Dynamics of pre- and post-choice behaviour: rats approximate optimal strategy in a discrete-trial decision task

Justine Fam¹, Fred Westbrook¹ and Ehsan Arabzadeh²,³

¹School of Psychology, University of New South Wales, Sydney, New South Wales, Australia
²Eccles Institute of Neuroscience, John Curtin School of Medical Research, Australian National University, Canberra, Australian Capital Territory, Australia
³ARC Centre of Excellence for Integrative Brain Function, Australian National University, Canberra, Australian Capital Territory, Australia

We simulate two types of environments to investigate how closely rats approximate optimal foraging. Rats initiated a trial where they chose between two spouts for sucrose, which was delivered at distinct probabilities. The discrete trial procedure used allowed us to observe the relationship between choice proportions, response latencies and obtained rewards. Our results show that rats approximate the optimal strategy across a range of environments that differ in the average probability of reward as well as the dynamics of the depletion-renewal cycle. We found that the constituent components of a single choice differentially reflect environmental contingencies. Post-choice behaviour, measured as the duration of time rats spent licking at the spouts on unrewarded trials, was the most sensitive index of environmental variables, adjusting most rapidly to changes in the environment. These findings have implications for the role of confidence in choice outcomes for guiding future choices.

1. Introduction

There are multiple aspects to a behavioural choice that cannot be captured by the choice action itself; this paper examines the various stages of choice behaviour as it unfolds in time. We trained rats in a discrete trial task that allowed firstly, a single choice to be parsed into various component stages, and secondly, for these stages to be mapped onto specific behaviours of the rat. We demonstrate that the complex environmental inputs that guide a choice are mirrored in behaviours preceding and following a choice.

Choice is central to foraging; how animals allocate their choices under various environmental conditions to optimize their procurement of food and other necessities such as water and mates [1–4]. According to this Darwinian perspective, a suite of adaptations allows animals to identify the distribution of responses across options that maximize the returns on their choices. In characterising how animals achieve this, we address the fundamental question of how organisms collect information from the environment and use it to behave adaptively [5].

One influential description of foraging behaviour is the Matching Law, which states that the ratio of choices to one option matches the ratio of rewards earned from that option [6]. If these ratios are expressed logarithmically, the slope and intercept of a best-fit regression line capture the sensitivity of choice to changes in reward ratios, and response bias, respectively [7]. This derivative of the Matching Law, termed the Generalized Matching Law (GML), characterises the effect of different types of reinforcement schedules on this seemingly innate tendency for animals to adjust their choices in response to environmental payoffs [1,8,9].

Here, we study two types of environments that require different optimal choice strategies. In one environment (termed reward Hold), the probability of reward is dynamic and contingent on previous choices. This simulates an environment where resources deplete and renew. Another environment (termed reward No-Hold) has a constant probability of reward and mimics resources which do
2. Material and methods

(a) Subjects
Subjects were 36 experimentally naive, adult, male Wistar rats with initial weights of 159–191 g. Rats were housed in a climate-controlled colony room on a 12 L:12 D cycle (lights on at 07.00 h). Rats were given 1 h free access to food and water after each session and body weights were monitored to ensure they did not drop below 85% of their ad libitum weight.

(b) Apparatus
Rats were trained in a chamber measuring 30 cm (length) × 20 cm (width) × 50 cm (height). A nose-poke aperture (4 cm × 4 cm) was located on the front wall of the chamber. To the right and left of the aperture were drinking spouts which delivered 5% sucrose solution. The spouts were fitted with sensors to detect licking at a temporal precision of 1 ms. The experiment was controlled and data recorded using custom-written programs in MATLAB (The MathWorks, Inc.).

(c) Task
Rats initiated a trial by performing a nose poke and maintaining it for a variable delay (100–600 ms, uniform distribution). Two diode lights located in the middle of front wall of the nose-poke aperture were lit after this delay to indicate the go-signal. The lights remained on until a choice was made. Rats indicated their choice by licking either the right or left spout. If a reward was programmed for the chosen spout, it was delivered after a variable delay provided that rats maintained licking (100–600 ms, uniform distribution). If rats abandoned licking prior to the delay allocated for that trial, then the trial was counted as an unrewarded trial. Each reward was delivered using mechanical pumps located outside of the testing chamber, and consisted of a 0.5 s delivery of 5% sucrose. If no reward was available for the chosen spout (unrewarded trial), there was no external event to indicate the absence of reward. There were no restrictions on initiation of the next trial. Each experimental session consisted of at least 250 trials, throughout which background noise (approximately 70 db in volume) was present to mask any extraneous sounds.

