University of Oxford

Hierarchical Decision Making with Dynamic Reward Structures

Mengxi Wang

Christ Church College
Department of Experimental Psychology



A thesis submitted for the degree of Doctor of Philosophy in Experimental Psychology

Hilary Term 2021

Abstract

Decision making is an interdisciplinary field that studies how the brain processes information to come to a commitment or choice behaviour. Over the past decades, a cascade of psychological experiments have provided key insights in establishing principles of how the brain forms a decision. The majority of evidence came from studies with simple value comparison or categorical judgement, whereas in real life decisions usually require multiple levels of inferences, or integration of evidence that changes across time. This thesis will first evaluate relevant theories (chapter 1) and present two experiments with novel task paradigms to examine hierarchical decision making (i.e. multiple levels of thought processes) with dynamic reward structures (i.e. underlying reward structures vary across time). The first experiment (chapter 2) investigates foraging, a type of patch leaving decision that emerged in Ecology, with a virtual foraging task in which human participants choose between gold mines and decide when to leave to re-enter or enter a different one. The second experiment (chapter 3) investigates hierarchical perceptual decisions with a visual perception task, in which human participants find out the target patch and report its rotation by integrating sensory evidence and tracking reward contingency. I will demonstrate how the two experiments, combined with computational models and neurophysiological recordings, probe into the cognitive mechanisms of how the brain tracks changes and integrates evidence across time for a specific choice or long term estimate of option values. Finally, I will discuss how the findings contribute to extant theories and expand our knowledge of the neural basis of decision making in relatively complex, real life scenarios (chapter 4).

Acknowledgements

I would like to thank my two amazing supervisors, **prof. Jill O'Reilly** and **Dr. Nils Kolling** for all the valuable advice and support throughout the last three years. Jill has always been so patient in teaching me even the basics and has kept the best interest of me in mind. Nils is a real genius full of inspiring ideas and contagious enthusiasm about science. I started out my PhD relatively inexperienced, and my first project happened to a huge one with a lot of technical stuff to learn, but Jill and Nils have been so responsible and willing to offer me help whenever I was in need. Their advice easily cleared my confusion when I got stuck. It was not an easy journey and I know I have stumbled my way here. Without their support, I could not have come that far.

I also owe a lot to people from the **Kolling-Hunt-O'Reilly joint lab**. It is one of few great places that can gather so many brilliant and wonderful young scientists with genuine passion for decision making. The weekly journal club is definitely inspiring and has kept me up-to-date of the most pioneering work in the field. Sorry that I have missed a lot of lab socials due to my Sulkido training, self-study plans and late-night-introversion, but each of you has a special place in my heart.

The Rushworth group set a path towards my PhD journey for me. I had a wonderful time there as an Erasmus student and the people I met and the things I learned will be forever kept in my mind. Especially I would like to thank **Dr. Lennart Verhagen** for teaching me how to collect EEG data and do the pre-processing; and **Dr. Miriam Klein-Flugge** for all the valuable advice and comments on my projects; and **prof. Matthew Rushworth**, the most charismatic psychologist I have ever met, for having me in the group in the first place. Had I not joined the group, I would not have had a chance to meet Nils and Jill or learn to be a scientist.

A special thank you to my comrades-in-arms from the **Oxford University Sulkido Club** and **Taekwondo Club**. The training we took and the competitions we fought together comprised the best part of my memories of Oxford. Being a martial arts guru is no easier than being a scientist. My three amazing instructors, **James, Dev and Vojciech**, your rigour and enthusiasm have endowed me with all the physical and mental strengths to pull through difficulties in my life. One day I will kick your asses in sparring. Be prepared.

Finally, a big thank you to two outstanding women, **my dear mum Shuying Chen and grandma Xiuying Cheng**, for their immeasurable support and love. Every time they watched me leave at the airport they tried so hard to hold back tears, a scene I will never forget, and has fuelled me and helped me overcome all the obstacles on my way. Science has taken me away from them, yet they never questioned my choice. I would also like to thank my friends, for all the alcohol, gossips and lame jokes, for all the wonderful time we spent together that relieved the loneliness and anxiety throughout the journey. I am so grateful for everything. I truly am.

