# Introduction

## Inference-Bound Decision-Making in Gerbils

In the realm of foraging, optimal decision making is important. The question isn't just "When does an animal eat?" but "How does an animal decide when to move on in search of more food?" One of the most prominent models of optimal foraging (i.e., maximizing the total energy or reward intake during time-constrained foraging) is the *marginal value theorem* (MVT). Introduced in 1976 by Charnov et al, this model predicts that a foraging agent should leave the current spout when the instantaneous energy intake rate (i.e., reward rate) within a given spout diminishes to the average reward rate in the entire environment (i.e., mean reward rate). This model assume access to fixed probabilities and an ability to gauge the current versus the average reward rate. However, this rule cannot be followed in a highly variable environment where the probability rules are not known. In our paradigm, although through experience, the animal may eventually understand that it receives less reward for every consecutive nose poke, the initial reward probability (spout quality) in every trial is not cued to the animal. Therefore, it is impossible for the animal to compute the current and average reward rate. So, what guides their decision in this uncertainty? Do they follow any specific rules or they act in a random manner? How do they adapt to our dynamic environment? To investigate the aforementioned questions, we focussed on the spout-leaving behaviour during the foraging task which reflects the animal’s dynamic decision-making i.e., when is the right time to leave the current spout and explore the other?

The variable resident times and rewards across different spout qualities (Results Fig) indicate that our gerbils did not conform to fixed-time or fixed-number rules, which prescribe consistent duration or reward counts before leaving a spout. These rules may have benefitted if the reward probability rules are constant across trials. However, in our case the starting reward probability is dynamic and hence a fixed time or fixed reward rule may end up failure. For example, in a fixed reward rule, the forager will tend to stay longer in a low-quality spout (Starting probability = 0.5) compared to a high-quality spout (Starting probability = 1) making them spend more energy for less reward by preventing them to optimally switch to another spout.

In highly variable environments it is rather difficult to reliably estimate the quality of the current spout at the time of entering. Instead, each reward capture may inform the quality of a spout thereby influencing the tendency to stay in the current spout. Following this principle, the incremental rule which states that the more the number of rewards the animal receives, more time it takes to leave the current spout. Although, the resident time and number of rewards per trial type (Results Fig) indicate towards this rule, the number of consecutive unrewarded poke before leaving gives a contradicting viewpoint (Results Fig). After the last rewarded poke in every trial, the number of rewards obtained thus far should have influenced the animal to stay longer or shorter based on whether it has received more or less rewards. Intriguingly, despite different starting reward probabilities, gerbils consistently left a spout after a certain number of consecutive unrewarded pokes, not purely driven by the quantity of rewards received before. The gerbils, despite not knowing the exact probabilities, behaved as if they were weighing their recent experiences. This behaviour aligns with the Giving Up Time (GUT) rule – a forager only stays for a certain period without a reward following the last successful forage (last reward). Exceeding this threshold prompts a switch to another spout, with each reward resetting the GUT. In other words, a single reward appeared to reset their attempt to explore, hinting that they valued this positive outcome more than the absence of it—a hallmark of inference over mere stimulus-response. This ability of the gerbils to form an inference about the action-outcome in the foraging task and alter their action based on recent experiences rather than fixed cues, hints at an underlying complex functional neural circuitry that guides this adaptive, experience-informed decision-making process.

Human and animal studies have revealed four major rules that is being followed during this situation (introduction figure). These are Fixed-time rule (ref), Fixed-N rule (ref), Incremental rule (ref), and Giving Up Time rule (ref). Our results showed that the gerbils showed a difference in their resident time and correspondingly the number of rewards obtained based on the spout quality (trials starting with different starting reward probability, Results Fig.1D and 1E). This indicates that the animals clearly didn’t follow the Fixed-time rule and Fixed-N rule as these would have required the animal to stay for the same time (Fixed-t-rule) or stay until they receive the same number of rewards (Fixed-N-rule) irrespective of the spout quality. In case of incremental rule, the more the number of rewards the animal receives, more time it takes to leave the current spout. Although, the resident time and number of rewards per trial type indicate towards this rule, the number of consecutive unrewarded poke before leaving gives a contradicting viewpoint. Think about it, after the last rewarded poke in every trial, the number of rewards obtained thus far should have influenced the animal to stay longer or shorter based on whether it has received more or less rewards. However, the animals have made consistent number of consecutive unrewarded pokes before leaving a current spout irrespective of the starting reward probability or the number of rewards obtained in a trial. This indicates that the animal didn’t follow incremental rule but followed a GUT rule. This is because, the animals for 25% of the trials left the current spout after experiencing two unrewarded pokes after the last rewarded poke irrespective of the rewards it obtained before in that trial.

