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 corresponding 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 the behavior as a single network change, this framework provides a new perspective on studying the structure of the behavior for experimental and theoretical research.

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.

Existing theories about habit formation are based on evidence from experimental or theoretical research in psychology and neuroscience. In the canonical view, responses are controlled by two different systems; goal-directed and habit systems. Such theories proposed that goal-directed and habit systems control responses by assigning different weights, and the difference in the weights determines whether the response is goal-directed or habit (Perez and Dickinson, 2020; Daw et al., 2011). In this assumption, habit formation can be viewed as losing control by the consequence of the response or reward sensitivity. However, 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).

In contrast to the canonical view, Dezfouli and Balleine (2012) proposed a new perspective that habit formation can be viewed as shaping or acquiring response sequences. In their model, an agent chooses their goal in a goal-directed manner and generates a response sequence to reach there. Although habits are viewed as a lack of reward sensitivity in the canonical view, their new model considers stereotyped behaviors as acquired response sequences. To what extent could this model change the way of viewing accumulating evidence of habit formation? Garr, Bushra, and Dalamater (2019) shows that rats acquired stereotyped response sequences did not lose reward sensitivity. In a series of studies reported by Dezfouli and Balleine (2012, 2013, 2014) dealt with only a few experiments on the reward sensitivity. These facts suggest that we still need much theoretical and experimental accumulations of evidence to confirm the validity of the new view of habit formation.

Here, instead of assuming two explicitly distinguished mechanisms as in the canonical views, we consider behavior as a network consisting of multiple responses, and show that changes in the structure of the network cause two behavioral features, goal-directed behavior and habit. By doing so, we could explain the lack of reward sensitivity in habit formation, which is a characteristic of the canonical view on habits.

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 press, key peck, nose poke, freezing, salivation, licking, eye blink, etc.) or put them into rigidly controlled situations where they can only engage in the responses to the well-defined stimulus. Another stream is an ethological approach that studies animal behavior under more natural and ecologically valid environments (Tinbergen, 1951). In this case, behavior that the organism is engaged in the real world could be observed, but the stimulus is difficult to control in terms of the strength, frequency, timing, etc. 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). Animals are engaged in more than lever press, key peck, or nose poke, they approach and orient to the stimulus, and walk or sniff around and explore in the given environment. Although the importance of observation and measurement of the behavior during learning was attempted in classic behavioral studies (Guthrie and Horton, 1946; Skinner, 1948; Staddon and Simmelhag, 1971; Jenkins and Moore, 1973), current behavioral quantification methods 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; Leon, et al. 2021). 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 quantification technique of the behavior. 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 experimenter-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.

Network science emerged in the mid to late 1990s and has spread to a wide range of fields. One of the important aspects of network science is handling the structure of the network. For example, in a network in which individual nodes are randomly connected, the distance between each node is large and the information transmission is slow. However, if there is a node called a hub in the network, which has acquired a large number of edges from other nodes, information can be rapidly transmitted through that node. In reality, this is like an influencer sending out information on a social networking service, which attracts the attention of a larger number of users and rapidly spreads the information. In this way, the structure of the network is closely related to the behavior of the entire system. We applicateintroduce this perspective of network structure to behavioral science. In this view, each response is assumed as a node, and behavior could be captured as a network of interconnected nodes. By doing so, we try to explain existing behavioral phenomena from a new perspective of the overall structure of behavior. Introducing the concept of network science to experimental analysis of behavior and the theory of habit formation has not been focused on so far.

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.

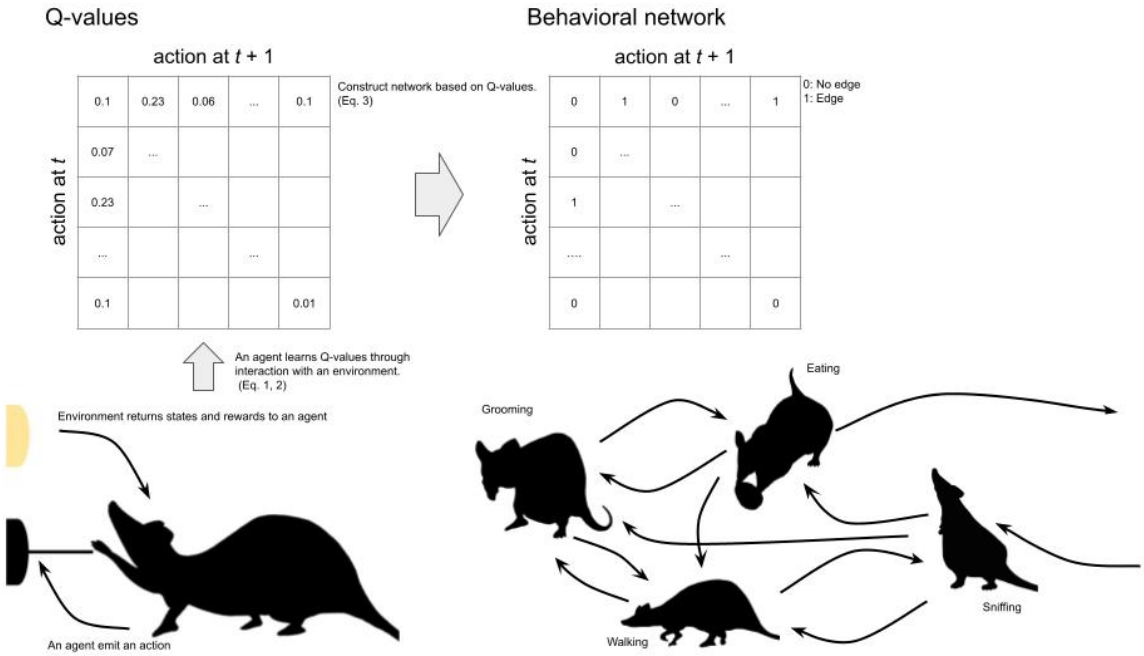


Figure 1. The scheme of the behavioral network.

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,” 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 0.001.

The shortest path search is used to find the shortest path between any two nodes in the network. We employed the Dijkstra’s algorithm (Dijkstra, 1959) 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

First, we generate a network based on the Q-matrix. We used an arbitrary Q-matrix to operate the degree of the edge concentration on the operant response. The Q-matrix is defined as the direct product of the Q-vector. The Q-vector contains scalars ranging from 0. - 1. and each element corresponds to each response. More specifically, The first element corresponds to the operant response and others correspond to the other responses. In simulation 1, we fixed the value for the other responses to 0.001 and varied the value for the operant response, $Q_operant$, from 0.0 - 1.0. 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. To reproduce the procedure, we set up the baseline and devaluation

phases where the value of reward obtained by the operant response is 1 and 0, respectively. In both phases, the first response that the agent engaged was randomly determined. Then, the agent chooses a response based on the reward value and searches for the shortest path to the response from the current engaging response. They engage in responses contained in the path and the agent reaches the chosen response. After it reaches the response, it repeats this process again. After several loops, we calculated the proportion of the operant response to the total number of responses to assess whether the operant response is habit or not.

Simulation result

Fig 2A and supplementary figure 1 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 2B 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 2C 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 2D shows that the probability that the operant response is included in the shortest path between any two nodes in the generated network. As $Q_{operant}$ increases, edges are concentrated into the operant response and it reduces the distance between any two nodes. Thus, regardless of which response the agent chooses, it will engage in the operant response while it reaches the chosen response. Fig 2E 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. These results were reproduced by varying the number of nodes included in the network (Supplementary figure 2), suggesting that the results were not depends on the parameter range.