Statement of Contribution

The work presented by this thesis reflects the collaborative nature of interdisciplinary fields and owes much to the contributions of people within and outside my supervisory team. The foraging experiment (chapter 2) was designed in collaboration with Nils Kolling and Jill O'Reilly. I wrote the experiment code jointly with Nils Kolling. I collected the data on my own and led the preliminary analyses. Advanced behavioural modelling and EEG decoding analyses were led by Nils Kolling. The EEG decoding analyses received valuable advice from Cameron Higgins. The metacognition experiment (chapter 3) evolved from an fMRI study conducted by Emma Lawrence and Jill O'Reilly. I adapted the experiment code that was originally written by Emma Lawrence to suit the purpose of our study. I collected the data by myself and led the analyses. The writing of this thesis is entirely my own work and was carried out under the supervision of Jill O'Reilly and Nils Kolling.

Contents

Abstract **i**

Acknowledgements **ii**

Statement of Contribution **iv**

1 General Introduction **1**

1.1 Overview 1

1.2 Neural Substrates 6

1.2.1 Neural mechanisms of patch leaving decisions 6

1.2.2 Neural mechanisms of perceptual decision making 10

1.2.3 Common and distinct neural correlates of the two studies 11

1.3 EEG Signals 13

1.3.1 Event-related potentials 13

1.3.2 Brain oscillations 16

1.4 Computational Modelling 18

1.4.1 Reinforcement Learning 20

1.4.2 Drift diffusion model 22

1.4.3 Bayesian Learner 24

1.5 Current Research 27

2 Study 1: Foraging **30**

2.1 Introduction 31

2.2 Materials and Methods 34

2.2.1 Participants and Procedure 34

2.2.2 Behavioural Task 34

2.3 Behavioural Analyses 40

2.3.1 Reinforcement Learning 41

2.3.2 Bayesian Learner . . . 42
2.3.3 Multiple regression analyses of patch selection and leaving decisions . . . 46
2.3.4 Model Comparison . . . 49
2.4 Behavioural Results . . . 50
2.4.1 Basic Breakdown . . . 50
2.4.2 Regression Results . . . 53
2.4.2.1 GLM for patch choice . . . 55
2.4.2.2 GLM for leave choice . . . 57
2.4.3 Summary of Behavioural Results . . . 62
2.5 EEG . . . 64
2.5.1 Recording and Pre-processing . . . 64
2.5.2 Decoding Analyses Overview . . . 64
2.5.2.1 Standard Decoding Approach . . . 66
2.5.2.2 Hidden Markov Model (HMM) . . . 66
2.5.2.3 Temporally Unconstrained Decoding Analysis (TUDA) . . . 67
2.5.3 EEG Results . . . 71
2.5.3.1 Duration of states . . . 72
2.5.3.2 States' relation to regressors . . . 73
2.5.4 EEG Discussion . . . 82
2.6 Conclusion . . . 88

3 Study 2: Metacognition 89

3.1 Introduction . . . 90
3.2 Materials and Methods . . . 95
3.2.1 Participants and Procedure . . . 95
3.2.2 Behavioural Task . . . 95

3.3 Behaviour . . . 99
3.3.1 Behavioural Results . . . 99
3.3.2 Summary of Behavioural Findings . . . 108
3.4 Drift Diffusion Model (DDM) . . . 110
3.4.1 DDM Methods . . . 110
3.4.2 DDM Results . . . 112
3.4.3 DDM Discussion . . . 118
3.5 Bayesian Model (BM) . . . 121
3.5.1 BM Methods . . . 121
3.5.2 BM Results . . . 122
3.5.3 BM Discussion . . . 124
3.6 EEG . . . 126
3.6.1 EEG Materials and Methods . . . 126
3.6.2 EEG Results . . . 126
3.6.2.1 Time-locked to onset of stimuli (S1 and S2) . . . 127
3.6.2.2 Results time-locked to response . . . 143
3.6.2.3 Time-locked to feedback . . . 155
3.6.3 EEG Discussion . . . 169
3.7 Conclusion . . . 175

