal of Neuroscience*, *35*(7), 1036-1051.

Dickinson, A. (1985). Actions and habits: the development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *308*(1135), 67-78.

Dickinson, A., Balleine, B., Watt, A., Gonzalez, F., & Boakes, R. A. (1995). Motivational control after extended instrumental training. *Animal Learning and Behavior*, 23, 197–206.

Dickinson, A., Nicholas, D. J., & Adams, C. D. (1983). The effect of the instrumental training contingency on susceptibility to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology*, *35*(1), 35-51.

Falk, J. L. (1966). Schedule - induced polydipsia as a function of fixed interval length 1. *Journal of the Experimental Analysis of Behavior*, *9*(1), 37-39.

Gentry, W. D. (1968). FIXED - RATIO SCHEDULE - INDUCED AGGRESSION 1. *Journal of the Experimental Analysis of Behavior*, *11*(6), 813-817.

Graybiel, A. M. (1998). The basal ganglia and chunking of action repertoires. *Neurobiology of learning and memory*, *70*(1-2), 119-136.

Graybiel, A. M. (2008). Habits, rituals, and the evaluative brain. *Annu. Rev. Neurosci.*, *31*, 359-387.

Guthrie, E. R. (1930). Conditioning as a principle of learning. *Psychological review*, *37*(5), 412.

Hagberg, A., Swart, P., & S Chult, D. (2008). *Exploring network structure, dynamics, and function using NetworkX* (No. LA-UR-08-05495; LA-UR-08-5495). Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Herrnstein, R. J. (1970). On the law of effect 1. *Journal of the Experimental Analysis of Behavior*, *13*(2), 243-266.

Hymowitz, N. (1971). Schedule-induced polydipsia and aggression in rats. *Psychonomic Science*, *23*(3), 226-228.

Jurado-Parras, M. T., Safaie, M., Sarno, S., Louis, J., Karoutchi, C., Berret, B., & Robbe, D. (2020). The dorsal striatum energizes motor routines. *Current Biology*, *30*(22), 4362-4372.

Killeen, P. R., & Fetterman, J. G. (1988). A behavioral theory of timing. *Psychological Review*, *95*(2), 274.

Kosaki, Y., & Dickinson, A. (2010). Choice and contingency in the development of behavioral autonomy during instrumental conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, *36*(3), 334.

Kruglanski, A. W., & Szumowska, E. (2020). Habitual behavior is goal-driven. *Perspectives on Psychological Science*, *15*(5), 1256-1271.

Levitsky, D., & Collier, G. (1968). Schedule-induced wheel running. *Physiology & Behavior*, *3*(4), 571-573.

Markowitz, J. E., Gillis, W. F., Beron, C. C., Neufeld, S. Q., Robertson, K., Bhagat, N. D., ... & Datta, S. R. (2018). The striatum organizes 3D behavior via moment-to-moment action selection. *Cell*, *174*(1), 44-58.

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, *21*(9), 1281-1289.

Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019). Habits without values. *Psychological Review, 126*(2), 292–311.

O'Hare, J. K., Ade, K. K., Sukharnikova, T., Van Hooser, S. D., Palmeri, M. L., Yin, H. H., & Calakos, N. (2016). Pathway-specific striatal substrates for habitual behavior. *Neuron*, *89*(3), 472-479.

Ostlund, S. B., & Balleine, B. W. (2005). Lesions of medial prefrontal cortex disrupt the acquisition but not the expression of goal-directed learning. *Journal of Neuroscience*, *25*(34), 7763-7770.

Perez, O. D., & Dickinson, A. (2020). A theory of actions and habits: The interaction of rate correlation and contiguity systems in free-operant behavior. *Psychological Review*, *127*(6), 945.

Niki, H., & Watanabe, M. (1979). Prefrontal and cingulate unit activity during timing behavior in the monkey. *Brain research*, *171*(2), 213-224.

Schultz, W., Apicella, P., Scarnati, E., & Ljungberg, T. (1992). Neuronal activity in monkey ventral striatum related to the expectation of reward. *Journal of neuroscience*, *12*(12), 4595-4610.