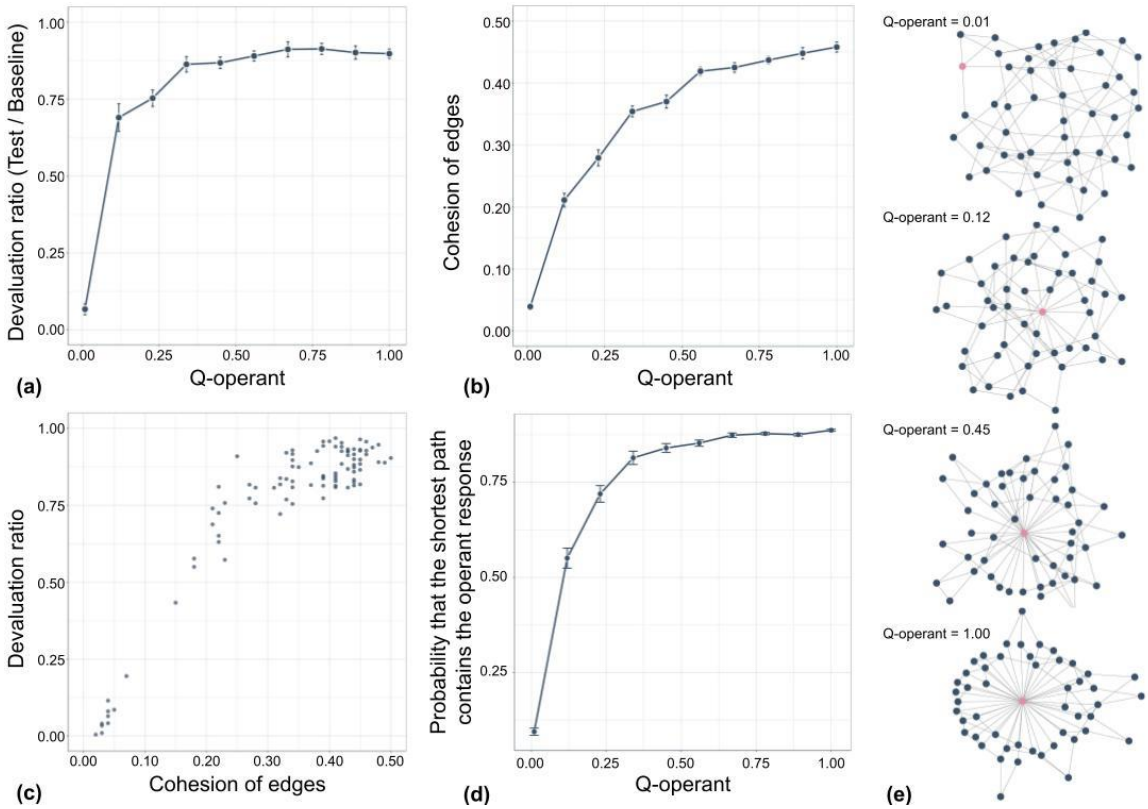


Figure 2. Results of Simulation 1

(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 devaluation 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) The probability that the operant response is included in the shortest path connecting any two nodes in the generated network. (E) 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-matrix and showed that habit formation occurs when edges from other responses are concentrated in the operant response. Fig 2A and Fig 2B show that as $Q_{operant}$ increased, the differences between the baseline and devaluation phases decreased, which implies that habit formation occurs under high $Q_{operant}$. Fig 2C and Fig 2E show that as $Q_{operant}$ increased, the

number of edges acquired by the operant response increased and Fig 2D shows the probability that the operant response is included in the shortest path between any two nodes increased. These results support the idea that habit formation can be explained by the concentration of the edges in the operant response.

As the $Q_{operant}$ increased, the number of edges acquired by the operant response increased. Nodes that acquired a large number of edges, hubs, are present on the network, thereby reducing the distance between nodes present on the network. In such a case, the probability that any shortest paths contain the operant response increased. Thus, if the value of the reward obtained by the operant response was reduced, the agent engaged 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 (more detail in Materials and Methods). After the training phase, the agent experienced the baseline and devaluation phases in the same way as in Simulation 1.

Simulation result

Fig 3A 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 3B 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 3C 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 3D shows that the probability that the operant response is included in the shortest paths in the generated networks under VI and VR schedule respectively. Under both schedules, the probability grew up as the amount of training increased but it is larger in the VI schedule than the VR schedule. Fig 3E shows the Q value of the self-transition of the operant response. It is larger under VR schedule than VI schedule and it implies that consecutive engaging the operant response is less likely to be reinforced in the VI schedule.

Fig 3F 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 3F 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. The number of nodes included in the network had no effect on these results (supplementary figure 3). These results were obtained when FR 1 was employed for other behaviors, but similar results were obtained when a VI schedule was employed (supplementary figure 4). To verify that the results did not depend on the specific algorithm, we employed SARSA instead of Q-learning. Even with the SARSA, we obtained the similar results as Q-learning (supplementary figure 5). Thus, we demonstrate that the results were not depend on the specific number of nodes, the schedule for other responses, and the algorithm used in the simulation.

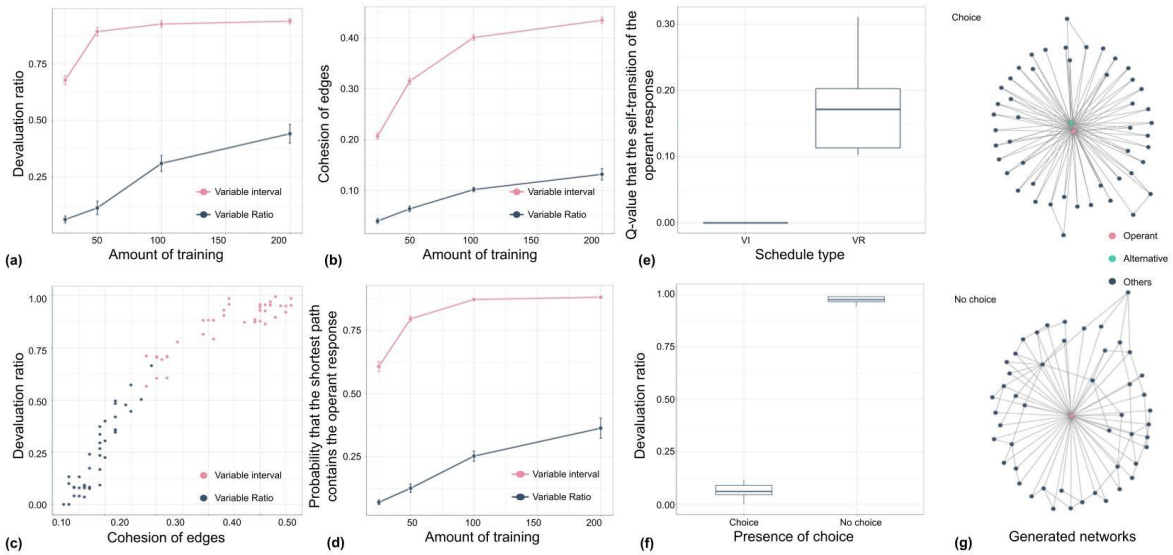


Figure. 3 The results of simulations under VI, VR and presence of choice or not.

(A) shows the changes in the devaluation ratio with increasing amounts 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) The probability that the operant response is included in the shortest path connecting any two nodes in the generated network. (E) Q value of the self-transition of the operant response. (F) Boxplot of devaluation ratio under the choice (left) and no choice (right) situations. (G) 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 3A 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. As we have seen in simulation 1, Fig 3B shows the operant response acquired most of the edges in the network under VI schedule, but not under VR schedule, and it turned out that the probability that the operant response included in the shortest paths grew up under VI schedule (Fig 3D). These results imply that the VI schedule and a large amount of training promote habit formation in the operant response. Fig 3F 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. These characteristics

of both schedules lead to the Q value of the self transition being larger under the VR schedule than VI schedule (Fig 3D) and under the VI schedule, the reward probability of the operant response increases as the agent spends more time on the other responses (Fig 3B). Thus, VI schedules promote habit formation but VR schedules do not.

One might suspect that, contrary to the experimental facts that the response rate is larger in VR schedule than VI schedule, if operant responses acquire more edge in the VI schedule, then the response rate would be higher in the VI schedule as well. However, Fig 3D shows the Q value of the self transition of the operant response is larger in VR schedule than VI schedule. It implies that once an agent starts to engage in an operant response, it will repeat the same response over and over again. In fact, it has been experimentally shown that the difference in response rate between VI and VR schedules is caused by such a mechanism (Shull, Gaynor, and Grimes, 2001; Tanno 2016; Matsui et al; 2018).

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.

Simulation 3: Correlation-based account vs contiguity-based account of habit formation

Here, we propose an experiment to directly test the response-reinforcer correlation, which has been considered as a factor leading to habit formation in the past, and our model's explanation: selective reinforcement of transitions from other behaviors to the operant response and the resulting structural changes in the network. This is a new experiment predicted by our model, which has not yet been examined in real animals, and will encourage future theoretical tests.

From canonical view, response-reward correlation, the operant responses remain goal-directed when animals experience a correlation between the operant responses and rewards, but become habits when they do not experience the correlation (Dickinson 1985; Perez and Dickinson, 2020). Under VR schedule, the more they engage the operant response, the more rewards they can obtain. It leads that they experience positive correlation between the operant response and rewards, and the operant response remains goal-directed. In

contrast, under VI schedule, since rewards availability is governed by time, such correlation is collapsed and they do not experience it. It results that the operant response becomes habit.

