RESEARCH ARTICLE

Influence of multiple action–outcome associations on the transition dynamics toward an optimal choice in rats

Noha Mohsen Zommara¹ · Muneyoshi Takahashi² · Johan Lauwereyns^{1,2}

Received: 10 May 2017 / Revised: 5 October 2017 / Accepted: 11 October 2017 / Published online: 16 October 2017
© Springer Science+Business Media B.V. 2017

Abstract When faced with familiar versus novel options, animals may exploit the acquired action–outcome associations or attempt to form new associations. Little is known about which factors determine the strategy of choice behavior in partially comprehended environments. Here we examine the influence of multiple action–outcome associations on choice behavior in the context of rewarding outcomes (food) and aversive outcomes (electric foot-shock). We used a nose-poke paradigm with rats, incorporating a dilemma between a familiar option and a novel, higher-value option. In Experiment 1, two groups of rats were trained with different outcome schedules: either a single action–outcome association (“Reward-Only”) or dual action–outcome associations (“Reward-Shock”; with the added opportunity to avoid an electric foot-shock). In Experiment 2, we employed the same paradigm with two groups of rats performing the task under dual action–outcome associations, with different levels of threat (a low- or high-amplitude electric foot-shock). The choice behavior was clearly influenced by the action–outcome associations, with more efficient transition dynamics to the optimal choice with dual rather than single action–outcome associations. The level of threat did not affect the transition dynamics. Taken together, the data suggested that the strategy of choice behavior was modulated by the information complexity of the environment.

Keywords Multiple action–outcome associations · Transition dynamics · Nose-poke paradigm · Rats

Introduction

Successful behavior requires animals as well as humans to select and process relevant stimuli, associated with reward or with the impending arrival of an aversive event, in order to prepare the appropriate response. We generally face novel stimuli in partially comprehended environments, including well-learned statistical regularities between actions and outcomes. Thus, we can choose to explore the new or to exploit the familiar. Humans, monkeys, rats, and other animals that show adaptive behavior stand to benefit from a tendency to engage in exploration under certain conditions, even when a particular context offers tried-and-true relations between stimuli, actions, and positively or negatively valued outcomes. Berlyne (1966) referred to this tendency to engage in exploration as “curiosity,” and suggested that the conventional approaches of Pavlovian and operant conditioning were unable to address the phenomenon. In the last decade, the study of decision-making with the computational framework of reinforcement learning has answered to the call to incorporate exploration as a critical component in learning by proposing that decision-making can be understood as an internal competition between different learning systems: a so-called “model-free system,” which incorporates the Pavlovian and operant mechanisms, and a “model-based system,” which involves a more cognitively-driven, information-oriented type of learning (Daw et al. 2005; Dayan and Daw 2008). Previous studies have shown the important role of the environmental information in decision-making and how

✉ Noha Mohsen Zommara
nohazommara@gmail.com

¹ Graduate School of Systems Life Sciences, Kyushu University, 3-1-1, Maidashi, Higashi-ku, Fukuoka 812-8582, Japan

² Brain Science Institute, Tamagawa University, Machida, Japan

spatial information in the brain being enhanced through learning (Yan et al. 2016; Hayakawa et al. 2015).

However, little is known about when, or in which contexts, an animal prefers switching toward a novel stimulus to staying with a familiar stimulus. Here, we develop a behavioral paradigm that allows us to systematically investigate the transition dynamics toward an optimal choice in rats. Based on the concept of model-based versus model-free reinforcement learning we hypothesized that the likelihood of exploration would be modulated by information-related factors in the context such as the expected value of the familiar choice option and the numbers of action–outcome associations. Such factors might increase or decrease the reinforcing level of behavior and thus affect the decision to explore and seek for the novel choice option. In both experiments, we used multiple action–outcome associations in certain conditions to address this hypothesis, by presenting rewarding outcomes (food pellets) in addition to manipulating the threat level, with a potential aversive outcome (an electric foot-shock) for erroneous actions.

Experiment 1: One versus two action–outcome associations

In Experiment 1, we employed both rewards and aversive events in order to contrast a hypothesis on the basis of “motivational significance” versus one on the basis of “information value.” Consider a situation in which a novel stimulus is presented concurrently with a familiar stimulus of a particular utility (i.e., associated with a particular action outcome). According to operant conditioning the utility of the familiar stimulus affects the reinforcement level of the behavior by increasing the chances of the familiar stimulus being selected in future situations. However, in a context where novel stimulus is offered with a new unknown utility a competition between previously acquired behavior associated with the familiar stimulus and the newly impeded novel stimulus will occur. Such stimuli competition depends on the degree of reinforcement associated with the previously learned association with familiar stimulus. If the behavior is strongly reinforced, the transition to the new stimuli might be less likely to occur. Thus, the transition dynamics to a novel choice option with higher utility might depend on the value of the familiar choice option; we call this the “motivational significance” hypothesis (compatible with traditional models of operant conditioning; for review, see Dickinson and Balleine 1994).