Shidara, M., & Richmond, B. J. (2002). Anterior cingulate: single neuronal signals related to degree of reward expectancy. *Science*, *296*(5573), 1709-1711.

Skinner, B. F. (1948). 'Superstition'in the pigeon. *Journal of experimental psychology*, *38*(2), 168.

Staddon, J. E., & Simmelhag, V. L. (1971). The "supersitition" experiment: A reexamination of its implications for the principles of adaptive behavior.

Tang, C., Pawlak, A. P., Prokopenko, V., & West, M. O. (2007). Changes in activity of the striatum during formation of a motor habit. *European Journal of Neuroscience*, *25*(4), 1212-1227.

Tinbergen, N. (1951). *The study of instinct*. Pygmalion Press, an imprint of Plunkett Lake Press.

Turner, K. M., Svegborn, A., Langguth, M., McKenzie, C., & Robbins, T. (2021). Opposing roles of the dorsolateral and dorsomedial striatum in the acquisition of skilled action sequencing. *bioRxiv*.

Yamada, K., & Kanemura, A. (2020). Simulating bout-and-pause patterns with reinforcement learning. *PLoS One*, *15*(11), e0242201.

Yin, H. H., Knowlton, B. J., & Balleine, B. B. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience*, 19, 181–189.

Yin, H. H., Ostlund, S. B., Knowlton, B. J., & Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience*, *22*(2), 513-523.

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, *7*(6), 464-476.

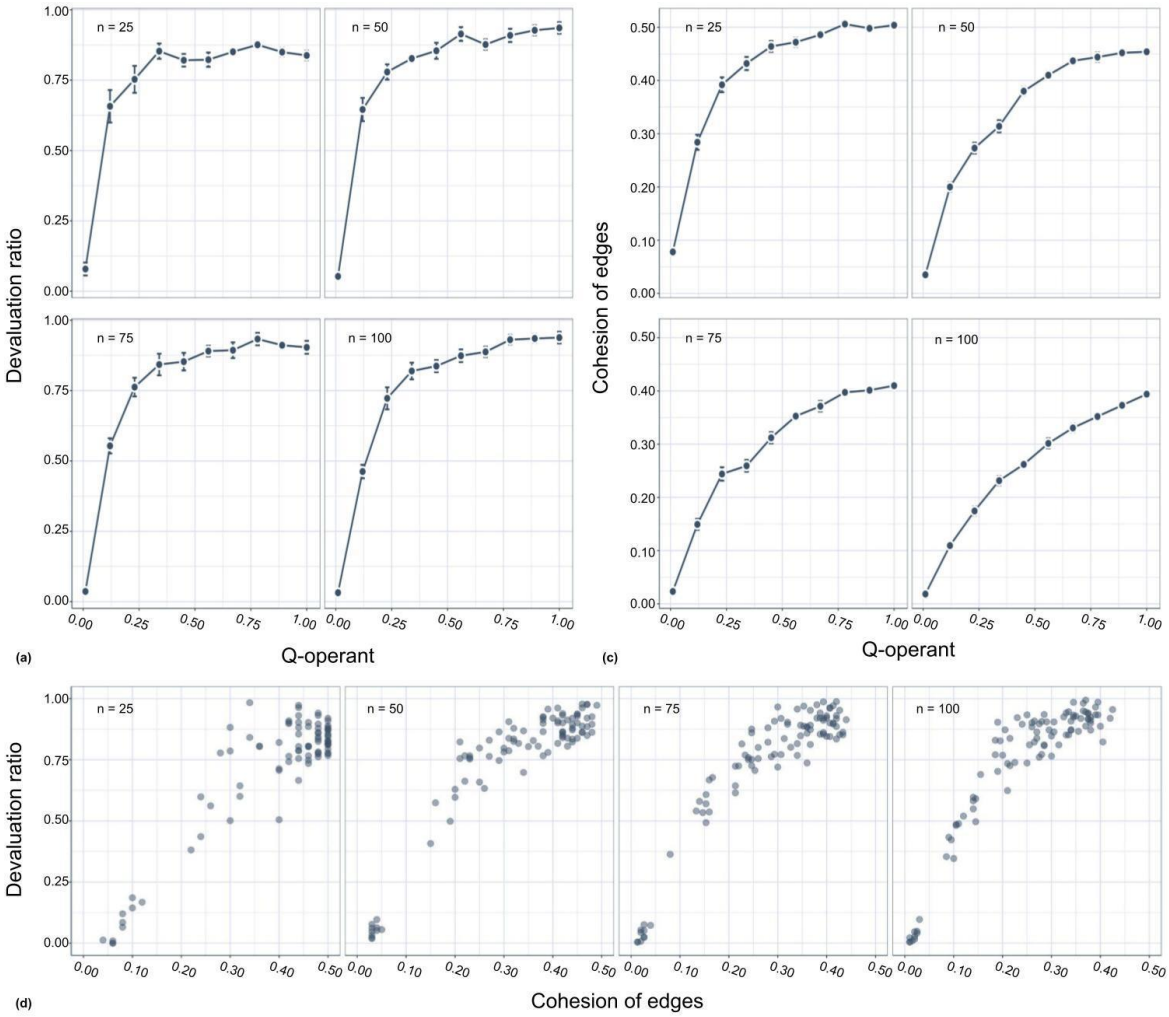
Yin, H. H. (2010). The sensorimotor striatum is necessary for serial order learning. *Journal of Neuroscience*, *30*(44), 14719-14723.