Why this way of behaviour is useful? Back it up with literature.

GUT does not require prior knowledge about the probability rules which in our case works well as the animals are not cued about the hidden starting probability rule in every trial. Therefore, in order to follow a GUT rule, the animal needs to form an inference about the task based on its accumulating experience. Therefore, in our study, the gerbils perform an inference-bound spout leaving behaviour.

## Neural Encoding of Reward Expectation and Evaluation in FrA

* Propose that the distinct AVREC patterns for unrewarded pokes, particularly the first unrewarded poke, imply a nuanced encoding of expectation and evaluation of reward outcomes in the FrA.
* Explain how these patterns go beyond motor preparation, suggesting a complex cognitive process where expectation and evaluation are intertwined (to be cited).
* Elaborate on the implications of heightened activity in the last poke and theorize its relationship to combined cognitive processes of reward expectation and motor planning

In our study, the spatiotemporal dynamics within the frontal region A (FrA) disclosed by current source density (CSD) profiles offer insights into the gerbil’s neural encoding of reward expectations and outcomes during a probabilistic foraging task. Notably, the AVREC peaks (Figures 3 and 4) denote more than just motor activity; they signal a complex interplay of action and anticipation, where reward expectation is intrinsically linked to subsequent motor actions.

The encoding of the expected (towards the end of pokes) and evaluation of reward (subsequently) shows different activity patterns during the exploitation (first poke until last rewarded poke) and exploration (last poke) phases of the trial. Both the rewarded pokes (first rewarded and last rewarded pokes) show a similar waveform where there is an early peak in amplitude immediately after the end of the poke that may encode for the expectation of a reward followed by a slightly increased peak occurring in less than 250 ms from the end of the pokes that correlates with the food pellet reward being received and hence, may encode for the reward evaluation.

The deviation observed in the first unrewarded poke, with an early peak preceding the end of the poke (Results Fig.4), along with a subsequent dip, suggests an encoding of prediction error rather than reward receipt. This nuanced representation implies that the FrA is involved in evaluating outcomes against expectations. Moreover, the pronounced increase in amplitude for the early peak in the last unrewarded poke—indicative of heightened neural processing—may reflect the gerbils' integration of reward expectation with the cognitive demand for movement planning towards a new foraging site.

These findings support the notion that the FrA contributes to a decision-making process that extends beyond reflexive reactions to stimuli, encompassing a learned anticipation of outcomes—a foundational aspect of inference-bound behavior.

## The Dynamic Transition of Expectation in FrA Activity

* Interpret the non-linear RMS trend in the early phase as indicative of evolving expectations during a trial.
* Illustrate the 'U' shaped pattern, discussing how it may reflect the cognitive shift from a state of expecting a reward to uncertainty and back to expecting no reward.
* Argue for the role of layer-specific cortical activity in managing these shifts, with an emphasis on how such dynamic changes underscore the cognitive complexity in gerbil decision-making (to be cited).

The AVREC of FrA encodes the reward expectation (<100 ms after end of the poke) and evaluation (>100 ms after the end of the poke) signalling a complex interplay between action and anticipation. The increased neural activity during the early phase of the last poke revealing the integration of reward expectation and complex movement planning makes the last poke unique (Results fig).

When the animal is still at the end of exploitation phase (first unrewarded poke after last rewarded poke), less activity is observed at the early phase but extended persistent activity patterns observed after 100 ms following the nose poke (reward evaluation) indicates that the animal is still focussed only on getting a reward and not at the nose poking activity itself. Conversely, at the onset of the exploration phase (last poke), heightened frontal activity is evident within 100 ms after the nose poke (early phase) indicating a focus shift from the reward evaluation towards the nose poking action that is tightly coupled with motor planning. This shift in activity pattern makes the last poke unique and hints that the early phase could be the crucial phase where the primary decision to explore may get encoded.