Now consider the same situation in terms of information processing. The value of the new stimulus (as a potential predictor of events, or a possible source to reduce the

ambiguity of the situation) should be higher when the familiar stimulus implies multiple associations than when the familiar stimulus provides a clear and simple implication. In the case of multiple associations, the additional information carried by the novel stimulus would allow the animal to learn more about the partially comprehended environment and thus to optimize its behavioral choices. On the other hand, in a well-understood situation, with a familiar stimulus that provides a clear and simple implication, the animal may be more likely to rely on habitual or “model-free” processing, exploiting a well-learned behavior regardless of the new stimulus. Several researchers have suggested that in variable environments animals tend to increase exploratory behavior and seek to promote their acquisition of information (e.g., Berlyne 1966; Catania 1975; Neuringer 1986, 2002; Page and Neuringer 1985; Roche et al. 1997). By this view, familiar stimuli with variable associations should produce more exploratory behavior than familiar stimuli with simple implications; we call this the “information value” hypothesis.

To pitch the “motivational significance” hypothesis against the “information value” hypothesis, we designed a nose-poke paradigm for rats with food rewards and electric foot-shocks as potential action outcomes. Initially the rat was trained to poke its nose into an illuminated hole at the center of the front wall in a Skinner box. The illumination of the central hole, then, was the familiar stimulus. We prepared two groups: One group was trained to obtain a reward only (i.e., “Reward-Only”), whereas the other group was trained to obtain a reward and simultaneously to avoid a foot-shock (i.e., “Reward-Shock”).

For the “Reward-Only” group, the association with the familiar stimulus was only of reinforcement “being able to obtain food pellets” and not of both reinforcement and punishment. That is, the onset of illumination (moment X) at the central hole (position Y) converged on the potential action outcome of food pellets: Two things lead to one. For the “Reward-Shock” group, on the other hand, the familiar stimulus held the double association of being able to obtain food pellets (event A) and avoid a foot-shock (event B). That is, moment X and position Y were associated with two events—a more critical situation, with plenty of unresolved associations in the sense that either moment X or position Y, or both, could be crucial for either the food pellets or the foot-shock avoidance, or both.

Following the initial training with reward-only or reward-and-avoidance learning, we introduced an alternative peripheral nose-poke hole in the same experimental apparatus and extended the task to a free-choice task. The rats could now choose either the original central nose-poke hole, and obtain the same reinforcement as before, or the novel peripheral hole, which was associated with a

larger reward magnitude (and remained unchanged with respect to foot-shock avoidance for the “Reward-Shock” group). The “motivational significance” hypothesis predicted that the “Reward-Only” group should show faster transition dynamics to the novel option because for these rats the familiar stimulus had less motivational significance than for the “Reward-Shock” group (who earned the same amount of reward and, additionally, avoided a foot-shock).

The “information value” hypothesis made the opposite prediction. Since the task situation was more complex for the “Reward-Shock” group, these rats should be more likely to explore the information carried by the novel stimulus (and thus to achieve faster transition dynamics) than the “Reward-Only” group.

Methods

Subjects and apparatus

Twenty-four experimentally naïve male Sprague–Dawley rats (weighing 250–270 g; 8 weeks old on arrival; Japan SLC Inc., Hamamatsu, Japan) served as subjects. Each rat was assigned to one of two groups (each with $n = 12$). They were housed individually and kept on a reversed 12 h light/dark cycle (lights on at 2000 h) with experimental sessions occurring during the dark period. Water was available *ad libitum* in the home cages. Subjects were fed individually after completion of each testing session to preserve their 85% free-feeding body weight throughout the duration of the experiment with allowances for growth. Four identical chambers of nine nose-poke holes (ENV-NPW-9L; Med Associates Inc., St Albans, VT), were used to conduct the experimental procedures. Each chamber was fitted within a sound-attenuating cubicle equipped with an exhaust fan, serving to ventilate the chamber and generate a background noise level of 65 dB (scale A). A 2.8-W house light was mounted on the ceiling of the cubicle, illuminated throughout the experimental sessions. The front and rear walls of each chamber were constructed of metal. The left and right walls and the ceiling were constructed of transparent Plexiglas. The left wall also functioned as the entrance to the chamber. All boxes contained an arc of 9 contiguous apertures set into the curved front wall. Each aperture was 2.5 cm square, 2.2 cm deep and 2.0 cm above the floor. Light-emitting diodes (LEDs) at the rear of each hole could be turned on and off automatically to provide visual cues specific to each hole. Infrared photo-beam detectors at the front of each nose-poke hole were used for the recording of the response timing and location. Each chamber was equipped with a recessed food pellet delivery trough fitted

with infrared photo-beam detectors to detect head entries, and with a 2.8-W lamp to illuminate the food trough. The trough, into which 20 mg food pellets (Bio-Serv, Frenchtown, NJ) were delivered, was located 2 cm above the floor in the center of the rear wall. The light in the trough was illuminated when the food pellets were delivered after correct responses and was extinguished when the food pellets were collected.