Watanabe, M. (1996). Reward expectancy in primate prefrental neurons. *Nature*, *382*(6592), 629-632.

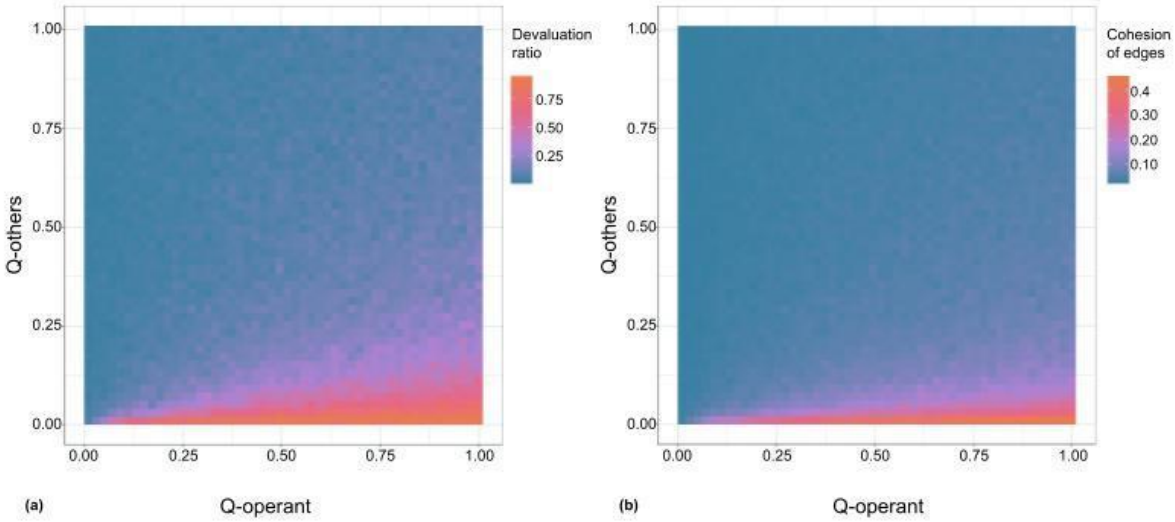
Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4), 279-292.

Wiltschko, A. B., Johnson, M. J., Iurilli, G., Peterson, R. E., Katon, J. M., Pashkovski, S. L., ... & Datta, S. R. (2015). Mapping sub-second structure in mouse behavior. *Neuron*, *88*(6), 1121-1135.