Reward probabilities were independent for each spout. Therefore, on any given trial, reward could be available on one, both or neither of the spouts. Three probability pairs were used: 60–40, 70–30 and 80–20; counterbalanced for position (left or right) within groups.

For group Hold, the reward contingencies had an additional manipulation; specifically, an uncollected reward from a previous trial remained available until collected, but did not accumulate in reward amount (figure 1a). This ‘holding’ of reward means that the probability of reward at a spout increases as a function of consecutive trials spent on the other spout (figure 1b). The reward probabilities for group Hold were thus dynamic, dependent on the choices that rats made. By contrast, for No-Hold groups, the probability of reward remained constant throughout the session. For both groups, the reward amounts were identical on every trial, hence only the probabilities of reward changed.

(d) Procedure

(i) Spout shaping
Rats were placed in the experimental chamber for 10 min and reward was freely available from both reward spouts. The nose-poke aperture was blocked at this stage.

(ii) Nose-poke shaping
Rats were rewarded only after performing a nose poke. The nose-poke delay and spout delay were gradually increased over three sessions (first session: 100 ms; second session: 100–400 ms; third session: 100–600 ms). During this stage of shaping, rats quickly learnt the task structure and the range of delays at which reward was delivered.

(iii) Training
Rats were presented with the assigned reward probabilities for five acquisition sessions. The reward probabilities were then reversed for a further five sessions (reversal sessions).
(e) Analysis

All analyses were carried out using MATLAB, GRAPHPAD and SPSS. Results are reported according to the reward contingencies that were allocated for that session; High during reversal refers to the reward spout that had a higher allocated probability of reward (which was Low during acquisition).

We decomposed a single trial into three time windows, corresponding to distinct stages of the choice process. The first time window was choice execution latency, defined as the duration of the interval between exit from the nose poke and arrival at a reward spout. We did not distinguish between rewarded and unrewarded trials when examining choice execution latencies because prior to arrival at a reward spout, there was no indication if a trial would be rewarded or not. The second time window was spout sampling duration, defined as the duration of the interval between the first and last spout contact on unrewarded trials only. This duration provides an index of how long rats maintained responding at a particular reward spout for an uncertain outcome [13]. The third time window was the inter-trial interval, defined as the duration of the interval between the last spout contact and the initiation of the next trial by nose poke. Here, we analysed rewarded and unrewarded trials separately in order to investigate how the experience of reward affected these latencies.

To apply the GML [7], data from each session were transformed with a base 2 logarithm in order to eliminate skew and enable quantification of the relationship between variables through a linear regression [14]. The following equation was used [7]:

$$\log \frac{R_{\text{High}}}{R_{\text{Low}}} = S \left( \log \frac{R_{\text{High}}}{R_{\text{Low}}} \right) + \log b,$$

where $B$ refers to response to the two alternatives indicated by the subscript. In this study, we examine choices to the alternatives, as well as time spent performing responses to the alternatives. $R$ refers to amount of rewards obtained (from High or Low). The parameters $S$ and $b$ indicate the sensitivity of responses to reward ratios and the bias in responding that arise independently of rewards obtained, respectively [7].

When comparing distributions of response latencies for High versus Low within groups, data were pooled across individual rats and tested for significance using the Mann–Whitney U test as data were not normally distributed. To compare spout sampling durations between groups in figure 6b, data from each rat were normalized according to individual means and standard deviations. To do this, we first combined High and Low spout sampling durations from every session and calculated the mean and standard deviation of this distribution. This mean was then subtracted from each of the spout sampling durations and the value was divided by the standard deviation. This was repeated for each rat, and the normalized spout sampling durations were then combined within groups and shown separately for High and Low. These latencies were not logarithmically transformed. For this figure, we also investigate how the recent reward history affects differential spout sampling durations by calculating the number of rewards rats received from High versus Low in 50-trial bins.