Both groups were able to obtain reward for every correct response. However, on erroneous trials an electric foot-shock was applied for only one group (i.e. Reward-Shock). The floor of the chamber was constructed of stainless-steel rods measuring 0.5 cm in diameter, spaced 1.5 cm apart center-to-center, and connected to a shock generator (Shocker/Scrambler: ENV-414, Med Associates Inc.) that delivered scrambled foot-shocks. All procedures, for the present and the subsequent experiment in the current study, were approved by the Kyushu University and Tamagawa University committees for animal care and use, and were in accordance with NIH guidelines.

Procedure

One group of rats was trained with reward only (the “Reward-only” group), whereas the other group was trained with both reward and avoidance of a shock (the “Reward-Shock” group; see Fig. 1 for a schematic representation of the design). In the initial training phase and baseline test, a trial started when the central nose-poke hole was illuminated. This light signaled that the rat was required to make a nose-poke response immediately in order to receive a food reward (2 pellets). If the rat did not make a nose-poke response within 5 s, the central light of the nose-poke hole was extinguished. For the “Reward-Shock” group, a 0.5 s foot-shock (0.18 mA) was delivered simultaneously with the offset of the light (i.e., an omission error); similarly, a foot-shock was delivered when the rat made a nose-poke response in the absence of visual stimulation (i.e., a commission error). For the “Reward-Only” group, no foot-shocks were delivered at any stage in the task. Nose-poke responses at the central hole during illumination were immediately rewarded with 2 food pellets (i.e., a correct response) and the central light was turned off. After an inter-trial interval (ITI) ranging from 13 to 18 s, the central light was re-illuminated to give the rat a new opportunity to proceed with a trial. For erroneous responses, the ITI ranged from 18 to 23 s (i.e., 5 s longer than the ITI after correct trials). An individual training session lasted for 45 min. The training was continued until the rats were able to reach more than 80% correct performance for at least 3 consecutive experimental sessions. The correct performance rate (CPR) was computed as follows:

Fig. 1 Experimental design of Experiment 1. The top panels illustrate the experimental procedures during training and baseline (left) and the transition test (right). In the transition test, the full line indicates the familiar choice option; the dashed line indicates the optimal, novel choice option. The table lists the reinforcement schedule for center and peripheral choice options for each group

$$\text{CPR} = \frac{N(\text{correct})}{N(\text{correct}) + N(\text{commission}) + N(\text{omission})}$$

After obtaining stable responses for the baseline test, all rats were presented with the transition test. For this test, a peripheral nose-poke hole was opened, and made available for responding, in addition to the central hole. We removed the metal shield from the second-most peripheral hole, and counterbalanced between the left and right side within each group (i.e., six rats in each group were tested with a left peripheral hole, whereas the other six were tested with a right peripheral hole). When a trial started, both nose-poke holes were illuminated for 5 s simultaneously. If the rats made a correct response to the central hole, they earned 2 food-pellets (i.e., a small reward), similar to the reinforcement schedule during the baseline training. However, if the rats chose the peripheral hole, they could collect three times the amount of reward (i.e., 6 food pellets) immediately after the response. An individual experimental session lasted for 45 min or until the rat had made 120 correct nose-poke responses; each rat participated in one session per day.

For the “Reward-Shock” group, in terms of avoidance behavior, both the central and the peripheral hole carried the same implications: When illuminated, both holes provided an opportunity for the rats to avoid a foot-shock. Conversely, the rats received a foot-shock if they responded to either the central hole or the peripheral hole before the LED was illuminated (i.e., commission error), and they also received foot-shocks for omission errors (i.e., when the rats failed to respond to either the central or the peripheral hole within the allotted time).