4 General Discussion **176**

4.1 Three Learning Models . . . 177
4.1.1 Learning through reinforcement . . . 177
4.1.2 Learning through sensory evidence integration . . . 178
4.1.3 Learning through probabilistic inference . . . 179
4.2 Neural Representations of Learning . . . 181
4.2.1 Context updating and choice commitment . . . 181
4.2.2 Feedback learning and action selection . . . 182

4.3 Future Research 187
4.4 Concluding Remarks 188

References **189**

1 General Introduction

1.1 Overview

Animals and humans live in complex environments, where a single choice can lead to many different outcomes. To achieve the best outcomes includes both learning from past experience and current environments as well as flexible use of the learned information. Imagine a monkey collects apples in an apple tree. As time goes the number of apples diminishes, and staying in the same tree may not be as profitable as before. Therefore, which apple tree to choose and how long to stay in the same tree become critical to the monkey's survival in this foraging scenario. Similarly, humans sometimes need to track continuous reward rates and anticipate structured changes in rewarding environments. Choosing the optimal driving path towards a cinema in London, for example, requires to consider the distance, traffic conditions, available parking slots among options. A decision maker needs first to prioritise the objectives (shortest distance versus the best traffic condition), recall from memory all possible paths, forecast the traffic condition in each path at the time of driving, and estimate the available parking slots near different entrances. Once a decision is made, the decision maker might change his mind and steer wheel towards another path for changes in the environment (e.g. busy traffic in the current path) or in needs (e.g. grab a burger before the movie starts). Does the decision maker integrate all sources of information to a single value and then compare and select the path with highest value? Or does he rank the importance of metrics and pick the path that maximises the most important ones? How does he weigh up costs and benefits to continue a course of action? How does he track the continuously changing information (e.g. traffic) and adjust the decision? These questions characterise the hierarchical nature of this type of decisions, for which the decision makers need to process multiple sources of information and decide whether to change

where they are or what they do. The thesis will introduce two projects that combine methods and theories to unveil the cognitive processes underlying such complex decisions.

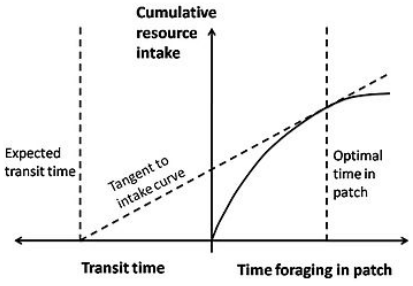


Figure 1: The canonical Marginal Value Theorem patch leaving problem determining patch leaving as a function of average reward rate and transit travel time (Charnov, 1976).

Considerable insights have probed into commonalities between complex decision faced by animals and humans. Adams et al., 2012 argued that similar neurobiological mechanisms with equivalent mathematical formulations may underlie dissimilar-seeming decision, emphasising the importance of thinking about which ecological functions are needed evolutionarily as well as what existing neural mechanism might be able to solve even modern seeming decision problems. Owing to the universal representation of complex decision faced by all organisms, it is worth studying the neural mechanisms of their behavioural flexibility. Within biology, Marginal Value Theorem (MVT; Charnov, 1976) has been crucial in describing reward based behaviours within continuous environments in animals, specifically predicting the optimal leaving time in a patch. MVT posits that a predator leaves the current patch when the marginal capture rate within the patch diminishes to the average capture rate of the overall environment. When taking into account transit time towards a new patch, the optimal leaving time is given by the transit time tangent to the resource intake curve within the current patch (**Figure 1**). As this class of model represents a universal patch leaving problem faced by all organisms, its neural implementation is the most obvious candidate

for using ecological ideas to understand neural mechanisms of behavioural flexibility. Note that MVT is distinct from value-based choices in general, which concern choices between discrete and simultaneously presented options.