In recent years, results have been reported that contradict the response-reward correlation (Corbit, Chieng, and Balleine, 2014; De Russo, et al., 2010; Garr , et al., 2020). For example, De Russo, et al. (2010) trained mice under VI and FI scehdules. FI and VI have a common molar relationship between response rate and rewards: in both schedules, animals can not obtain more than the determined number of rewards within a certain duration, no matter how much they engage in the operant response. Under such a condition, the response-reward correlation view predicts that both schedules guide the same level of habit formation. However, the operant response of mice trained under FI schedule remains goal-directed but under VI schedules, the operant response becomes habit. DeRusso, et al. (2010) conclude that the contiguity, which is defined by average temporal distance between responses and successive rewards, disrupts habit formation. In the FI schedule, animals tend to emit more response as they approach the time when rewards are presented. In contrast, animals do not know when the reward becomes available, they emit responses uniformly during inter-reward intervals in VI schedule. Thus, under the FI schedule, animals emit many responses just before rewards and the contiguity of responses and rewards becomes higher but, under the VI schedule, operant responses are distributed uniformly and the contiguity becomes lower.

A similar discussion has been made for VI-VR response rate difference and there are two kinds of accounts. One explains the difference by the difference in interresponse time that is likely to be rewarded (Wearden and Clark, 1988; Tanno and Silberberg, 2012). In VI schedule, probability of reward availability increases as the elapsed time from last response increases and it results that longer IRTs are more likely to be rewarded than shorter ones. In contrast, such characteristics are not found in the VR schedule or shorter IRTs are more likely to be reinforced. (Fig 4A). Thus, response rate is lower in VI schedule than VR schedule. Especially, the copyist model (Tanno and Silberberg, 2012) explains the difference by average of interresponse times between successive rewards and this is the similar to contiguity-based account of habit formation (DeRusso, et al., 2010; Garr, et al., 2020). Second account is based on the molar relationship between response rate and reward rate (Baum, 1973, 1981). The more animals emit responses under VR schedule, the more rewards they can obtain (Fig 4B). In contrast, under VI schedule, animals can not obtain more rewards than experimentally defined, no matter how they emit responses under the schedule (Fig 4B). This account underlies the response-reward correlation account of habit formation (Dickinson, 1985; Perez and Dickinson, 2020).

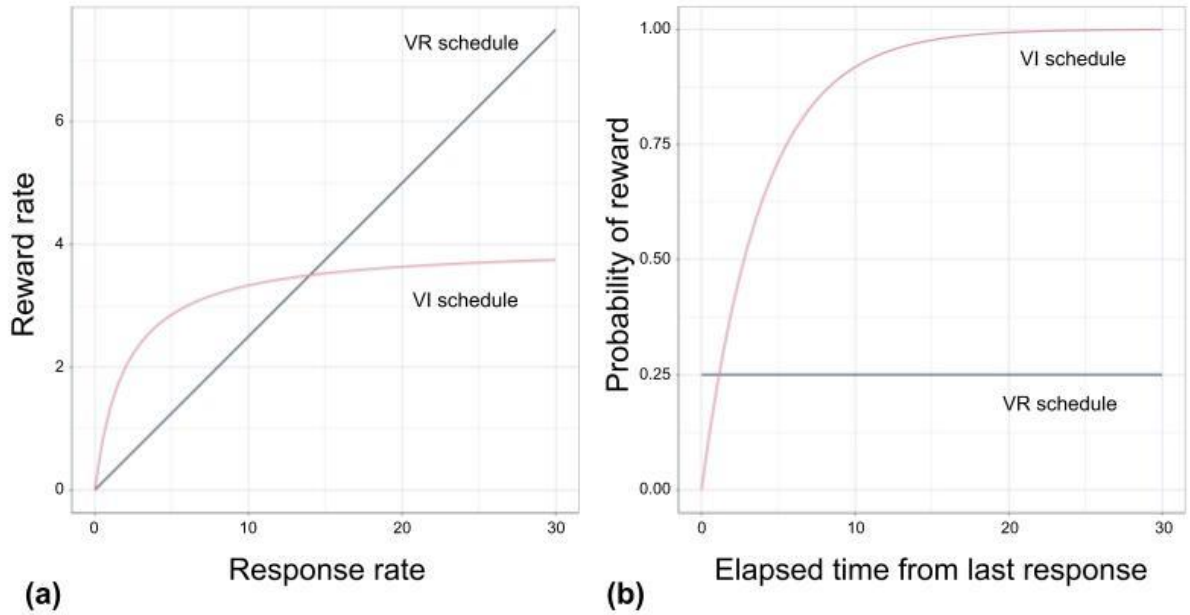


Figure 4. Response rate and reward rate correlation (A) and reward probability as function of elapsed time last response (B) in VR and VI schedules. In VR schedule (black line), reward rate is proportional to response rate, in contrast, reward rates reach a plateau as response rate increases in VI schedule (red line). Reward probability is constant independent from elapsed time from last response in VR schedule, in contrast, it increases exponentially as the time increases.

Our model is positioned similarly to the contiguity-based account in these discussions. As we show in simulation 2, the VI-VR response rate difference can be explained by which transitions are likely to be rewarded: In VI schedule, the transitions from other responses to the operant response are more likely to be rewarded but not in VR schedule (Fig 3E). Viewing the cause of long IRTs as engagement in other responses (Shull, 2011, Yamada and Kanemura, 2020), differential reinforcement of long IRTs can be interpreted as differential reinforcement of the transition from other response to the operant response. Considering these discussions, our model suggests that the same discussions for VI-VR response rate difference can be applied to habit formation.

Here, we mimic an experiment which is conducted to reveal that the VI-VR response rate difference is caused by IRTs immediately followed by rewards (Peele, et al., 1984). In the experiment, pigeons are trained under tandem VI VR and tandem VR VI schedules. The former schedule, tandem VI VR, shares a molar relationship between response rate and reward rate with VI schedule. However, VI schedule is immediately followed by short VR schedule and longer IRTs are less likely to be rewarded than simple VI schedule. The later one is tandem VR VI, it's molar relationship between response rate and reward rate is similar to the simple VR schedule. However, since VR schedule is followed by VI schedule, longer IRTs are more likely to be rewarded. In this schedule, pigeons showed higher response rate

in tandem VI VR schedule and lower in tandem VR VI schedule (Peele, et al., 1984). These findings contradict the account based on response rate and reward rate correlation but well explained by differential reinforcement of IRTs (Tanno, 2012). Will habit formation occur under these schedules? From the view of response-reward correlation, tandem VI VR schedule leads habit but not in tandem VR VI schedule because there is lower response-reward correlation under the former schedule but higher than the later one. In contrast, our model makes the opposite prediction that habit formation will be guided under tandem VR VI schedule but not under tandem VI VR schedule. This is because, in the former schedule, transitions from other responses to the operant response are more likely to be rewarded, and the operant response acquired more edges. In the later schedule, transitions from other response to the operant response and the self transition of the operant response are rewarded in the same probability so the operant response acquired not so many edges.

Simulation result

Fig 5A shows devaluation ratios simulated under VI, tandem VI VR, VR, and tandem VR VI schedules. They were higher under VI and tandem VR VI schedules than VR and tandem VI VR schedules. Although the response-reward correlation account suggests that habit formation is promoted under tandem VR VI schedule and not under tandem VI VR schedule, the results were the opposite, habit formation was promoted under tandem VR VI and but not under tandem VI VR. Fig 5B shows the proportions of edges that the operant response acquired to the overall number of edges in the network and the operant response acquired more edges under VI and tandem VR VI schedules. Fig 5C 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. This is because, under VI and tandem VR VI schedules, the probabilities that the operant response is included in the shortest paths were higher than VR and tandem VI VR schedules.