Results

After training, all rats from both groups were able to perform the nose-poke responses to the central hole when the LED was illuminated, at a correct performance rate (CPR) of more than 80% during 45-min sessions across last 3 successive days of training (see Fig. 2). Mean correct performance rate of “Reward-Shock” group is ($M = 0.92$, $SD = 0.037$) and “Reward-Only” group is ($M = 0.88$, $SD = 0.079$). Two-way repeated measures ANOVA between-subjects effect on mean correct performance rate

Fig. 2 Mean correct performance rates during the baseline test in Experiment 1. The graph illustrates the group averages (with error bars reflecting the standard error of the mean) for each daily session during the last 3 days of the training for the “Reward-Only” and “Reward-Shock” groups

showed that there was no significant group differences $p > 0.1$. To assess the effect of reinforcement context on the transition behavior, we measured the optimal choice rate, by dividing the number of correct peripheral choices by the total number of correct choices. This index ranged from 0, if the rats always chose the central hole, to 1, if their preference had completely shifted to the peripheral option.

The optimal choice rate in the transition test (OCR) is illustrated for both experimental groups across 5 successive days in Fig. 3. The OCR gradually increased over the course of 5 days, for both groups. Individual data are shown for day 5 in Table 1, and over the course of 5 days in Fig. 4. Only one subject in the “Reward-Shock” group failed to switch to the optimal, novel option. Rats trained with dual action–outcome associations (Fig. 4a) were able to achieve higher peripheral responding rates and faster adaptation than rats trained with a single association (Fig. 4b).

Two-way repeated measures ANOVA on OCR, with the between-subjects factor Context (“Reward-Only” versus “Reward-Shock”) and the within-subjects factor Day (5 days), revealed a significant main effect of Day, $F(4,88) = 65.16$, $p < 0.001$, $\eta^2 = 0.784$ but not of Context ($F < 2$). However, there was a significant interaction

Fig. 3 Mean optimal choice rates as a function of Day and Context in Experiment 1. There was a significant difference between contexts on the first and second days of the transition test. Error bars reflecting the standard error of the mean

Table 1 Individual data (OCR) on Day 5 of the transition test in Experiment 1

Rat	1	2	3	4	5	6	7	8	9	10	11	12
Reward-Shock												
OCR	0.93	0.90	1	0.99	0.99	0	1	1	0.99	0.98	0.01	0.80
Reward-Only												
OCR	0.69	0.80	1	0.93	1	0.93	0.98	0.98	0.78	0.94	0.81	0.97

between Day and Context, $F(4,88) = 7.009$, $p < 0.001$, $\eta^2 = 0.242$. Post-hoc pair-wise comparisons (with Bonferroni correction) revealed that there were statistically significant differences between Contexts on the first day ($p = 0.016$) and on the second day ($p = 0.020$). The group trained with two associations, “Reward-Shock,” made the transition to the novel option significantly faster than the group trained with one association, “Reward-Only.”

Two-way repeated measures ANOVA on commission error rates revealed a significant interaction between Day and Context, $F(4,88) = 2.563$, $p < 0.05$, $\eta^2 = 0.104$ as well as a significant main effect of Context, $F(1,22) = 22.556$, $p < 0.001$, $\eta^2 = 0.506$ (the commission error rates are shown in Fig. 5).

Discussion

As predicted by the “information value” hypothesis, but not the “motivational significance” hypothesis, the rats in the “Reward-Shock” group showed faster transition dynamics to the optimal choice, particularly during the first two days of the transition test. Miller and Escobar (2002) reported that when there are two implications with one stimulus, the context becomes more critical. Animals tend to engage in active exploration to disambiguate the environment by searching for any new informative stimuli (Blaisdell 2008). Presently, the variability in the context appeared to promote attention to the newly introduced option. The acquisition of new information might be intrinsically rewarding in that the information value is higher in a complex context (Behrens et al. 2007). However, before accepting the “information value” hypothesis as a more valid approach to the present data than the “motivational significance” hypothesis, one might propose that other factors modulated the operant conditioning: particularly, Pavlovian conditioning of place preference (for review, see Huston et al. 2013), and/or arousal (e.g., Miller 1948).

In terms of a modulation by Pavlovian conditioning, we ought to consider that the central nose poke might acquire a negative value for the rats due to the association with foot-shocks, despite the fact that the rats can perform an operant response to avoid the foot-shocks. This negative association might then contribute to the attractiveness or

Fig. 4 Individual data of optimal choice rates during 5 days of the transition test in Experiment 1. The left panel shows the individual data for the “Reward-Shock” group. The right panel shows the individual data for the “Reward-Only” group

Fig. 5 Mean commission error rates in Experiment 1. Vertical axis shows commission error rates as a function of 5 days of transition test. Error bars reflecting the standard error of the mean

“salience” of the new alternative, and so facilitate the instrumental learning. Alternatively, the threat of foot-shocks might have produced a heightened level of arousal, which would then facilitate any instrumental learning. Interestingly, both of these alternative explanations imply that foot-shocks of larger amplitude should further strengthen the tendency observed in Experiment 1. However, the “information value” hypothesis does not align with that prediction. We conducted Experiment 2 to address the different predictions.