In experimental and cognitive literature, this type of decision is often cast as exploration-exploitation or sometimes foraging. In a study that examined how people balanced the desire to accumulate rewards with a familiar option (exploitation) and to choose a less familiar one that might improve future decision (exploration) with a multi-arm bandit task, it was found that people picked the option with the highest payoffs most of the time but occasionally substituted a random action to explore other options (Daw et al., 2006). Another study discovered evidence integration for patch leaving on primates as predicted by MVT with a virtual foraging task (Hayden et al., 2011). Monkeys chose whether to stay with the current option which yielded diminishing rewards or to leave, pay some travel time, and re-enter the replenished patch by making corresponding saccadic movements. The study revealed that the primate brain computes a decision variable based on reward magnitudes associated with staying and re-entering the patch, as well as leaving cost (travel time), and the decision processes were likely to take place in the dorsal anterior cingulate cortex (dACC). As not every exploration brings useful information, it opens a question if an exploratory behaviour follows a directed strategy that people deliberately make to collect information or it is merely an outcome of decision noise. A study that examined the two strategies with a horizon task where participants make exploitation-exploration choices in contexts that differ in the number choices showed that people mixed both strategies to solve the dilemma but tended to be more information seeking when given more options to explore (Wilson et al., 2014). However, the mechanism of information seeking remains unknown. To tackle this problem, a study on rhesus macaque monkeys (*Macaca mulatta*)

identified a common mechanism of reward seeking and information seeking behaviour in a subpopulation of lateral habenula neurons. In other words, the midbrain dopamine system generates values to reward informative cues just as to primary rewards and generate positive and negative information prediction error for actions following unexpected and expected information cues, respectively (Bromberg-Martin and Hikosaka, 2011). Back to the above mentioned foraging example, the decision of staying or leaving the current apple tree depends on the reward level on the current tree and the average reward level in the surrounding environment, involving an exploitation-exploration trade-off. A study with human participants in a foraging scenario demonstrated distinct neural mechanisms in ventromedial prefrontal cortex (vmPFC) for comparative decision making and anterior cingulate cortex (ACC) for foraging (Kolling et al., 2012). Note that the study also introduced energy cost as a trade-off for foraging.

The thesis will introduce two studies that extend hierarchical decision making in two ways. The first study designed a novel *Gold Rush* task where participants decided when to quit a patch in a continuous time framework. Previous literature has characterised two types of decisions involved in foraging paradigm: the *within-state* decisions where people compare values of options and choose between them; the *state-change* decisions, where people leave the current state for a different set of options (Kolling and O'Reilly, 2018). As within-state decisions require computing and comparing reward rates of the options, a key feature is tracking the temporal changes of reward structures. State-change decisions involve comparing the expected values of the entire environment with the average reward rate within the current state based on past experiences, as well as taking into account the potential travelling cost. Learning reward structures requires integrating past experiences. It has been found that humans

adjust learning rates to volatility of the state and such adaptation is reflected by topographical representation of neurons in dorsal anterior cingulate cortex (dACC) (Meder et al., 2017). It has also been shown that the decision of leaving the current environment is largely guided by the estimation on environment turning 'good' or 'bad' according to reward trends and prediction errors in reinforcement learning, and the comparison process as well as the switch behaviour are reflected by dACC signals (Wittmann et al., 2016). Our *Gold Rush* project aims to extend these previous findings to a virtual foraging scenario that reflects the real life foraging decisions in a continuous time framework and with dynamic reward structures. Participants selected between two patches with different reward rates and travel distances, decided when to leave a patch given diminishing reward trend, and chose to re-enter the same patch or switch to a different patch. The reward rate was displayed in a continuous format and the reward structure changed gradually from trial to trial. We hope to examine the learning processes of estimating reward rates, tracking reward changes, incorporating reward rates and travel cost to make optimal decisions that maximise reward within a given time frame.