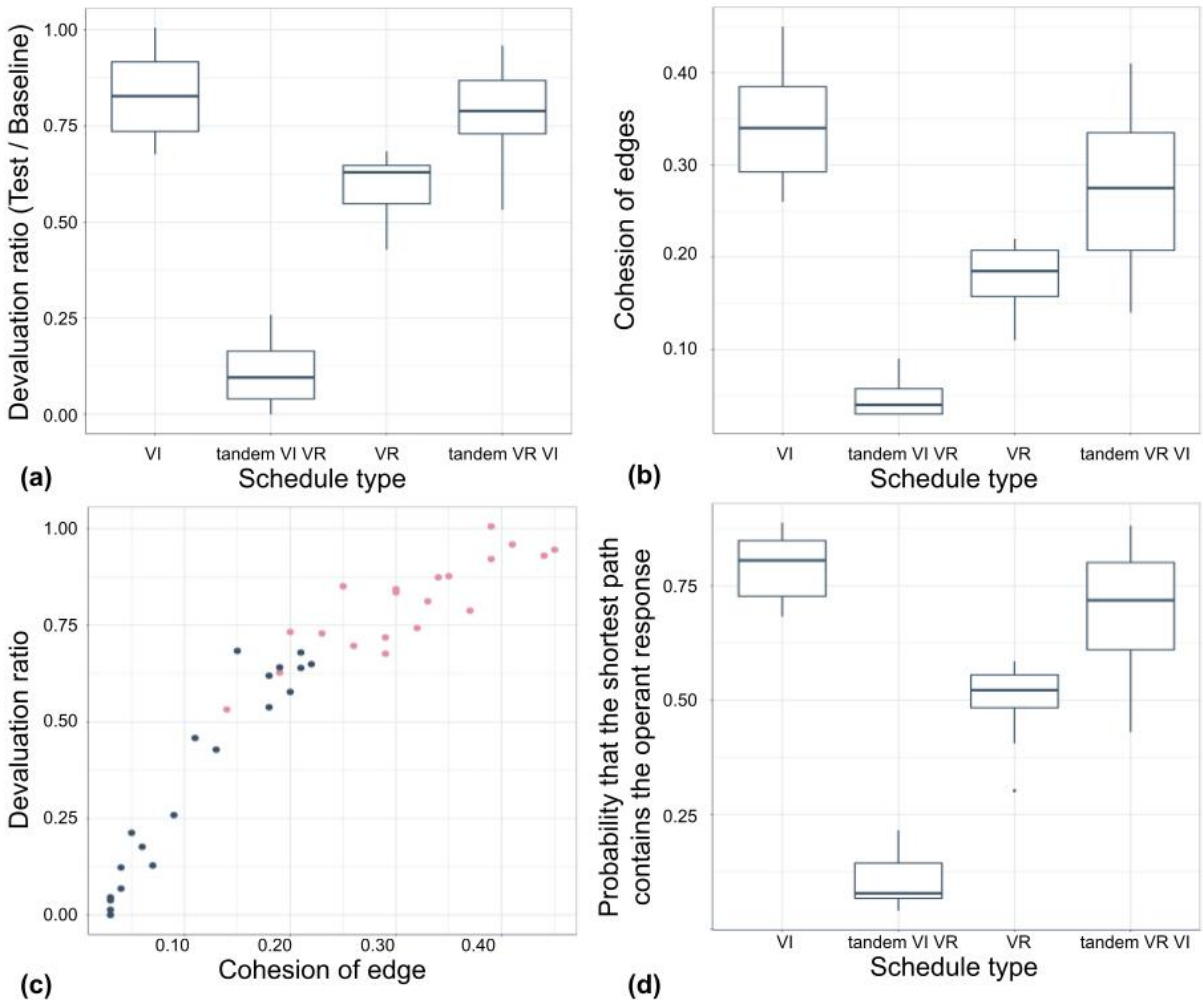


Figure. 5 The simulation results in tandem VI VR and tandem VR VI schedules.

Interim Discussion

In simulation 3, we mimicked the schedules employed by Peele et al. (1984) to reveal what characteristics of schedules promote habit formation. Traditional accounts suggest that lack of the correlation between the response and reward promotes habit formation (Dickinson, 1985; Perez and Dickinson, 2020). In contrast, other researches suggest that contiguity between responses and reward is crucial for habit formation but not the correlation (DeRusso, et al., 2010; Garr, et al., 2020). These two accounts make different predictions in the schedules we employed here. Tandem VR VI and tandem VI VR schedules have different characteristics simple VI and VR schedules have. Tandem VR VI schedule has a common molar relationship between response rate and reward rate with simple VR schedule (black line in Fig 4A) but it also has a time-dependent property, which is found in VI schedule (red line Fig 4B), that the probability of obtaining rewards increases as time elapses. In contrast, tandem VI VR schedule lacks both a molar relationship between response rate and reward rates and time-dependency (red line in Fig 4A and black line in Fig 4B). In the such schedule, animals can not obtain more than the determined number of rewards within a certain duration,

no matter how much they engage in the operant response but the transition from other responses to the operant response is less likely rewarded. If habit formation is determined by the response-reward correlation, habit formation is promoted in tandem VI VR schedule but not in tandem VR VI schedule. However, in contrast to the traditional view, our model predicts that habit formation is more likely promoted in tandem VR VI schedule (Fig 5A). Because of time-dependency tandem VR VI schedule have, transition from others response to the operant response is more likely to be rewarded in the schedule and acquired more edges that simple VR schedule and tandem VI VR schedule (Fig 5B). Thus, as we showed in simulation 1 and 2, the probability that the operant response is included in the shortest paths increased and habit formation occurred (Fig 5C & Fig 5D).

Our model supports the account that the contiguity between responses and rewards promotes habit formation (DeRusso., et al., 2010; Garr. et al., 2020). In tandem VI VR and simple VR schedule, the self transition of the operant response is more likely rewarded than transition from others responses to the operant response. This is because, the operant response occurred as a bout, a burst of responses is followed by long pauses, and this implies that animals emit more responses just before reward presentation. In tandem VR VI and simple VI schedule, the self-transition of the operant response is less likely rewarded because of the time-dependent property between response and reward (Fig 4B red line). This result implies that animals emit less response just before the reward presentation. Thus, response reward contiguity is higher in the tandem VI VR and simple VR schedule than tandem VR VI and simple VI schedule.

Discussion

In this research, 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 goal-directed behavior and habits. 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. These 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. In simulation 3, we analyzed the behavior of the proposed model in an experimental situation where the canonical theory (Dickinson, 1985; Perez and Dickinson, 2020) and the proposed model make different predictions. The results suggest that our model supports the view of reward-responses contiguity promoting habit formation (DeRusso, et al., 2010; Garr, et al., 2020) but not the canonical view of reward-

response correlation.

Relationship to other theoretical models of habit formation

Although there are many models of habit formation, most of them are viewed as goal-directed behavior and habits as interactions between two distinctive associative structures. Here, we succeeded in providing a novel explanation by taking a more molar view of behavior. Specifically, the proposed model substantially differs from existing models in two ways. First, the proposed model does not consider 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 molar 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 responses are determined by 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 any two responses, resulting in habit formation.

Although the proposed model deals with experiments on habit formation in rodents' operant situations (Adams, 1982; Dickinson, et al., 1995; Dickinson, Nicholas, and Adams, 1983; Colwill and Rescorla, 1985; Kosaki and Dickinson, 2010), most of the experiments discussed here are also dealt in Perez and Dickison (2020). Both models reproduce results that are consistent with the experimental results. Perez and Dickison (2020) provide an explanation based on reward-response correlations. In their model, the lower the correlation between response and reward, the more habit formation is promoted. On the other hand, the proposed model provides an explanation based on contiguity between response and reward (DeRusso, et al., 2010; Garr, et al., 2020). Contiguity is defined by the temporal distance between the reward and the emitted response to obtain it. The lower the contiguity, the longer the temporal distance between the response and the reward, the more habit formation is promoted. Although the proposed model does not explicitly incorporate contiguity as a variable in the model, it allows for a similar representation by dividing the agent's behavior into the operant responses and other responses, and separating transitions to the operant responses into self-transitions and transitions from other responses. For example, in a schedule with low reward-response contiguity, such as the VI schedule, transitions from other behaviors to the operant are more likely to be reinforced, while in a VR schedule with high contiguity, transitions from other behaviors are less likely to be reinforced. As a result, the operant response obtains

more edges and promotes habit formation in schedules with low contiguity. As an experiment in which these two factors can be more clearly separated, we employed the procedure of Peele et al (1984). Under this procedure, correlation-based and contiguity-based explanations provide opposite predictions. The proposed model reproduced the same results as predicted by the contiguity-based explanation. Whether habit formation occurs under this experimental procedure has yet to be examined, but it does provide useful insights for updating the theory of habit formation.

The proposed model may seem similar to the model of Dezfouli and Balleine (2013). In fact, their model and our proposed model have two common assumptions. First, instead of treating the agent's behavior as a single response, the two models explicitly assume other responses. They explain habit formation in terms of the acquisition of those sequences or the structure of the network. The second point is that the agent generates sequences or searches for the shortest path based on the value of the reward. On the other hand, there are two differences between the two models. Dezfouli and Balline (2013) uses hierarchical reinforcement learning to explicitly learn a sequence of responses. The proposed model uses simple Q-learning and Sarsa (supplementary figure 5), which does not explicitly include the process of learning sequences. Another point is the treatment of other responses. Dezfouli and Balline (2013) designed their model with a two-stage Markov decision task for humans, in which other responses are limited to those that are both distinct and experimentally possible. On the other hand, the proposed model treats not only the experimentally assumed reactions but also any reactions that the organism can engage in as other reactions, and does not refer to the specific properties of each reaction. By ignoring specific properties, the proposed model emphasizes the molar characteristics of the network.

Dezfouli and Balline (2012) applied their model to the free operant situation and reproduced the effect of amount of training on habit formation. However, they did not treat how other factors, schedule types and presence of alternatives, affect habit formation. The proposed model, which shares common assumptions with their model, can reproduce the results reported in empirical study (Adams, 1982; Dickinson, et al., 1995; Dickinson, Nicholas, and Adams, 1983; Colwill and Rescorla, 1985; Kosaki and Dickinson, 2011). Moreover, the model clarifies the difference between the canonical correlation-based account and common points with the contiguity-based account.