Experiment 2: Varying the level of threat

Using exactly the same paradigm as in Experiment 1, we contrasted the predictions from the “information value” hypothesis versus proposals that combine operant conditioning with a modulation by Pavlovian conditioning and/

or arousal. We did so by comparing the transition dynamics between two groups of rats, varying only the level of amplitude of the foot-shocks, which were applied only in error trials. According to hybrid proposals that combine operant conditioning with Pavlovian conditioning and/or arousal, increasing the amplitude of the foot-shock, and thus the level of threat, should strengthen the modulation and therefore exacerbate the tendency seen in Experiment 1. Rats in a high-threat condition should show faster transition dynamics to the optimal choice than rats in a low-threat condition. Conversely, the “information value” hypothesis makes no such prediction. Both groups are presented with the same number of action–outcome associations but different levels of stimulus threat. There should be no difference in information value for the novel choice option.

However, it may be argued that the information value in a variable context reflects a higher-order form of operant conditioning, in which information has an intrinsic reinforcing potential. Variability can reinforce behavior and may be considered as an operant dimension (De Souza Barba 2012). One might reason that different operant mechanisms could become integrated or summated. Specifically, in a high-threat condition, the basic operant conditioning mechanisms should lead to a more strongly reinforced central nose-poke response than in a low-threat condition. New information might therefore be less likely to attract rats with strongly reinforced nose-poke response behavior. Thus, the higher-order operant mechanism, relating to information, might be less likely to control the rats’ behavior than the lower-order operant mechanism, relating to the threat of foot-shock.

The “information value” hypothesis, then, would predict either no difference as a function of the level of threat (because the number of action–outcome associations stays the same), or a slower shift to the optimal choice in a high-threat condition than in a low-threat conditioning (because

the central nose-poke response would be more strongly reinforced).

Methods

Subjects and apparatus

Twenty-four experimentally naïve male Sprague–Dawley rats served as subjects. Each rat was assigned to one of two groups (each with $n = 12$). The experiments were conducted using the same four nose-poke chambers as in Experiment 1.

Procedure

One group of rats was trained, and performed the transition test, with both reward and avoidance of shock with intensity of 0.18 mA for 0.5 s (“Reward-Shock”—0.18 mA); the other group of rats was trained, and performed the transition test, with both reward and avoidance of shock with intensity of 0.36 mA for 0.5 s (“Reward-Shock”—0.36 mA) (see Fig. 6 for a schematic representation of the design). In all other respects, the procedures were exactly the same as in Experiment 1.

During training, we initially applied foot-shocks of 0.54 mA in error trials for only one group. With this level of shock, most rats failed to achieve a high performance rate on the nose-poke task, tending to avoid responding to the illuminated central hole. The rats were then retrained without any foot-shock, before introducing foot-shocks with amplitude of 0.36 mA in error trials. With the threat of 0.36 mA foot-shocks, the rats were able to reliably acquire the central nose-poke response. For the other

group, the level of the foot-shock was 0.18 mA throughout the entire training.

In Experiment 2, we initially aimed to perform the transition test over the course of 5 days, as in Experiment 1. However, upon noticing the ongoing dynamics in the rats’ choice behavior, we decided to extend the transition test in Experiment 2 for another 10 days (for a total of 15 days).

Results

All rats were able to show stable responses during training. Both groups performed nose-poke responses to the central hole when the LED was illuminated, at a correct performance rate of more than 80% during individual 45-min sessions across 3 successive days (see Fig. 7). Mean correct performance rate for “Reward-Shock” 0.36 mA and “Reward-Shock” 0.18 mA are ($M = 0.84$, $SD = 0.13$) and ($M = 0.87$, $SD = 0.07$) consequently. There were no statistically significant differences between groups in the correct rates.

The “Reward-Shock” 0.36 mA group did not significantly differ from the “Reward-Shock” 0.18 mA group in terms of optimal choice rates during the 5 days of the transition test (Fig. 8). Two-way repeated measures ANOVA revealed that there was a statistically significant effect of Day on the OCR, $F(4, 88) = 12.149$, $p < 0.001$, $\eta^2 = 0.356$. However, we did not obtain a main effect of Context ($F < 1$) or any interaction between Context and Day ($F < 2$).