The second study investigated metacognition (i.e. evaluating the quality of perceptual evidence and one's own confidence, to inform how much it should be taken into account in switching decisions) with a perceptual judgement task. The research question was derived from a simple real life scenario. Suppose one made a maths mistake, for instance: one plus one should be two, but the person said 'three'. The person came to realise the mistake and figured out the correct answer. However, if the person took a step further and thought about why he or she made the simple mistake: have I not got enough maths practice recently? Was it due to a bad sleep last night? Or am I experiencing cognitive deficit, perhaps early Alzheimer? The hierarchical levels of inference from unitary feed-

back is metacognition. Purcell and Kiani, 2016 tackled the same question with a hierarchical random dot kinematograms task and demonstrated that people evaluated sources of error by integrating expected accuracy and outcomes over multiple trials and the results of a behaviour model suggested people switched decision strategy when accumulated evidence hit a threshold. With a collective low-level perceptual decision making task, Bahrami et al., 2010 showed people used perceptual confidence to weight decisions when two people had to negotiate a joint decision, and the communication of decision confidence improved collective decision unless the group members had very different visual sensitivities. Another study with rhesus monkeys also showed that they used perceptual confidence to decide whether to 'opt out' of a gamble, and the choice confidence was indicated by firing rates of neurons in lateral intraparietal cortex (LIP) (Kiani and Shadlen, 2009). We designed a gabor task where participants reported rotation of a selected gabor patch which they believed to be deterministic on reward. Reporting the wrong rotation or betting on the irrelevant patch led to no reward, and since participants only got one feedback (i.e. reward vs. no reward), it would be important to identify the right source of an error. Thus, participants took into account their perceptual uncertainty together with reward relevance when evaluating the decision. They needed to decide whether reward has switched side (as in normal foraging tasks) or whether they were just failing to get the reward due to poor perceptual judgement.

1.2 Neural Substrates

1.2.1 Neural mechanisms of patch leaving decisions

Patch leaving decisions can be regarded as a combination of two types of decisions as formulated by Kolling and O'Reilly, 2018: a *within-state* decision involving comparing and choosing between options, analogous to the traditional

economic decisions; a *state-change* decision involving leaving the current environment for a new environment or re-entering the replenished one. A cascade of research has shed light on brain areas related to certain decision processes involved in patch leaving decisions. One core brain area related to patch leaving decisions is the ventromedial prefrontal cortex (vmPFC), a part of the prefrontal cortex in the mammalian brain (**Figure 2**). Human imaging studies have identified association between blood oxygen level dependent (BOLD) signal in vmPFC and reward expectations and value comparisons (reviewed by Rushworth et al., 2012). In Daw et al., 2006's exploitation-exploration dilemma study with humans, neuroimaging results revealed correlation between vmPFC activity with the probability assigned by the RL model to the action chosen on a given trial, and the whole frontopolar cortex as well as intraparietal sulcus region were particularly active during exploratory decisions.

The anterior cingulate cortex (ACC), located in the frontal part of the cingulate cortex (**Figure 3**), has been suggested to signal conflicts of information as well as decision adjustments and thus is critical to cognitive control (reviewed by Botvinick et al., 2004). The area has also been believed to perform value comparison as the BOLD signal in ACC was found to be inversely correlated with value difference (reviewed by Rushworth et al., 2012).

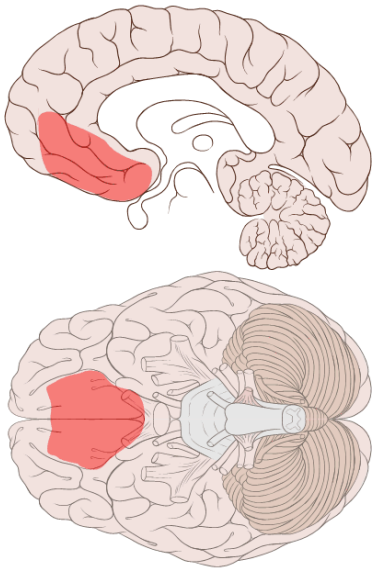


Figure 2: Ventral and medial views of the human brain with vmPFC highlighted (Lynch, n.d.).