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). 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 its importance in the learned behavior, DLS also encodes innate response sequences (Aldridge and Berridge, 1998). These facts imply that habit formation and 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 the reward presentation (Falk, 1966; Hymowitz, 1971; Gentry, 1968; Levitsky, 1968). 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 assume 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, our model 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 need to 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 function 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

Overview

We conducted three simulations in this article and they contain four steps (Fig 6). In the first step, our agents are given a hypothetical Q matrix in the simulation 1 or learn the Q matrix in the given environments in the simulation 2 and 3. In the second step, the agents generate a network based on the Q matrix. The way to generate the network is the same in all simulations. In the third step, baseline, the agents travel in the network and engage responses. Here, the agents choose their responses based on reward values and the reward value obtained by the operant response is set to 1.0. The agents no longer update the Q matrix nor reconstruct the network. In the final step, devaluation, the agents behave in the same way as the baseline. However, the reward value of operant response is reduced to 0.0. The only difference between baseline and devaluation is the reward value of operant response. We explain the procedures conducted in the four steps in detail after sections. Our simulation codes are available at: https://github.com/7cm-diameter/bnet.

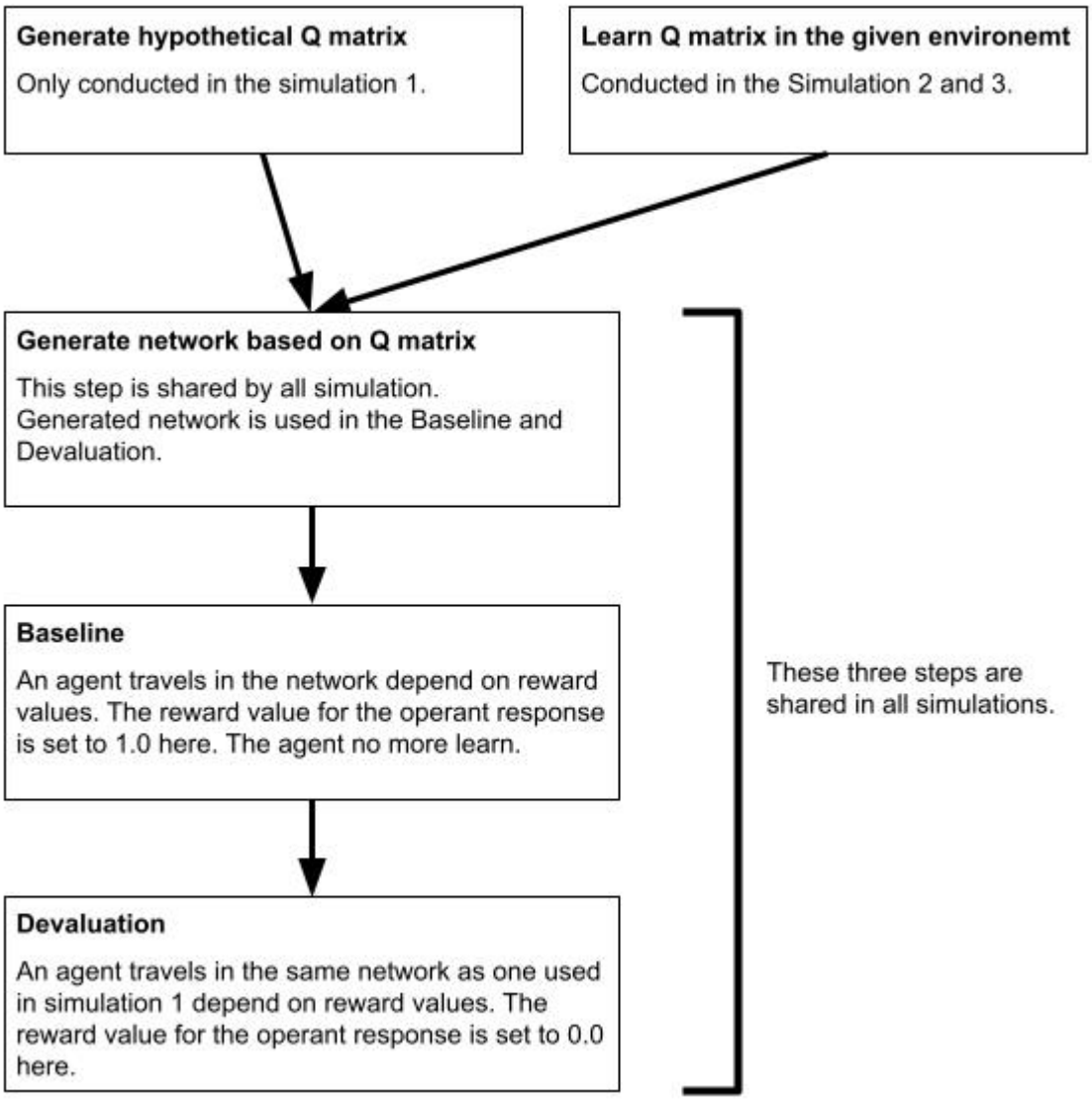


Figure 6. Overview of simulations.

Generate hypothetical Q matrix

Here, the agents are given a hypothetical Q matrix instead of learning it through interactions between an environment. First, we determine the number of nodes contained in a network. We assign a scalar value for each node and it is represented by a vector. The first element of the vector denotes the operant response and other elements denote other responses. The values of the other responses are fixed to 0.01. The value of the operant response is ranged from 0.01 to 1.0. The Q matrix is then defined as the direct product of the Q vector.

Learn Q matrix in the given Environment

In simulation 2 and 3, agents learn the Q-matrix in an experimental environment. In the simulation 2, 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. In simulation 3, we conducted simulations with tandem VI VR and tandem VR VI.

In all these simulations, the agent chooses a response and the environment provides a reward based on the response. Then, the agent updates the Q-matrix according to the response and the reward. In all simulations, we employ fixed ratio (FR) 1 for other responses, where the agent can obtain rewards every time it engages in the responses and the reward values are 0.001 for all other responses. These flows are the same in all simulations. The only difference between the simulations is the schedule in which the environment gives rewards to the agents. Algorithm 1 describes general flow of all simulations. In the following sections, we explain the differences in the schedules for each simulation.

Algorithm 1 Learn Q-matrix in the given environment

```
INPUT: N, schedule, amout_training
OUTPUT: Q
Q = generate an N×N matrix with all elements set to zero.
a_{t-1} = choose a response randomly from [0, 1, 2, …, N]
response_durations = sample N samples from exponential distribution with λ = 1 / 2.5.
r_operant = 0
while r_operant < amout_training:
    a_t = choose a response randomly
    τ = engage the chosen response for response_durations[a_t] seconds.
    r_t = schedule receive a_t and τ and return a reward
    update Q(a_t, a_{t+1}) according to Eq. 1
    if a_t == 0 and r_t == 1:
        r_operant += 1
    a_{t-1} = a_t
return Q
```

VR VI comparison and amount of training

The VR schedule presents rewards depending on the response of the agent. At each response, the reward is presented at a given 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. We employ VR 15 in simulation 2.

Algorithm 2 Variable ratio schedule

```
INPUT: p, rewards, amout_training
OUTPUT: schedule
q = Divide the range 0 - 1 into N equal parts
reward_count ← 0
required_response = q[reward_count] th quantile of geometry distribution with parameter p.
def schedule(a_t, _τ):
    if a_t == 0:
        required_response -= 1
    else:
        return rewards[a_t]
    if required_response <= 0:
        reward_count += 1
        required_response = q[reward_count] th quantile of geometry distribution with parameter p.
        return rewards[a_t]
return schedule
```

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. Moreover, we examined the effect of the amount of training on habit formation by manipulating the number of rewards in both schedules. We calculated the average of inter-reward intervals in the VR schedule and used them as the parameter of VI schedule.

Algorithm 3 Variable interval schedule

```
INPUT: λ, rewards, amout_training
OUTPUT: schedule
q = Divide the range 0 - 1 into N equal parts
reward_count = 0
required_time = q[reward_count] th quantile of exponential distribution with parameter λ.
def schedule(a_t, τ):
    required_time -= τ
    if a_t == 0 and required_time <= 0:
        reward_count += 1
        required_time = q[reward_count] th quantile of exponential distribution with parameter λ.
        return rewards[a_t]
    elif a_t != 0:
        return rewards[a_t]
return schedule
```

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.

We mimicked these experiments. In the choice condition, two of the responses were treated as operant responses, and assigned two VI schedules with the same value and the reward values obtained from both were set to 1.0. In the no-choice condition, the operant response was assumed to be one, but the reward was presented independently of the response in order to control the reward amount. We assigned a variable time schedule to the rewards that are presented independent from the agent responses. We employ concurrent VI x VI x in the choice condition, and concurrent VI 60 VT 60 in the no choice condition.