The different levels of shock appeared to produce considerable individual variation (see Table 2 and Fig. 9). Five out of 12 rats trained with the lower shock amplitude

Fig. 6 Experimental design of Experiment 2. Same format as Fig. 1

Fig. 7 Mean correct performance rates during the baseline test in Experiment 2. There was no significant effect between contexts during the baseline. Error bars reflecting the standard error of the mean

Fig. 8 Mean optimal choice rates as a function of Day and Context in Experiment 2. Vertical axis shows the optimal choice rates as a function of 5 days of transition test. Error bars reflecting the standard error of the mean

of 0.18 mA still ignored the new option by Day 5. Among rats trained with the higher shock amplitude of 0.36 mA, eight out of 12 rats still ignored the new option by Day 5.

Two-way repeated measures ANOVA on commission rate showed a main effect of Day, $F(4,88) = 11.808$, $p < 0.001$, $\eta^2 = 0.349$ and a main effect of Context,

$F(1,22) = 22.556$, $p < 0.001$, $\eta^2 = 0.246$ but there was no interaction between Day and Context, $F < 2$ (see Fig. 10).

Given the null results in the OCR, further analyses were conducted to examine whether the two groups showed behavioral differences in reaction time (see Fig. 11). We obtained a main effect of Context, $F(1,24) = 5.721$, $p < 0.05$, $\eta^2 = 0.193$ but no main effect of Day ($F < 2$) and no interaction between Context and Day ($F < 2$). The “Reward-Shock” 0.36 mA group responded to an illuminated hole faster (at an average reaction time of 911 ms) than the “Reward-Shock” 0.18 mA group (an average reaction time of 1157 ms).

Since the behavioral trends did not appear to have stabilized after 5 days, we extended the transition test in Experiment 2, and performed a second set of analyses with data from 15 days (see Figs. 12, 13, 14). The pattern of data, and results from ANOVA analyses, showed the same trends as with the data set from 5 days.

Discussion

The finding that there was no significant difference in the transition dynamics as a function of the level of threat did not match with the predictions based on proposals that combine operant conditioning with a modulation by Pavlovian conditioning and/or arousal. The null result, however, went accompanied by changes as a function of the level of threat in the pattern of errors and in the reaction times. Rats in the “Reward-Shock” 0.36 mA group made faster responses and more commission errors than Rats in the “Reward-Shock” 0.18 mA group, suggesting that the increase in the level of threat caused the rats to be more active, possibly more alert, or in a higher state of arousal. In this sense it is important to note that the increased activity or alertness did not imply an improvement in learning, or the ability to optimize choices. The rats showed evidence of being influenced by the amplitude of the foot-shock, yet this did not stimulate a switch to the new option; instead, if anything, it appeared to make the rats more reluctant to discontinue the old choice option.

In Experiment 2, as in Experiment 1, we observed a gradual optimization over the course of multiple days, with highly significant effects of Day on OCR. In this sense, the

Table 2 Individual data (OCR) on Day 5 of the transition test in Experiment 2

Rat	1	2	3	4	5	6	7	8	9	10	11	12
Reward-Shock (0.18 mA)												
OCR	0.79	0.90	0.40	0.90	0.98	0.00	0.00	0.94	0.00	0.00	0.95	0.00
Reward-Shock (0.36 mA)												
OCR	1	0	0.01	0	0.01	0.99	0.01	0	0.98	0.85	0.02	0

Fig. 9 Individual data of optimal choice rates during 5 days of the transition test in Experiment 2. The left panel shows the individual data for the “Reward-Shock” 0.18 mA group. The right panel shows the individual data for the “Reward-Shock” 0.36 mA group

Fig. 10 Mean commission error rates during 5 days of the transition test in Experiment 2. Error bars reflecting the standard error of the mean

Fig. 12 Mean optimal choice rates as a function of Day and Context during 15 days of the transition test in Experiment 2

Fig. 11 Mean reaction times during 5 days of the transition test in Experiment 2. The graph shows significant effect of context in reaction time during the transition test. Error bars reflecting the standard error of the mean

data replicated very well. The inter-individual variation was comparable in both experiments, as can be gleaned from the error bars in the graphs, and the statistics as reported. To the extent that the data showed slight variation

Fig. 13 Mean commission error rates during 15 days of the transition test in Experiment 2

across the two experiments, it should be noted that they were conducted at different institutions, with rats supplied by different animal breeders. Factors relating to rearing and acclimatization might have affected the behavior. Critically, the group comparison in Experiment 2 was done between two groups of rats supplied by the same animal

Fig. 14 Mean reaction times during 15 days of the transition test in Experiment 2

breeder and subjected to the same acclimatization; thus, these two groups in Experiment 2 can effectively be compared directly, with all factors the same, except for the experimental manipulations. Such direct numerical comparison is more problematic across experiments due to additional factors that are not directly related to the experimental conditions (e.g., rearing, housing, acclimatization).