Different nomenclatures have been proposed to break down the functional subregions of ACC (reviewed by Stevens et al., 2011). The more anterior part is referred to

as rostral or ventral ACC, and the part adjacent to the posterior cingulate cortex (PCC) as caudal or dorsal ACC. The dorsal ACC has been proposed to be split off as the middle cingulate cortex (MCC) (Vogt, 2005). The MCC can be further divided into anterior (aMCC) and posterior (pMCC) parts and the ACC can be divided into pregenual (pACC) and subgenual (sACC) regions (reviewed by Stevens et al., 2011). This thesis will adopt the pgACC-dACC nomenclature and be precise in the subregions of ACC if they are specified in the cited studies. However, the present discussion on ACC's functional role mainly serves as a general picture of the neural mechanisms of patch leaving decisions, whereas the exact functional anatomy of ACC is beyond the scope of the thesis, as the current research aims to examine the temporal dynamics of decision elements with electroencephalogram (see 'EEG Signals' below).

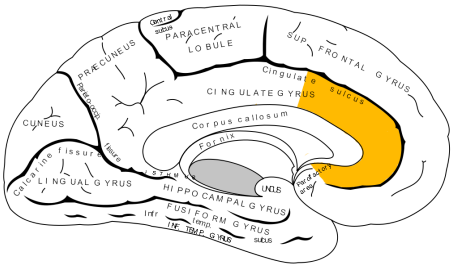


Figure 3: Anatomical structure of the left cerebral hemisphere of human brain, with ACC highlighted (Brodmann, n.d.).

In human foraging scenarios, ACC has been found to code values for switching to an alternative option as opposed to staying in the current environment (Kolling et al., 2012). In foraging tasks with primates, dorsal anterior cingulate cortex (dACC) has been found to encode value of leaving a depleting environment for a new one, as the the decision variable was represented by the firing rates of neurons in the dACC (Hayden et al., 2011). However, it was unclear whether the ACC encodes preference for the alternative option or resentment for the ongoing option (i.e. negative value of staying). In a foraging study with rodents, it was found that ACC activity represented two separate computations, one related to abandoning the current strategy

and the other related to the actual pursuit of the alternative strategy (Tervo et al., 2020). Another human functional imaging study provided evidence for dACC signals indexing adopting alternative choices and the actual taking of such choices (Kolling et al., 2014). Moreover, other evidence suggests ACC activity reflects time interval discrimination as well as dopaminergic projections, indicating the area is able to integrate rate-related information about states and actions, which is a crucial step in foraging decisions (reviewed by Kolling and O'Reilly, 2018).

Although vmPFC and ACC are adjacent in the frontal lobe, Kolling et al., 2012 discovered functional dissociation between the two in their foraging task: vmPFC encodes values of available options, whereas ACC signals search value and the average value of the environment, both areas contributing significantly to foraging decisions. Shenhav et al., 2016 extended the finding by conducting a similar foraging task with a larger group of human participants. They contended that dACC is more likely to encode choice difficulty alone (within-state decision) rather than in combination with foraging value (state-change decision), and that the vmPFC activity, inverse to dACC, serves a common function for both economic and foraging choices.

Note that both vmPFC and ACC activities correlate with option values or reward expectations, regardless of whether the choice is actually made by the participants. This implies that after participants commit to a choice, the signal must be carried on and translated into a behaviour. Signals in premotor cortex, parietal cortex (LIP in monkeys), frontal eye fields (FEF) have been found to index action selection and response execution (reviewed by Platt and Plassmann, 2014). Specific regions depend on task scenario and which response the subject is required to make (i.e. button press or saccadic movements). Outcome valuation is another important step in foraging so that the agent can evaluate and adjust

their decision strategy to changes in the environment. Orbitofrontal cortex (OFC) has been suggested to be a critical brain region in outcome evaluation, and its signals correlate with not only primary rewards (i.e. pleasant stimuli such as juice and music) but also secondary rewards (e.g. money) (reviewed by Platt and Plassmann, 2014). The classical three stages model (option valuation - action execution - outcome valuation) proposed in traditional economic decision making may also explain key building blocks of foraging decisions, although it is important to note the latter involve interleaved within-state and state-change considerations.