For the autocorrelation analysis of the choice sequences, we coded each individual rat’s choice sequence as 1 for every High choice and 0 for every High choice. Autocorrelations ($r$) across lags ($k$) were calculated according to the following formula:

$$r_k = \frac{\sum_{i=1}^{N} (y_i - \bar{y})(y_{i+k} - \bar{y})}{\sum_{i=1}^{N} (y_i - \bar{y})^2},$$

where $y$ refers to the choice on the ith trial and $y$ refers to the mean choice across the total $N$ trials. Autocorrelations were calculated from normalized choice sequences in MATLAB using the ‘autocorr’ and ‘zscore’ functions. To establish the consistency of patterns across rats, we calculate the mean autocorrelation

Figure 2. Group mean (± s.e.m.) High choice proportions (a) and performance (b) across five acquisition sessions (1–5 on x-axis) and five reversal sessions (6–10 on x-axis). Grey vertical dashed lines indicate the switch from acquisition to reversal probabilities.

3. Results

Figure 1a shows the discrete-trial structure of the task, where trials were subject-initiated and rewards were delivered probabilistically after a variable delay. After five acquisition sessions, probabilities were reversed for five reversal sessions. For group Hold, an uncollected reward from a previous trial remained available (but did not accumulate in amount) across trials (figure 1b). Figure 1c shows that by choosing High on consecutive trials, the probability of reward on Low eventually exceeds the probability of reward on High and indicates the trial number at which this occurs for each probability condition. This represents the optimal strategy for group Hold: repetitions of a sequence of consecutive High choices followed by one Low choice [15]. This optimal Hold strategy requires that rats keep track of the average probability of reward, as well as the distribution of the past choices which control the dynamics of renewal and depletion of reward. For No-Hold groups, the optimal strategy is simply to respond exclusively on High since the probability of reward on Low is static over the entire session (see the electronic supplementary material, figure S1 for more detail on strategies).

Figure 2a shows the mean proportion of High choice for each experimental group. Across five acquisition sessions, High choice proportions for 60-H decreased and remained stable at approximately 0.5 over reversal sessions. All other groups increased their choices to High across acquisition and this proportion matched the reward probabilities presented for groups 70-H and 80-H. Proportion High choice for 70-NH and 80-NH increased to above 0.9 over acquisition, approaching the optimal strategy of exclusive High choice. On the first reversal session, proportions of High choice for
all groups were low (between 0.45 and 0.6 for Hold groups and between 0.35 and 0.55 for No-Hold groups). These choice proportions increased over reversal sessions in a similar fashion to the increase seen during acquisition, but did not reach the same terminal proportions.

Since groups differed in the number of rewards that were potentially available for collection (electronic supplementary material, figure S1), we quantified performance by expressing the number of rewards that rats obtained as a proportion of the rewards that could have been obtained with the optimal strategy. Figure 2b shows that all subjects approximated the optimal strategy with a terminal performance of above 0.8 at acquisition and reversal for all groups. To further assess this degree of optimality, we performed autocorrelation analyses on the choice sequences of rats in group Hold to identify past trials which were most correlated with the current trial. For comparison, the same choice sequence was shuffled 100 times for each rat. This method of shuffling removes the sequential contribution of strategy yet preserves overall choice proportions. Figure 3a, left, shows example autocorrelations of Low choice sequences for one rat from each Hold group. Across a window of 20 trials, the correlation coefficients and the shape of autocorrelations indicate the periodic pattern with which rats visited spouts. By contrast, correlation coefficients were highest for a trial lag of 3 for 70-H, and trial lag of 5 for 80-H. The gradual changes in correlation coefficients which cycle over multiple trials indicate the periodic pattern with which rats visited Low. These cyclical choice sequences are present in the averaged autocorrelations (figure 3a, right column) and indicate that rats had identified the periodic nature of the optimal strategies despite not being a perfect replication of the optimal strategies (autocorrelation functions for each optimal strategy are shown in the electronic supplementary material, figure S2).