Algorithm 4 Concurrent VI VI schedule

```
INPUT: λ, rewards, amout_training
OUTPUT: schedule
q_1 = Divide the range 0 - 1 into N equal parts
q_2 = Divide the range 0 - 1 into N equal parts
reward_count_1 = 0
reward_count_2 = 0
required_time_1 = q_1[reward_count_1] th quantile of exponential distribution with parameter λ.
required_time_2 = q_2[reward_count_2] th quantile of exponential distribution with parameter λ.
def schedule(a_t, τ):
    required_time_1 -= τ
    required_time_2 -= τ
    if a_t == 0 and required_time_1 <= 0:
        reward_count_1 += 1
        required_time_1 = q_1[reward_count_1] th quantile of exponential distribution with parameter λ.
        return rewards[a_t]
    if a_t == 1 and required_time_2 <= 0:
        reward_count_2 += 1
        required_time_2 = q_2[reward_count_2] th quantile of exponential distribution with parameter λ.
        return rewards[a_t]
    elif a_t != 0:
        return rewards[a_t]
return schedule
```

Algorithm 5 Concurrent VI VT schedule

```
INPUT: λ, rewards, amout_training
OUTPUT: schedule
q_1 = Divide the range 0 - 1 into N equal parts
q_2 = Divide the range 0 - 1 into N equal parts
reward_count_1 = 0
reward_count_2 = 0
required_time_1 = q_1[reward_count_1] th quantile of exponential distribution with parameter λ.
required_time_2 = q_2[reward_count_2] th quantile of exponential distribution with parameter λ.
def schedule(a_t, τ):
    required_time_1 -= τ
    required_time_2 -= τ
    if a_t == 0 and required_time_1 <= 0:
        reward_count_1 += 1
        required_time_1 = q_1[reward_count_1] th quantile of exponential distribution with parameter λ.
        reward = rewards[a_t]
    elif a_t != 0:
        reward = rewards[a_t]
    if required_time_2 <= 0:
        reward_count_2 += 1
        required_time_2= q_2[reward_count_2] th quantile of exponential distribution with parameter λ.
        reward += rewards[1]
return schedule
```

Tandem VI VR and tandem VR VI

The tandem schedule is a schedule that presents multiple schedules in temporal succession. For example, tandem FR 5 VI 30 means that VI 30 will start after the agent has responded 5 times, and the reward will be presented at the end of VI 30. In addition, since tandem does not provide any explicit cues about the components it consists of, the agent can not know which schedule it is under. In tandem VI VR, the agent is first placed under a VI schedule, and after it is finished, it is moved to a VR schedule. In tandem VR VI the order of components are reversed, starting with the VI schedule and followed by the VR schedule. We employ tandem VI 15 VR 3 and VR 10 VI 5.

Algorithm 6 Tandem VR VI schedule

```
INPUT: p, λ, rewards, amout_training
OUTPUT: schedule
q = Divide the range 0 - 1 into N equal parts
reward_count = 0
required_response = q[reward_count] th quantile of geometry distribution with parameter p.
required_time = q[reward_count] th quantile of geometry distribution with parameter λ.
def schedule(a_t, τ):
    if a_t == 0:
        required_response -= 1
    else:
        return rewards[a_t]
    if required_response <= 0:
        required_time -= τ
    if required_time <= 0. and a_t == 0:
        reward_count += 1
        required_response = q[reward_count] th quantile of geometry distribution with parameter p.
        required_time = q[reward_count] th quantile of geometry distribution with parameter λ.
        return rewards[a_t]
return schedule
```

Algorithm 7 Tandem VI VR schedule

```
INPUT: p, λ, rewards, amout_training
OUTPUT: schedule
q = Divide the range 0 - 1 into N equal parts
reward_count = 0
required_response = q[reward_count] th quantile of geometry distribution with parameter p.
required_time = q[reward_count] th quantile of geometry distribution with parameter λ.
def schedule(a_t, τ):
    required_time -= τ
    if a_t == 0 and required_time <= 0:
        required_response -= 1
    elif a_t != 0:
        return rewards[a_t]
    if required_response <= 0:
        reward_count += 1
        required_response = q[reward_count] th quantile of geometry distribution with parameter p.
        required_time = q[reward_count] th quantile of exponential distribution with parameter λ.
        return rewards[a_t]
return schedule
```

Baseline and Devaluation

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 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.

Devaluation

In the devaluation phase, the agent behaved in the same way as in the baseline phase. The difference between the devaluation 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. Algorithm 8 describes the procedure of the baseline and devaluation phase.

Algorithm 8 Procedure of baseline and devaluation phases

```
INPUT: Network, N, rewards, loop
OUTPUT: propotion_operant
s = choose an initial response from 0 - N randomly
operant = 0
total = 0
for _ in 1:loop:
    t = choose a response according to Eq. 4
    shortest_path = find a shortest path from s to t on the Network
    total += number of response contained in shortest_path
    if 0 in shortest_path:
        operant += 1
    s = t
return operant / total
```

Acknowledgement

This research was supported by JSPS KAKENHI 20J21568 (KY), 18KK0070(KT), 19H05316 (KT), 19K03385 (KT), 19H01769 (KT), Keio Academic Development Fund (KT), Keio Gijuku Fukuzawa Memorial Fund for the Advancement of Education and Research (KT).

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.

Baum, W. M. (1973). The correlation-based law of effect 1. *Journal of the experimental analysis of behavior*, *20*(1), 137-153.

Baum, W. M. (1981). Optimization and the matching law as accounts of instrumental behavior. *Journal of the experimental analysis of behavior*, *36*(3), 387-403.

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.

Corbit, L. H., Chieng, B. C., & Balleine, B. W. (2014). Effects of repeated cocaine exposure on habit learning and reversal by N-acetylcysteine. *Neuropsychopharmacology*, *39*(8), 1893-1901.

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.

DeRusso, A., Fan, D., Gupta, J., Shelest, O., Costa, R. M., & Yin, H. H. (2010). Instrumental uncertainty as a determinant of behavior under interval schedules of reinforcement. *Frontiers in integrative neuroscience*, *4*, 17.

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

Dezfouli, A., & Balleine, B. W. (2013). Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized. *PLoS computational biology*, *9*(12), e1003364.

Dezfouli, A., Lingawi, N. W., & Balleine, B. W. (2014). Habits as action sequences: hierarchical action control and changes in outcome value. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130482.

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.

Dijkstra, E. W. (1959). *Communication with an automatic computer* (Doctoral dissertation, Excelsior).

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.

Garr, E., Bushra, B., Tu, N., & Delamater, A. R. (2020). Goal-directed control on interval schedules does not depend on the action–outcome correlation. Journal of Experimental Psychology: Animal Learning and Cognition, 46(1), 47.

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.

Guthrie, E. R., & Horton, G. P. (1946). Cats in a puzzle box.

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.

Jenkins, H. M., & Moore, B. R. (1973). THE FORM OF THE AUTO-SHAPED RESPONSE WITH FOOD OR WATER REINFORCERS 1. *Journal of the experimental analysis of behavior*, *20*(2), 163-181.

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.

Leon, A., Hernandez, V., Lopez, J., Guzman, I., Quintero, V., Toledo, P., ... & Escamilla, E. (2021). Beyond single discrete responses: An integrative and multidimensional analysis of behavioral dynamics assisted by Machine Learning. *bioRxiv*.

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.

Peele, D. B., Casey, J., & Silberberg, A. (1984). Primacy of interresponse-time reinforcement in accounting for rate differences under variable-ratio and variable-interval schedules. *Journal of experimental psychology: Animal behavior processes*, *10*(2), 149.

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.

Shull, R. L. (2011). Bouts, changeovers, and units of operant behavior. *European Journal of Behavior Analysis*, *12*(1), 49-72.

Shull, R. L., Gaynor, S. T., & Grimes, J. A. (2001). Response rate viewed as engagement bouts: Effects of relative reinforcement and schedule type. *Journal of the experimental analysis of behavior*, *75*(3), 247-274.

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.

Tanno, T. (2016). Response-bout analysis of interresponse times in variable-ratio and variable-interval schedules. *Behavioural processes*, *132*, 12-21.

Tanno, T., & Silberberg, A. (2012). The copyist model of response emission. *Psychonomic Bulletin & Review*, *19*(5), 759-778.

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.