General discussion

In two behavioral experiments with rats performing in a nose-poke paradigm, we investigated the transition dynamics toward an optimal choice in a situation that pitches a familiar choice option with known reinforcement schedule against a novel choice option with unknown, but superior reinforcement schedule. In Experiment 1 and 2 we varied the level of threat, and found that with multiple action–outcome associations the rats showed enhanced transition dynamics. However, the transition dynamics did not depend on the magnitude of the electrical foot-shock. The pattern of data suggested that the number of action–outcome associations was the factor that influenced the transition dynamics, consistent with the “information hypothesis.” It has been suggested that the more uncertain and variable the situation, the higher will be the value of new information (Dayan et al. 2000; Courville et al. 2006). Animals tend to explore and resolve the ambiguity of the context by collecting information (see also Molet et al. 2010). In both experiments presentation of shock suppressed the commission errors for the group trained with 0.18 mA but increased the error rates for the group trained with 0.36 mA, consistent with the notion that the threat of an electric shock is more salient for the group trained with 0.36 mA. The data suggests increased magnitude of shock affects the performance during choices. This observation

lends further support to the information hypothesis (here, in line with operant conditioning) rather than classical conditioning.

The data are best understood with the theoretical framework based on “model-based” versus “model-free” reinforcement learning. While this framework clearly builds on the work from earlier generations, it can be considered an upgrade in the sense that it incorporates a clear role for exploration and information-seeking behavior. Timberlake (2004) has suggested that operant contingencies can be improved by linking it to other concepts and approaches. It has been recognized that incentives do not operate by converting into a single currency that drives behavior, but that the nature of the incentives activates specific behavioral systems (see Domjan 2005). More generally, the present data may be considered compatible with an explicit role for cognitive factors in the learning (e.g., Foote and Crystal 2007; Kepecs et al. 2008). By this view, the transition dynamics reflect the role of exploratory behavior as a function of cognitive factors such as the internal representation of utility and the complexity of information present in the given context. Animals often prefer variable to fixed reinforcement ratios, even when the fixed ratio provides a better payoff (e.g., Field et al. 1996), again connecting exploratory behavior with variability. From a cognitive perspective, such tendencies suggest that the exploratory behavior may be driven by a need for information as a potential source of disambiguation.

By this interpretation, information would serve as an intrinsic attractor, particularly in partially understood settings. The notion that information may be desirable in and of itself was illustrated in a study by Bromberg-Martin and Hikosaka (2009), in which monkeys chose to obtain cues that provided reliable advance information about future events rather than cues that provided random information, even though the information changed nothing with respect to the actual amounts of reward that could be obtained in the experiment. Apparently, the monkeys valued the advance information. Similarly, in our current experiments, rats that were faced with a complex setting may have been particularly responsive to the attraction of new information that might shed light on the relationship between different stimulus features and outcomes. Future studies can build on the current paradigm to systematically investigate the role of information with context variability, and the role of different types of reinforcers, in the dynamics of transition.

Acknowledgements This work was supported by Human Frontier Science Program award RGP0039/2010, Grant-in-Aid for Scientific Research on Innovative Areas “Neural creativity for communication” (24120710), “Comprehensive Brain Science Network,” Tamagawa Global Center of Excellence Program of the Ministry of Education, Culture, Sports, Science, and Technology, Japan, and the Narishige Neuroscience Research Foundation. We thank Minoru Tsukada, Gary

D. Bird, and David N. Harper for advice and comments on the research.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All applicable international, national, and/or institutional guidelines for the care and use of animals were followed. All procedures were approved by the Kyushu University committee for Animal Care and Use, and were in accordance with NIH guidelines.