To further characterize the periodicity of choices, we measured the number of consecutive choices made to High and Low during the last session of acquisition and reversal. The frequency of a single choice was highest for 60-H, followed by 70-H and 80-H (figure 3b). The similarity across the four panels (acquisition versus reversal, and choices to High versus Low) for 60-H provides additional evidence that rats in this group treated the two spouts equivalently. By contrast, most trials consisted of consecutive High choices of more than 3 for 70-H, and an even greater proportion for 80-H. This was not the case for Low as, during acquisition and reversal, 70-H never chose it more than three times consecutively, and 80-H never more than twice. For No-Hold rats, the frequency of less than 10 consecutive choices to High was low and a large proportion of trials consisted of more than 10 consecutive choices to High (figure 3c, note the different scale of y-axis from figure 3b).

To further investigate how behaviour was controlled by reward contingencies, we applied the GML to the ratio of rewards obtained from High against ratios of High choice, High choice execution latencies and High spout sampling durations (figure 4; electronic supplementary material, table S1). With the exception of spout sampling durations during reversal for group 60-H, the ratio of rewards obtained from High accounted for a significant amount of variance for all

**Figure 3.** Choice patterns. (a) Autocorrelation functions of choice sequences for an example rat from each Hold group (left column) and averaged autocorrelation functions across rats (right column). Grey shaded bands in left column are mean and s.e.m. of 100 shuffles of each rat’s choice data. Grey bands in right column are the s.e.m. for the averaged autocorrelation functions. (b) Group mean and s.e.m. of frequency of consecutive choices during the last sessions of acquisition and reversal for Hold groups. (c) Group mean and s.e.m. of frequency of consecutive choices during the last sessions of acquisition and reversal for No-Hold groups.

<table>
<thead>
<tr>
<th></th>
<th>60-H</th>
<th>60-NH</th>
<th>70-H</th>
<th>70-NH</th>
<th>80-H</th>
<th>80-NH</th>
</tr>
</thead>
<tbody>
<tr>
<td>60-H</td>
<td>0.50</td>
<td>0.40</td>
<td>0.60</td>
<td>0.50</td>
<td>0.70</td>
<td>0.60</td>
</tr>
<tr>
<td>70-H</td>
<td>0.60</td>
<td>0.50</td>
<td>0.70</td>
<td>0.60</td>
<td>0.80</td>
<td>0.70</td>
</tr>
<tr>
<td>80-H</td>
<td>0.70</td>
<td>0.60</td>
<td>0.80</td>
<td>0.70</td>
<td>0.90</td>
<td>0.80</td>
</tr>
</tbody>
</table>


three variables (electronic supplementary material, table S1; all $p < 0.01$). This indicates that the ratios of obtained rewards determined the distribution of choices between High and Low, the total time spent moving towards choices (choice execution latencies, figure 4b, left), as well as the total time spent responding at reward spouts on unrewarded trials (spout sampling durations, figure 4b, right). Within No-Hold groups, $r^2$ values decreased systematically according to reward contingencies (60-NH largest and 80-NH smallest), with no such pattern for Hold groups.

Since the duration of time spent moving towards the spouts, and the duration of time spent at the spouts on unrewarded trials both correlate with the ratio of choices, the GML analysis of figure 4b does not reveal the direct relationship between reward ratios and these behaviours. We then asked whether the differential status of reward spouts (High versus Low) directly impact the temporal dynamics of the rats’ behaviour on individual choice trials (i.e. choice execution latency, spout sampling duration and inter-trial interval). Figure 5 compares spout sampling durations between High and Low spouts for Hold (top) and No-Hold (bottom) during acquisition, where the lateral shift indicates a systematic difference between the distributions. Here, spout sampling duration for High was significantly longer for all groups (all $p < 0.01$; Mann–Whitney U test). The differences in spout sampling durations were robust as they were present for all groups over acquisition and for all groups except 60-NH at reversal. We, therefore, examined how quickly subjects adjusted this differential sampling time following reversal. Figure 6 summarizes the differences between distributions by plotting the medians for choice execution latencies and spout sampling durations, and the corresponding analysis for inter-trial intervals is in the electronic supplementary material, figure S4.

For choice execution latencies (figure 6a) during acquisition, with the exception of 60-H and 60-NH, all groups were significantly faster to execute a High choice than a Low choice ($p < 0.01$; Mann–Whitney U test). Following reversal, only 80-H and 70-NH showed a corresponding reversal in their latencies, with significantly faster latencies to execute a High choice as compared with Low ($p < 0.01$; Mann–Whitney U test.)