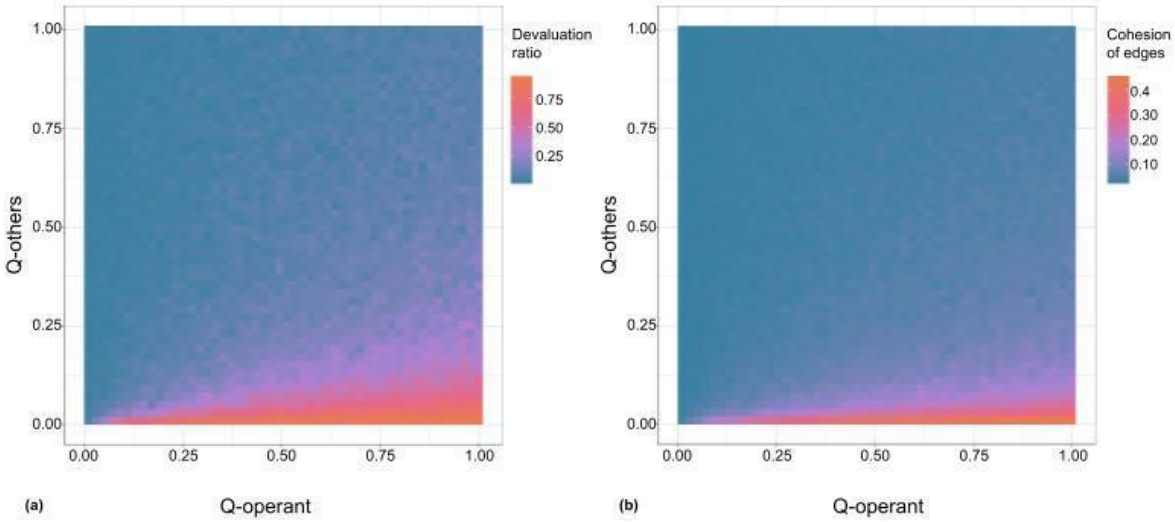
Watanabe, M. (1996). Reward expectancy in primate prefrental neurons. *Nature*, *382*(6592), 629-632.

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

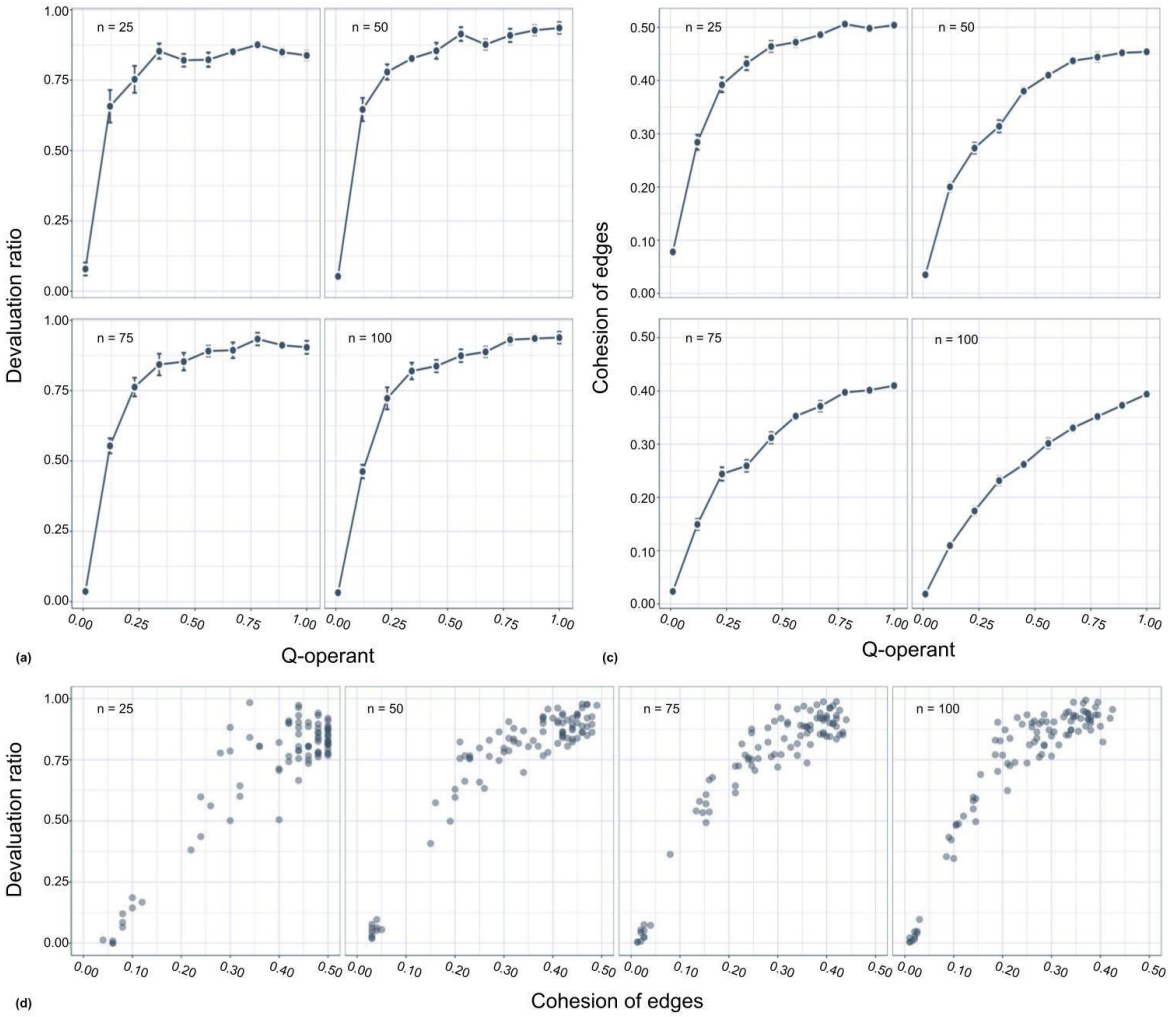
Wearden, J. H., & Clark, R. B. (1988). Interresponse-time reinforcement and behavior under aperiodic reinforcement schedules: A case study using computer modeling. *Journal of Experimental Psychology: Animal Behavior Processes*, *14*(2), 200.

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.