References

- Behrens TE, Woolrich MW, Walton ME, Rushworth MF (2007) Learning the value of information in an uncertain world. *Nat Neurosci* 10:1214–1221. doi:[10.1038/nn1954](https://doi.org/10.1038/nn1954)
- Berlyne DE (1966) Curiosity and exploration. *Science* 153:25–33
- Blaisdell AP (2008) Cognitive dimension of operant learning. In: Menzel R (ed) *Learning and memory: a comprehensive reference*. Elsevier, Oxford, pp 173–195
- Bromberg-Martin ES, Hikosaka O (2009) Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron* 63:119–126
- Catania AC (1975) Freedom and knowledge: an experimental analysis of preference in pigeons. *J Exp Anal Behav* 24:89–106. doi:[10.1016/j.neuron.2009.06.009](https://doi.org/10.1016/j.neuron.2009.06.009)
- Courville AC, Daw ND, Touretzky DS (2006) Bayesian theories of conditioning in a changing world. *Trends Cogn Sci* 10:294–300. doi:[10.1016/j.tics.2006.05.004](https://doi.org/10.1016/j.tics.2006.05.004)
- Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci* 8:1704–1711. doi:[10.1038/nn1560](https://doi.org/10.1038/nn1560)
- Dayan P, Daw ND (2008) Decision theory, reinforcement learning, and the brain. *Cogn Affect Behav Neurosci* 8:429–453. doi:[10.3758/CABN.8.4.429](https://doi.org/10.3758/CABN.8.4.429)
- Dayan P, Kakade S, Montague PR (2000) Learning and selective attention. *Nat Neurosci* 3:1218–1223. doi:[10.1038/81504](https://doi.org/10.1038/81504)
- De Souza Barba L (2012) Operant variability: a conceptual analysis. *Behav Anal* 35:213–227
- Dickinson A, Balleine B (1994) Motivational control of goal-directed action. *Anim Learn Behav* 22:1–18. doi:[10.3758/BF03199951](https://doi.org/10.3758/BF03199951)
- Domjan M (2005) Pavlovian conditioning: a functional perspective. *Annu Rev Psychol* 56:179–206. doi:[10.1146/annurev.psych.55.090902.141409](https://doi.org/10.1146/annurev.psych.55.090902.141409)
- Field DP, Tonneau F, Ahearn W, Himeline PN (1996) Preference between variable-ratio and fixed-ratio schedules: local and extended relations. *J Exp Anal Behav* 66:283–295. doi:[10.1901/jeab.1996.66-283](https://doi.org/10.1901/jeab.1996.66-283)
- Foot AL, Crystal JD (2007) Metacognition in the rat. *Curr Biol* 17:551–555. doi:[10.1016/j.cub.2007.01.061](https://doi.org/10.1016/j.cub.2007.01.061)
- Hayakawa H, Samura T, Kamijo TC, Sakai Y, Aihara T (2015) Spatial information enhanced by non-spatial information in hippocampal granule cells. *Cogn Neurodyn* 9:1–12. doi:[10.1007/s11571-014-9309-x](https://doi.org/10.1007/s11571-014-9309-x)
- Huston JP, de Souza Silva MA, Topic B, Müller CP (2013) What's conditioned in conditioned place preference? *Trends Pharmacol Sci* 34:162–166. doi:[10.1016/j.tips.2013.01.004](https://doi.org/10.1016/j.tips.2013.01.004)
- Kepecs A, Uchida N, Zariwala HA, Mainen ZF (2008) Neural correlates, computation and behavioural impact of decision confidence. *Nature* 455:227–231. doi:[10.1038/nature07200](https://doi.org/10.1038/nature07200)
- Miller NE (1948) Studies of fear as an acquirable drive: I. Fear as motivation and fear-reduction as reinforcement in the learning of new responses. *J Exp Psychol* 38:89–101. doi:[10.1037/h0058455](https://doi.org/10.1037/h0058455)
- Miller RR, Escobar M (2002) Associative interference between cues and between outcomes presented together and presented apart: an integration. *Behav Process* 57:163–185. doi:[10.1016/S0376-6357\(02\)00012-8](https://doi.org/10.1016/S0376-6357(02)00012-8)
- Molet M, Urcelay GP, Miguez G, Miller RR (2010) Using context to resolve temporal ambiguity. *J Exp Psychol Anim Behav Process* 36:126–136. doi:[10.1037/a0016055](https://doi.org/10.1037/a0016055)
- Neuringer A (1986) Can people behave “randomly”? the role of feedback. *J Exp Psychol Gen* 115:62–75. doi:[10.1037/0096-3445.115.1.62](https://doi.org/10.1037/0096-3445.115.1.62)
- Neuringer A (2002) Operant variability: evidence, functions, and theory. *Psychon Bull Rev* 9:672–705. doi:[10.3758/BF03196324](https://doi.org/10.3758/BF03196324)
- Page S, Neuringer A (1985) Variability is an operant. *J Exp Psychol Anim Behav Process* 11:429–452. doi:[10.1037/0097-7403.11.3.429](https://doi.org/10.1037/0097-7403.11.3.429)
- Roche JP, Timberlake W, McCloud C (1997) Sensitivity to variability in food amount: risk aversion is seen in discrete-choice, but not in free-choice, trials. *Behaviour* 134:1259–1272. doi:[10.1163/156853997X00142](https://doi.org/10.1163/156853997X00142)
- Timberlake W (2004) Is the operant contingency enough for a science of purposive behavior? *Behav Philos* 32:197–229
- Yan C, Wang R, Qu J, Chen G (2016) Locating and navigation mechanism based on place-cell and grid-cell models. *Cogn Neurodyn* 10:353–360. doi:[10.1007/s11571-016-9384-2](https://doi.org/10.1007/s11571-016-9384-2)