1.2.2 Neural mechanisms of perceptual decision making

Brain mechanisms of perceptual decision making are associated with sensory evidence integration. Early evidence emerged from electrophysiological experiments on primates and rodents. In several vibrotactile frequency (VTF) discrimination tasks with primates, activities in the primary somatosensory cortex (S1), the second somatosensory cortex and the dorsolateral prefrontal cortex (dlPFC) were found to closely resemble a decision variable. In random dot motion (RDM) direction discrimination tasks, neurons in the middle temporal area (MT/V5), LIP, superior colliculus (SC), FEF and dlPFC were found to code motion direction and preparation for response (reviewed by Gold and Shadlen, 2007). Hence, although specific brain areas depend on stimuli and response type used in the tasks, a general consensus can be drawn on motor and prefrontal areas involved in evidence integration. It is hard for human imaging studies to track neural correlates of perceptual evidence integration with relatively high temporal dynamics due to fMRI's low temporal resolution. Despite the limitations, an fMRI study using a simple categorisation task where human participants decided whether a presented image was a face or a house, showed that left dorsolateral prefrontal cortex covaried with face- and house-selective

regions in ventral temporal cortex and predicted behavioural performance in the task (Heekeren et al., 2004).

In addition to evidence integration, another key aspect of our metacognition project is the decision confidence. Decision confidence has usually been formulated as a probability that captures how certain the subject is about the response. Previous studies have established a connection between the firing rates of LIP in primates brain and probabilistic judgements. In a probabilistic classification task, neurons in monkey LIP showed spatially selective persistent activity predicting eye movements towards the chosen option, as the firing rates of LIP were proportional to the log-likelihood ratio of the chosen option (Yang and Shadlen, 2007). Another motion direction discrimination task with monkeys demonstrated that the weighting of prior probability, mirroring bias signals found in speed-accuracy setting with human subjects, was reflected by the firing rates of neurons in the LIP of rhesus monkeys (Hanks et al., 2011). A neuroimaging study using RDM task revealed perigenual anterior cingulate cortex (pgACC) in human tracked expected performance, an aggregated signature of decision confidence, and is linked to between-subject variation in explicit confidence estimates (Bang and Fleming, 2018).

1.2.3 Common and distinct neural correlates of the two studies

Depending on the task scenarios, decisions have been categorised into value-based and perceptual decision making (Gold and Shadlen, 2007; Glimcher and Fehr, 2014). Value-based decision making is analogous to classic economic decisions including computing and comparing each option's value based on corresponding reward magnitudes and probabilities, whereas perceptual decision making mainly involves integrating perceptual evidence and making judgement given a fixed threshold compared against the accumulated evidence. Although we do not aim to simply classify our studies into these categories due to the more

complicated, hierarchical nature of our tasks, as the first level of inference in each project more or less mirrors value-based or perceptual decision making (i.e. patch choice decision in *Gold Rush* and rotation judgement in metacognition), commonalities and differences in neural substrates of the two decision types are still of our interests. According to the literature, regions like LIP and ACC have been found to encode values in both decision categories and regions like FEF and dlPFC are suggested to represent response preparation. However, conclusions on common neural mechanisms have faced serious challenges. It has been argued whether the brain computes a 'common currency' as decision variables across different evidence types and task paradigms, or if the individual components such as reward magnitude and probability are represented separately (Gold and Shadlen, 2007). Moreover, it is unclear whether neurophysiological findings from the animal brain can be mapped to the exactly same areas in the human brain. For example, the human ACC is more analogous to dlPFC than ACC in the rat's brain (Kolling and O'Reilly, 2018).

Despite using seemingly dissimilar task paradigms, our two projects may help establish common neural mechanisms underlying foraging and perceptual judgement decisions. Note that both projects involve hierarchical decision elements: the within-state versus state-change decisions in the *Gold Rush* task, and the visual perception versus reward contingency inference in the metacognition task. The decision stages are not separate but interleaved and may change according to the highly dynamic task structures. Thus, we do not aim to pinpoint the exact functional localisation of related brain areas, but to understand neural signatures underlying specific events (e.g. stimuli processing, value computation and comparison, action execution and outcome evaluation) and to read out the temporal dynamics relating neural signals to distinct decision elements. Built upon pioneering work in the field, we will focus on relating signals of the frontal

pole and the parietal cortex to relevant task features and behavioural characteristics with Electroencephalography (EEG), an electrophysiological monitoring method with excellent temporal resolution that we will next introduce in detail.