How do I say that the expectation gets altered over series of poke?

Individual poke analysis revealed that the FrA shows a non-linear evolution of activity patterns during the shift from exploitation to exploration indicating an alteration of expectation over time (Results Fig – U curve). Initially, at the end of exploitation (first unrewarded poke after last rewarded poke), the FrA produces a prediction error signal as the animal still expects a reward after a nose poke but didn’t get one (Results Fig – dip in neural activity after 100 ms (brown curve)). As the unrewarded pokes keep accumulating, the animal moves into a fuzzy state of not knowing what to expect exactly thereby a decrease in overall frontal activity as seen in U figure (n-6 to n-2). Once the animal realizes that the reward is exhausted at this spout, it shifts its focus on the nose poking activity and start to expect a no reward as an outcome. As this gets confirmed (n-2 to nth poke), the activity again starts to increase and finally the animal leaves the spout. This intricate alteration of expectation over time guided by the current and previous experience may indicate the possibility of value alteration to reward and no-reward which forms the basis for the inference guided decision-making behaviour that is seen.

This alteration of expectation from expecting a reward to expecting a no reward for the same nose-poke action demands the involvement of a differential layer activity in FrA. (En apdi solra!?)

## Functional Implications of Layer-Specific Activity During Foraging

* Postulate why increased activity in supragranular layers during exploration may be linked to their involvement in long-range cortical connections and attentional resource allocation.
* Relate the role of deeper cortical layers in exploitation, drawing parallels to working memory and possibly dopaminergic influences (to be cited).
* Bridge observed results with existing literature to support the proposed functions of layer-specific activity during different phases of the foraging task.

Laminar CSD recordings allowed us to investigate the underlying local layer-specific FrA micro-circuit activity during the foraging behaviour (Results Figures). Our results disambiguated that the allocation of exploitation or exploration strategies differentially orchestrated the FrA circuitry in a layer-dependent manner (Results Fig.6).

Looking at the whole foraging paradigm from an attentional resource allocation problem inside the brain, when the animal decides to explore other spout after exploiting the current one, there clearly requires a change in attentional resources to shift its focus from current spout and move towards another one. According to reinforcement learning theories, goal-directed behaviour requires the estimation of the reward expected from a particular stimulus or action. Such value estimates can be identified by the prediction error: the difference between expected reward and reward actually obtained (Schultz, 2015). Such reward-coding in the brain is based on reinforcement-evaluating brain structures, like the ventral tegmental area (VTA), which convey information about stimulus salience and valence to target areas distributed throughout the brain including the frontal cortex (Bromberg-Martin et al., 2010).

Transition of overall FrA activity revealed an underlying neural circuitry that alters the expectation of rewards over time thereby guiding the gerbils to shift from exploitation to exploration (Results Fig – U curve).

Studies focussing on other parts of neocortex such as sensory cortices (ref) have informed that the supragranular layers have anatomical connections that enables cross-columnar activations and long-range inter-cortical connections. This offers a potential platform for reallocating attentional resources required for exploratory behaviour. On the contrary, the recurrent corticoefferent feedback loops originating in infragranular layers (Avery and Kirchmar, 2015, 2017) updates the persistent working memory content through reward-prediction error signal thereby, offering a potential service of the maintenance of exploitation strategies. Based on this, we hypothesized that an exploratory behaviour i.e., change in foraging site, will be correlated with strong recruitment of upper layers, deeper layer activity will coincide with the exploitation of current foraging site.

Layer analysis showed that during the shift from exploitation to exploration, the biggest difference seen in overall FrA activity (results fig) was mainly contributed by the differences observed in the upper layers (Layer I/II and III/IV). Close observation revealed that the upper layers showed increased neural activity during the exploratory phase compared to the deeper layers. Furthermore, individual poke analysis revealed that the activity of upper layers showed a similar non-linear trend as observed in overall FrA activity and also as expected, the activity escalated during the early phase just before the animals decided to abandon the current site and explore the other. These results support our hypothesis that the FrA indeed adopts a layer-dependent processing to adequately allocate attentional resources in the cortex contributing effectively to the dynamic change in search strategies during the foraging behaviour.