Spout sampling durations showed that rats maintained responding on High during unrewarded trials for a significantly longer duration than for Low (figure 6b, $p < 0.01$; Mann–Whitney U test). The differences in spout sampling durations were robust as they were present for all groups over acquisition and for all groups except 60-NH at reversal. We, therefore, examined how quickly subjects adjusted this differential sampling time following reversal. Figure 7a shows that, on the first reversal session, despite the rats persisting in choosing the reward spout which had the higher probability of reward during acquisition (left panel), their spout sampling durations had already adjusted to the reversal probabilities (right panel). For this individual session analysis (figure 7a), all groups exhibited longer spout sampling durations for High than Low. Although 60-NH show a small reversal in spout sampling duration on the first session of reversal, this reversal did not persist throughout the whole of reversal (figure 6b). Figure 7b plots quantiles (2% steps) from the High and Low spout sampling durations during the first reversal session. The Q–Q plots for all groups show that nearly all data points fall above the diagonal line, indicating that equivalent quantiles occur at longer spout sampling durations for High compared with Low.
Figure 6. Group medians for choice execution latencies (a) and spout sampling durations (b).

For comparison, Q–Q plots from the middle of acquisition are also shown.

The differential status of reward spouts was also reflected in the latencies with which rats moved away from a past choice in the inter-trial interval (electronic supplementary material, figure S4). Although there was variability depending on trial outcome (rewarded or not) and session type, overall, rats took longer to leave High and initiate a new trial than they did for Low.

We then sought to identify the relationship between ratio of rewards received and differential latencies between High and Low. In order to see how this relationship withstood variable conditions, we combined acquisition and reversal data as well as collapsed across probability groups. For this analysis, we again transformed ratios with a base two logarithm so that the linear relationship between variables could be examined. The ratio of obtained rewards was negatively correlated with the ratio of medians of choice execution latency distributions (for group Hold: Pearson’s r = −0.47, p < 0.01; for group No-Hold: Pearson’s r = −0.58, p < 0.01; figure 8a, left). Greater ratios of rewards obtained from High made the difference in median choice execution times more negative. A negative value for the median difference indicates longer latencies to execute a Low choice compared with a High choice. By contrast, the ratio of obtained rewards was positively associated with the difference between the medians of spout sampling duration distributions (for group Hold: Pearson’s r = 0.24, p < 0.01; for group No-Hold: Pearson’s r = 0.39, p < 0.01; figure 8a, right). This indicates that the more rewards obtained from High, the longer were the spout sampling durations on High compared with Low.

Finally, we examined whether group differences in the number of rewards obtained in 50-trial bins characterised group differences in spout sampling durations. We normalized spout sampling durations to allow for comparisons between groups. The difference between the numbers of
The GML was a good description of the data in terms of choice ratios and the ratio of time spent performing various aspects of choice behaviour. Deviations from matching (sensitivity parameters larger or smaller than 1) indicate that choice is disproportionally allocated to one option [14]. In this regard, a striking pattern in the data was that group 60-H showed the most deviation from matching across all three variables as well as the smallest $r^2$ values. This might reflect the optimal strategy of alternating for both acquisition and reversal. According to this suggestion, it was advantageous for 60-H rats to ignore the differences in rewards obtained from High versus Low and the changes in reward probabilities over time, and instead allow the alternating optimal strategy to exert the greatest control over behaviour. Our measure of performance (figure 2b) indicates that 60-H rats obtained almost all of rewards that were available, further emphasizing that 60-H rats closely adhered to the optimal strategy.