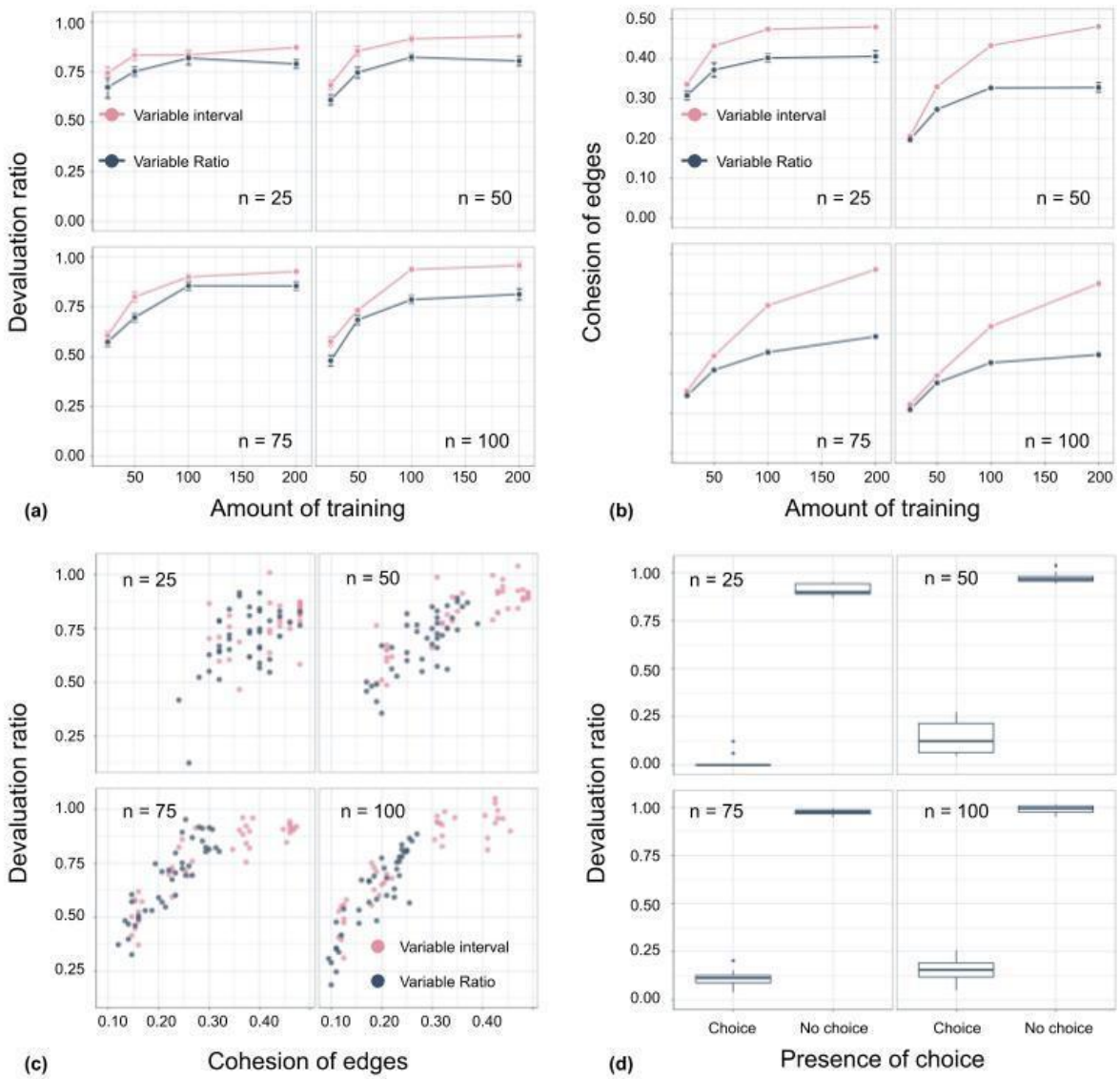
Supplementary materials



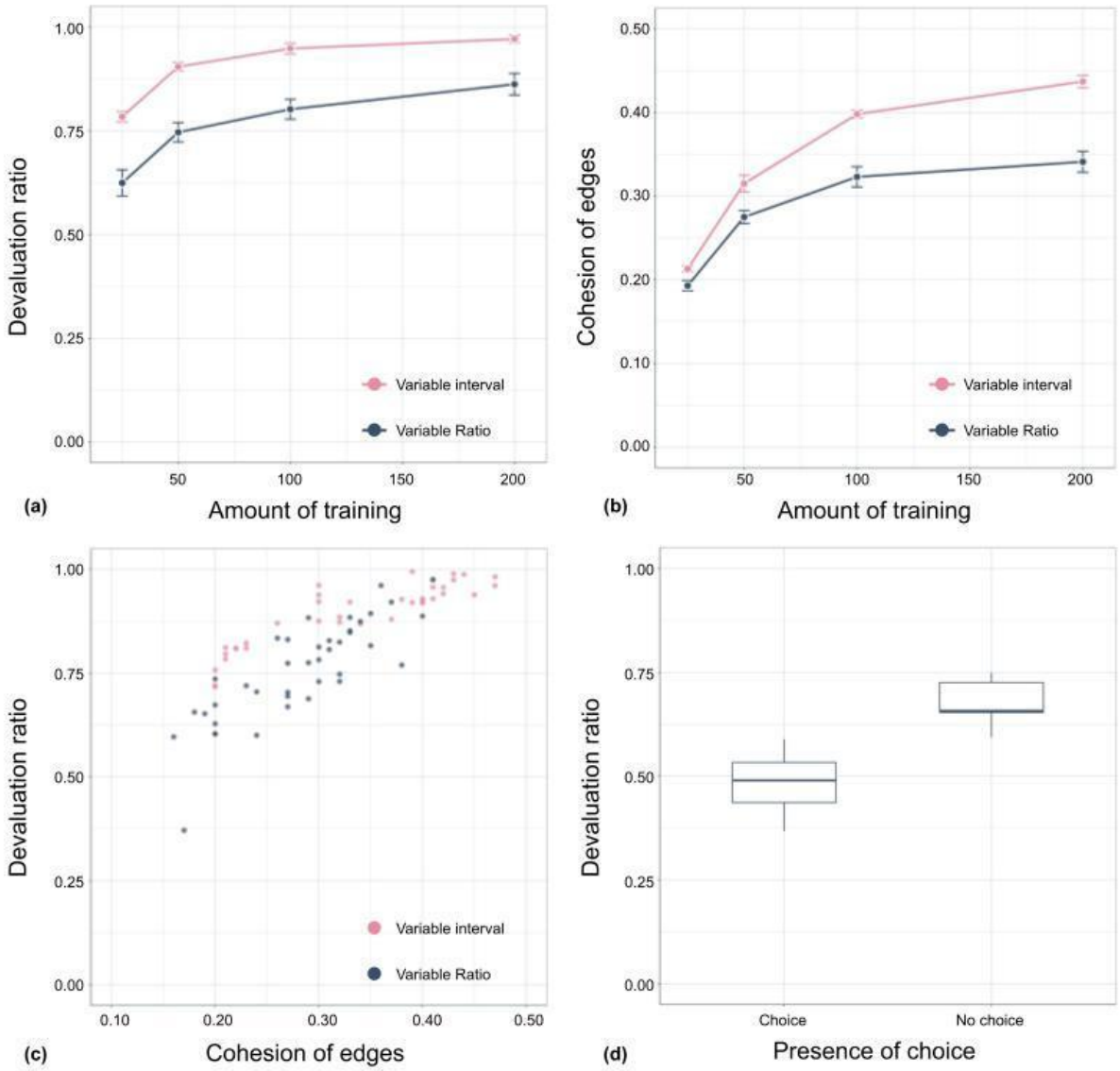
Supplementary figure 1. Simulation results replicating Figure 2 (simulation 1), with the different numbers of nodes (25–100).