## Performance of the foraging behaviour

In order to see if the animal has learnt the foraging task and reduce the possibility of random exploration, we looked at their travel time. Here, travel time is defined as the time between the end of last poke in a trial and the start of the first poke in the succeeding trial (Fig1B, Eq.2). Lesser and consistent travel time indicate that the animal is goal-directed and not randomly exploring. The consistent and lower travel time indicates that after each trial, the animals were directed towards the other spout and not randomly exploring the cage (Fig.2A). Based on this, only sessions from 6 to 20 were considered for further behavioural and electrophysiological analyses.

### Animals make inference-based decisions

To evaluate how well the animal learnt to make decisions during the exploitation-exploration dilemma, we focussed on the time point where the animal decided to shift from exploitation to exploration. This is because every trial begins with exploiting the current spout for rewards and slowly as the rewards gets exhausted, the important decision is to decide when to leave the current spout to explore the other. This can be interpreted by the consecutive number of unrewarded pokes the animal makes before leaving the current spout.

Following the experimental design (Fig.1C), Fig.2B shows that the distribution of total rewards shifts towards right for trials starting with higher initial reward probability. This indicates that the if the animal makes same number of pokes in each trial, it will receive a greater number of rewards for trials starting with higher initial reward probability. On the contrary, this distinction between trials starting with different starting reward probabilities gets lost when it comes to the consecutive unrewarded pokes made before switching (Fig.2C) In other words, it shows that the animals irrespective of the starting reward probability and the total number of rewards received in a trial maintains a consistent number of consecutive unrewarded pokes before leaving a particular spout. This consistency in decision making may reflect the possibility that the animals form an inference about the hidden reward structure based on their learning of the task.

## Distinct motor and reward related activity patterns in the frontal field A

To confirm if the target region FrA encodes distinct activity patterns that reflects the animal’s behaviour, four different time points that are unique and represents crucial stages during the foraging session were selected based on the animal’s performance (Fig1B and Fig2). These are first poke (rewarded and unrewarded), last rewarded poke and the last poke. The time between first poke until last rewarded poke can be considered as the exploitation phase because during this phase, even after experiencing unrewarded pokes, the animal insists on staying on the same side expecting more rewards. The last rewarded poke is a critical point as it represents the end of exploitation phase and beginning of transition towards exploration. This is because after the last rewarded poke, the animal slowly starts to alter its expectation and move towards exploration phase which is best represented by the last poke in a trial. Fig.3 shows distinct spatiotemporal neural activity within frontal field A that encode both the poke (motor) and the ensuing reward related information. The selected epochs represent -1 to +2 seconds from the end of the poke (black dashed line, t=0) to compare the neural activity during decision-making phase. Further, the differential reward related activity patterns helped identify and distinguish infragranular layers from superficial layers in the laminar recordings and perform channel-layer specification.

In order to visualize the overall frontal cortical activity, the CSD signals were rectified and averaged across the laminar electrodes to lose the spatial information. The average rectified signals (AVREC) from the CSD profiles also show distinct motor (Fig.4, peak around end of the poke (t=0)) and reward related signals (Fig.4, peak around 250 ms from the end of the poke). From the perspective of expectation and evaluation of reward, it can be seen that in FrA, the encoding of the expected (towards the end of pokes) and received reward (subsequently) shows different activity patterns during the exploitation (first poke until last rewarded poke) and exploration (last poke) phases of the trial. Both the rewarded pokes (first rewarded and last rewarded pokes) show a similar waveform where there is an early peak in amplitude immediately after the end of the poke that may encode for the expectation of a reward followed by a slightly increased peak occurring in less than 250 ms from the end of the pokes that may encode for the reward evaluation. On the other hand, in the unrewarded pokes (first unrewarded and last poke), the reward evaluation shows a dip followed by a later peak in amplitude around 500 ms that may encode for a prediction error from the deeper layers. Compared to all other pokes, the last poke which is the starting point of exploration phase where the animal decides to switch to the other spout, shows a heightened frontal activity immediately after the end of the poke (<100 ms). These distinct activation patterns for different time points seen in Fig.3 and Fig. show that the frontal field A encodes not just the motor activity but also the reward related activity (expectation and prediction error).