In previous studies, over-matching has been observed when choices were costly, as when subjects have to traverse a barrier between two options [17]. This competes for control over behaviour as it produces a cost for switching between options. In this study, we also observed over-matching for choice allocation during reversal for Hold rats. This over-matching might reflect the computationally costly periodic optimal strategy as compared with the simpler strategy of exclusive High choice for No-Hold rats. Clearly, the nature of the cost in the current study differs from the cost imposed when traversing a barrier but, taken together, these findings suggest that the ‘difficulty’ of choice mediates the sensitivity of behaviour to reward contingencies.
(c) Behavioural latencies contain environmental information that guide choice

The differential reward status (High versus Low) of spouts had dissociable effects on the temporal structure of behaviour across different components of a single trial. The most robust observation here was the longer spout sampling durations during unrewarded trials for High compared with Low. It is striking that this differential spout sampling duration is evident even for group 60-H during acquisition. Despite 60-H rats making an equivalent number of choices to High and Low, the slight advantage of the High spout is reflected in this measure of commitment to a choice. This is further demonstrated by the replication of our previous finding that differential spout sampling durations emerge even on the first day of reversal, prior to the adjustment in choice proportions [13]. These two findings are clear indications that behaviour immediately subsequent to an explicit choice can be a more sensitive reflection of the information in the environment acquired by the rats. Binary choice is a relatively crude measure, whereas temporal duration allow for gradients in preferences to manifest.

Choice execution latencies showed a contrasting pattern from spout sampling durations; a greater number of rewards obtained from High was associated with shorter choice execution latencies to High compared with Low. While this was observed during acquisition, a reversal in reward contingencies did not result in the same consistent reversal in choice execution latencies as observed for spout sampling durations. It is possible that choice execution latencies are a less sensitive measure of the relative status of reward spouts. Since it is more distant from the actual reward than spout sampling durations, choice execution latencies might reflect information about reward history from previous sessions in addition to the immediate High-Low status of reward spouts. Alternatively, individual rats may have different points during a trial at which they had made a left or right choice. Because these self-initiated trials were uncued with respect to which reward spout to move towards, there was not a specific point within a trial at which rats had to make a choice. It is also possible that this choice time point changed across sessions concomitantly with rats becoming more familiar with the structure of the task. Finally, the well-practiced choice execution response of selecting High at acquisition could remain intact and still control behaviour at reversal. In other words, while the reversal of reward probabilities produced changes in behaviour to the spout which was Low during acquisition, well-learned, habit-like responses to the other spout (High during acquisition) may have remained intact.

The shorter choice execution latencies to High compared with Low can also be seen as an index of learned value, where the reward history at the High spout encourages approach behaviour. The same mechanisms of learned value also accounts for the differential latencies to initiate a new trial (electronic supplementary material, figure S4). Overall, rats took longer to initiate a new trial after just visiting High than Low regardless of whether that visit was rewarded or not, indicating that the reward history at High encouraged rats to remain there.

Reinforcement learning has been used to explain how animals acquire adaptive responses. Models of reinforcement learning, such as temporal difference learning, hold that a critical aspect which underpins learning is the evaluation and updating of one’s expected action outcomes with the experienced action outcomes [18]. The discrepancy between expected and actual outcomes, or prediction error, is governed by the amount of surprise elicited by such discrepancies [19–21]. In the present experiment, the inter-trial interval may be when such choice evaluation occurs. One way in which the mechanisms that underpin such group differences in inter-trial intervals (electronic supplementary material, figure S4) might be made clearer is through electrophysiological recordings. A comparison of neural activity at different time points within a trial might help identify the different brain areas responsible for various aspects of decision making. While the basal ganglia, with its rich dopaminergic innervation, is implicated in choice evaluation [22,23], the orbitofrontal cortex is implicated in representations of value [24]. Neural activity can then be tracked over reversal sessions to determine how and when neural changes occur in relation to changes in behaviour.

The effect of environmental contingencies on choice behaviour can be summarized according to the three discrete components of a trial we have identified here: a highly valued option excites approach and shortens the time taken to move towards it, prolongs the amount of time spent trying to obtain the desired outcome, and discourages leaving the location of the highly valued option. We observed differences in behaviour across the three time windows of choice execution latencies, spout sampling durations and inter-trial intervals. This emphasizes that these three components within a trial index different aspects of a single choice which need to be considered when trying to obtain a complete picture of the dynamics of choice.

Ethics statement. All procedures were approved by the Animal Care and Ethics Committee at the University of New South Wales.

Data accessibility. All data used for the analyses have been uploaded to Dryad at doi:10.5061/dryad.j55sn.

Acknowledgements. We thank Nathan Holmes for valuable discussions.

Funding statement. This work was supported by the Australian National Health & Medical Research Council Project grant no. 1028670.

References