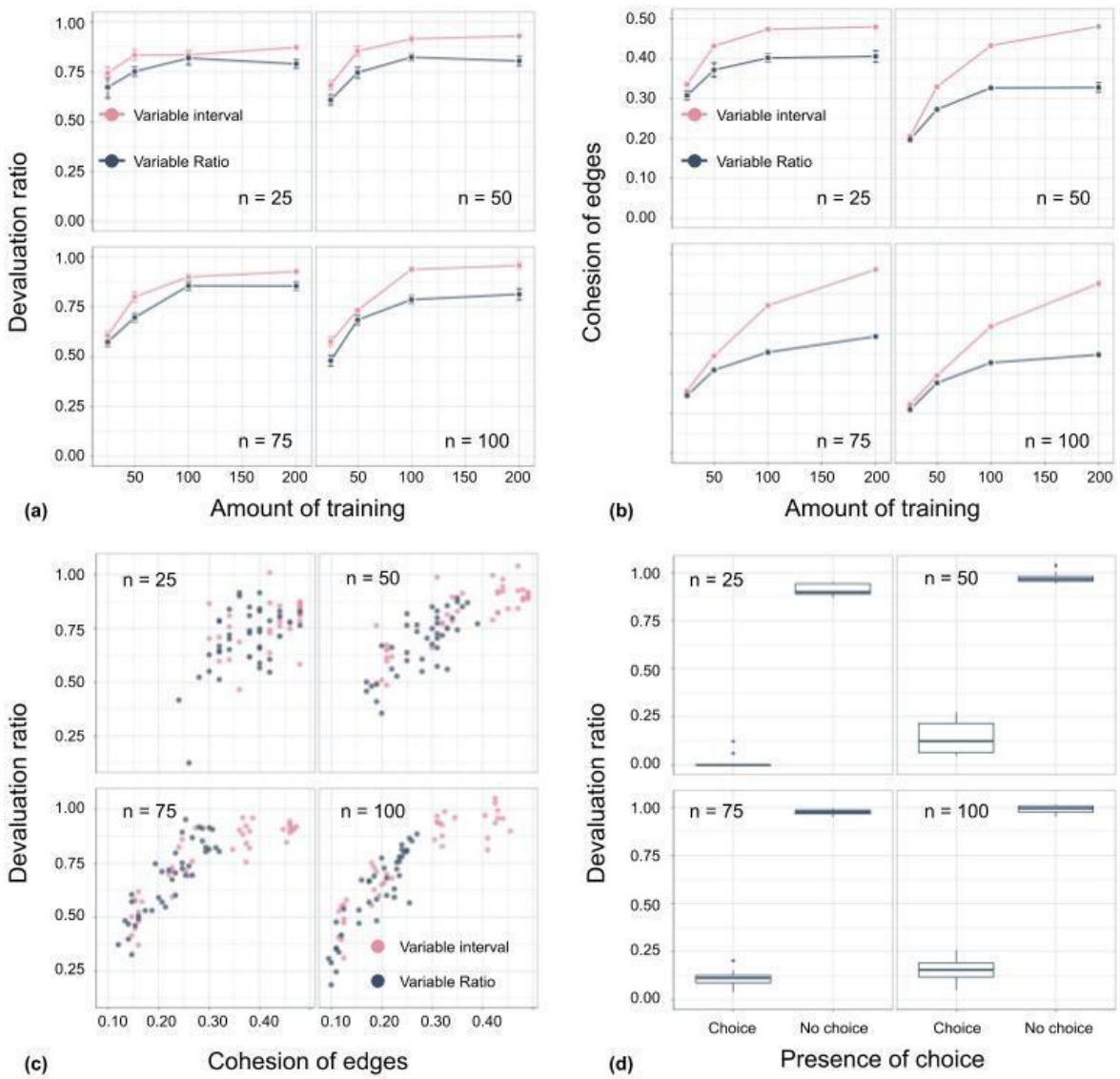
Supplementary materials



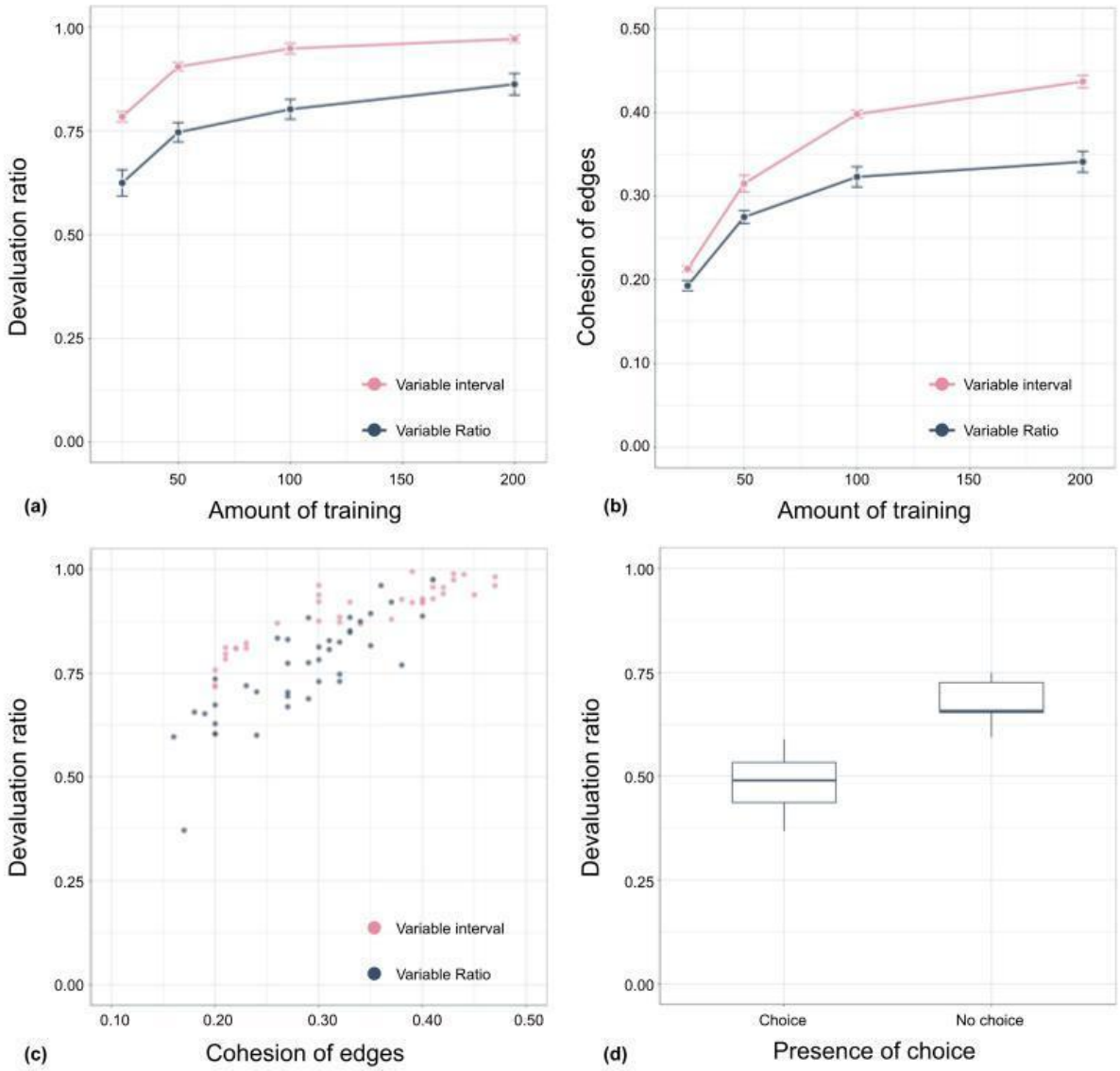
Supplementary figure 1. The dependencies of the devaluation ratio and cohesion of edges in the operant response on the $Q_{operant}$ and Q_{others}.



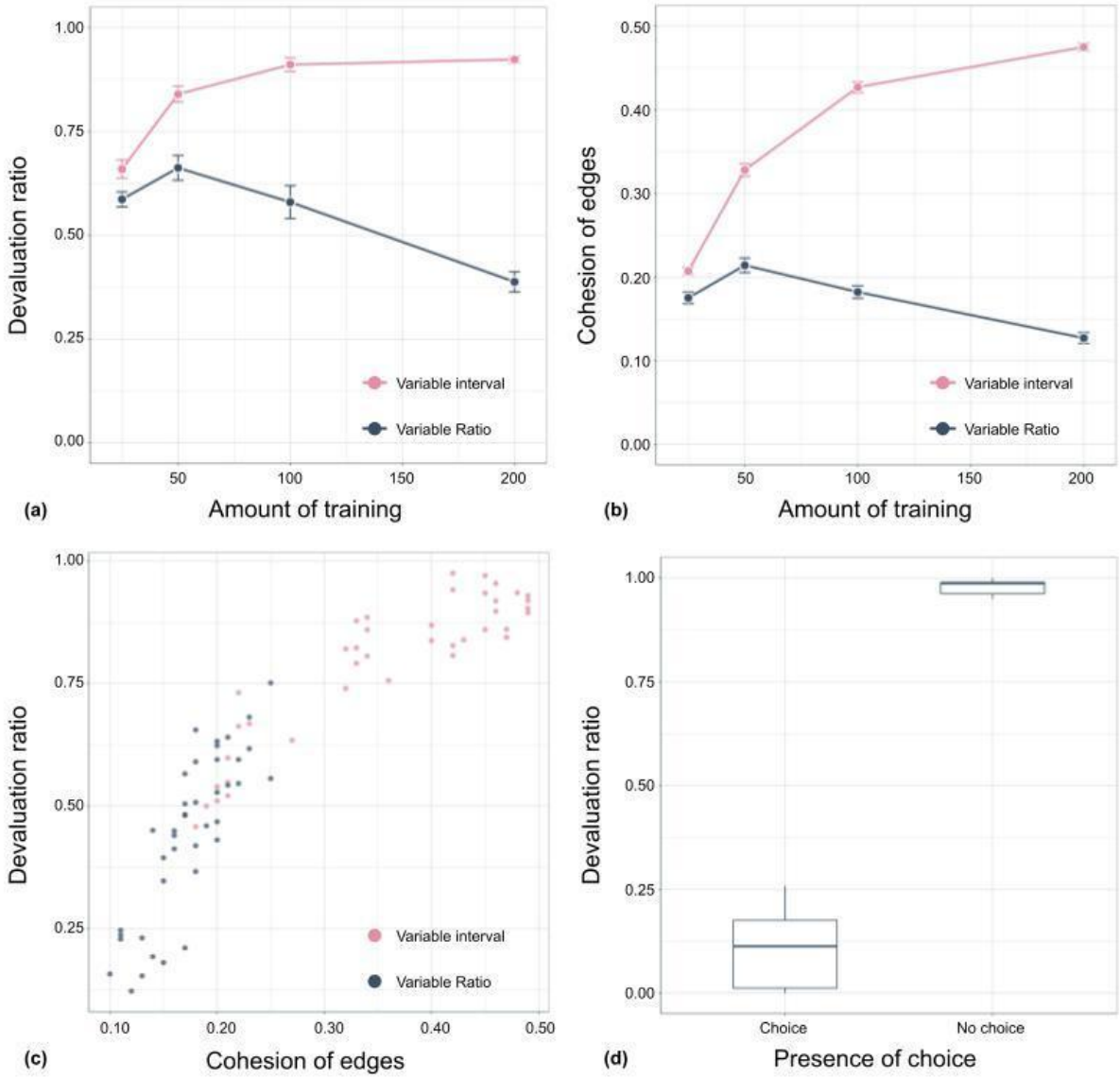
Supplementary figure 2. Simulation results replicating Figure 2 (simulation 1), with the different numbers of nodes (25–100).



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.