## Shifts in frontal activity patterns: Exploitation to exploration

### What makes the last poke the last one?

As we have confirmed that the frontal field A encodes distinct activity patterns for different phases (Fig.3 and Fig.4), it is essential to understand how the animal decides the last poke in a trial. This is essential as the animals may encounter multiple consecutive unrewarded pokes after the last rewarded poke (Fig.5A). Hence, the last poke also being an unrewarded poke needs to be distinguished from its preceding unrewarded pokes.

Fig.5B evidently shows that the last unrewarded poke has a distinct activity pattern compared to that of first unrewarded poke succeeding the last rewarded poke. In order to decode when the decision making exactly occurs, first 500 ms from the end of the poke was considered and two distinct phases were selected. Early phase represents the first 100 ms (Fig.5B, yellow region) and the late phase represents 100-500 ms (Fig.5B, light pink region) from the end of the poke. This way, we can verify if the decision making is happening immediately after the nose poke action or it also requires a reward evaluation information. When the animal is still at the end of exploitation phase (first unrewarded poke after last rewarded poke), less activity is observed at the early phase but extended persistent activity patterns are observed after 100 ms following the nose poke (reward evaluation) indicating that the animal is still focussed on getting a reward. Conversely, at the onset of the exploration phase (last poke), heightened frontal activity is evident within 100 ms after the nose poke (early phase) indicating a focus shift from the reward expectation towards the nose poking action. This shift in activity pattern makes the last poke unique and hints that the early phase could be the crucial phase where the primary decision to explore may get encoded.

### The evolution of frontal activity from exploitation to exploration

To get a deeper understanding of the evolution of change in frontal activity patterns from exploitation to exploration, we focussed on the individual pokes. We computed and compared the root mean square (RMS) of the AVREC signal for all the unrewarded pokes between last rewarded poke and last poke (Fig.5B and 5C) and z-normalized it within each animal. Over the transition from exploitation to exploration (Fig.5B, n-7 to nth poke), during the early phase, the overall frontal activity initially decreased (n-6 to n-2) and then increased just before the animal decides to leave the spout (n-2 to nth poke) (One way ANOVA with Bonferroni correction, \* p < 0.05, \*\* p < 0.01, \*\* p<0.001). However, during the late phase, as it mainly encodes for reward evaluation, it constantly decreases as all the pokes considered here are unrewarded. The non-linear trend of activity change during the early phase indicates the involvement of a complex processing strategy underlying the inference-based decision-making behaviour seen in the animals (Fig.2C).

## Layer specific motor and reward related activity patterns in the frontal field

Five distinct cortical layers were identified from the current source profiles of the laminar recordings (Fig.3). The source signal was removed and only the sinks were considered to ensure that the signal is contributed only by layer specific local excitatory synaptic populations. Comparison of layer-wise averaged sink activity between the last poke and first unrewarded poke succeeding the last rewarded poke indicates that the large difference in activity patterns observed in overall frontal activity (Fig,5A) is mainly contributed by the large difference seen in the superficial layers (layer I/II and III/IV) (Fig.6A). Further, shortly before the decision of the animal to explore the other spout, particularly layers I/II and III/IV showed two prominent activity peaks in amplitude directly at the retraction of the spout (early phase, 0-100 ms) and afterwards (late phase, 100-500 ms). In the individual poke RMS analysis, similar to overall frontal activity, there is an initial decrease in all cortical layers. However, at the beginning of the exploration phase (towards the last poke (n)), the activity of upper layers I/II and III/IV, as well as Vb, increases. The increase is mainly observed in the early phase following the nose poke, and then with regard to the evaluation of the absence of rewards (Fig.6B). In the late phase, with the exception of layer I/II and layer III/IV which showed a steep increase in activity just before the last poke, all the cortical layers showed a decrease in activity across all the pokes similar to that of overall frontal activity.