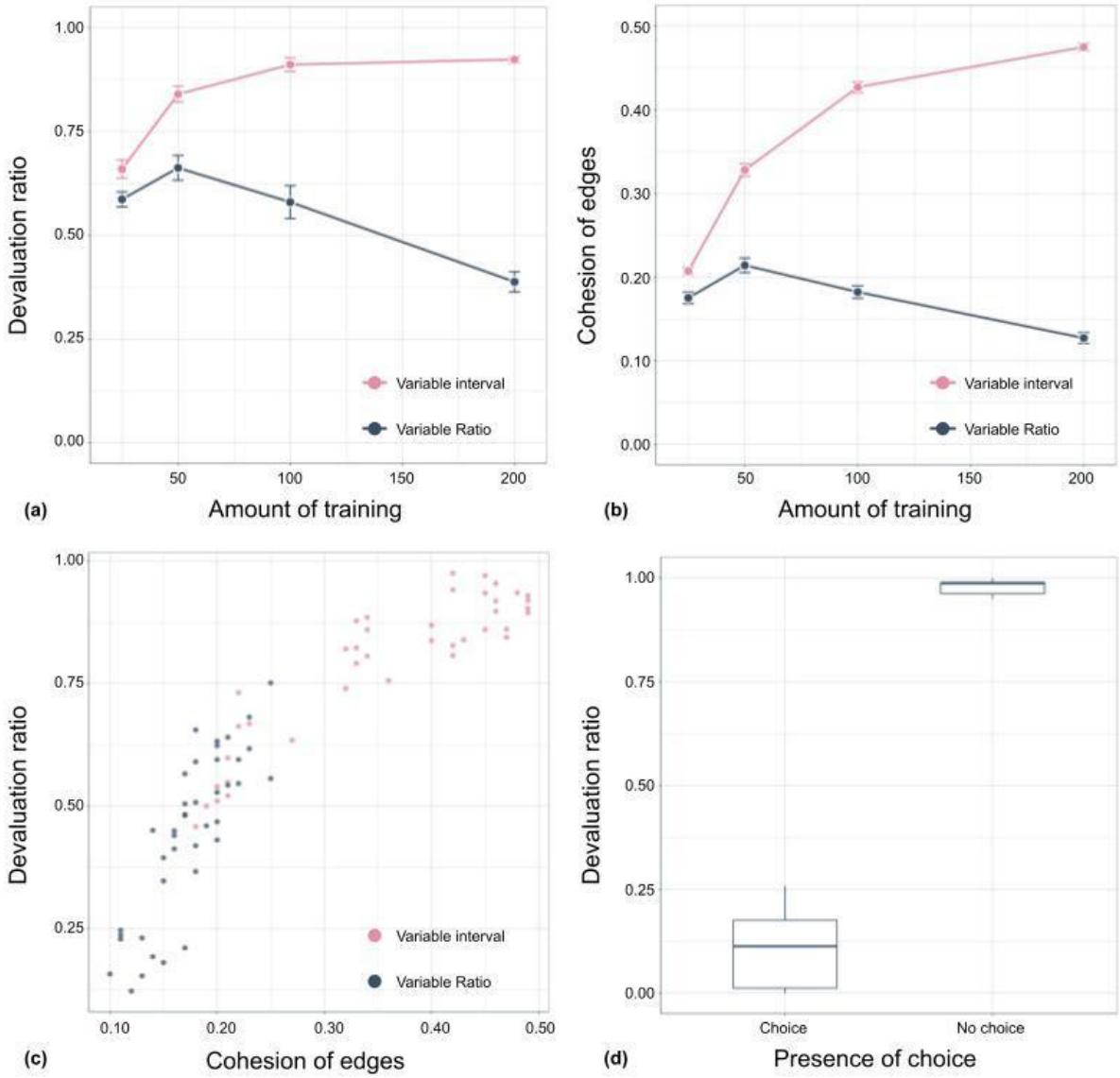
Supplementary figure 2. The dependencies of the devaluation ratio and cohesion of edges in the operant response on the $Q_{operant}$ and Q_{others}.



Supplementary figure 3. Simulation results replicating Figure 3 (simulation 2), with the different numbers of nodes (25–100).



Supplementary figure 4. Simulation results replicating Figure 3 (simulation 2), with a different schedule for other responses from the original simulation. We employed the VI 360 s schedule instead of FR 1 for other responses. We set their reward values as 1 / 50.



Supplementary figure 5. Simulation results replicating Figure 3 (simulation 3), with the different algorithm SARSA from the original algorithm Q-learning.