1.3 EEG Signals

A cascade of neuroimaging studies with human participants have identified the key brain areas related to decision making with an emphasis on the prefrontal and parietal areas. However, since the fMRI technique averages BOLD signals from brain areas comprised of many millions of neurons within a single voxel usually across a few seconds, it is hard to capture time-to-time variations related to key decision elements. As our *Gold Rush* task involves instantaneous reward rate changes at one-second level and the metacognition task examines anticipatory attention and working memory allocation both occurring on a millisecond scale, it is very important to probe into transient neural activities before and after a decision event. Therefore, we decide to use EEG for its excellent temporal resolution to find out the neural signature underlying behaviour and task characteristics in our studies.

1.3.1 Event-related potentials

An EEG component P300 (P3) has been demonstrated to reflect motivation guided information processing, which serves as an important antecedent for response generation (Nieuwenhuis et al., 2005). P3 is a large positive component across the central-parietal cortex occurring around 300-400 ms after stimuli presentation (S. Sutton et al., 1965). It has been suggested to co-vary with response accuracy and its latency is correlated with reaction time. Specifically, stimuli eliciting a larger P3 have a higher chance of being accurately discriminated and P3 latency is correlated with trial-to-trial variability of reaction times.

However, the latter finding mainly comes from two-alternative forced choice tasks and there is no consensus in the literature whether P3 should be stimulus or response-locked. P3 might be deemed as selective attention to task relevant decision processes, a mediator between detection of sensory stimuli and selective response facilitation (reviewed by Nieuwenhuis et al., 2005). A study with an auditory stimuli discrimination task revealed that P3 matched a gradual evidence accumulation process up to a fixed threshold level and the signal build-up precisely captured the reaction time as well as accuracy of the perceptual decision making task, and that P3 showed equivalent dynamics across auditory and visual modalities (Twomey et al., 2015). Although previous evidence on P3 was predominately obtained from perceptual decision making tasks, we hypothesise that if P3 reflects in general motivation-guided stimuli processing and subsequent response preparation, we would observe it as well as its correlation with key task and behavioural characteristics in our two studies.

Another key component we are interested in is the one related to error monitoring. Error-related negativity (ERN), a negative deflection in EEG occurring around the frontal cortex (mostly at the electrode FCZ) after people commit an error in reaction-time studies, has been found to be relevant to error processing (reviewed by J. R. Wessel, 2012). ERN is believed to be generated by ACC and mainly elicited by detection of error itself as a mechanism to prevent future errors (reviewed by Yeung et al., 2004). Specifically, it has been argued that ERN occurs when a negative reinforcement learning signal is carried forward to ACC through the mesencephalic dopamine system (Holroyd and Coles, 2002). This theory may explain previous findings from neuroimaging studies on the role of ACC in performance evaluation and decision adjustments. Alternatively, Yeung et al., 2004 proposed a conflict monitoring theory, arguing that ERN reflects continued processing of response conflict that may help detect errors during the

post-response period. We do not seek to establish ERN in our studies, as evidence about ERN was predominately obtained from reaction-time tasks where participants were explicitly informed of the correct response and immediately realised an error, whereas due to the complexity of our tasks, participants would not realise an error until a feedback was given.

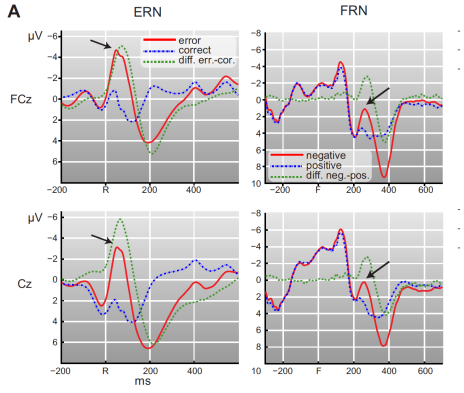


Figure 4: Event-related potentials related to response-locked errors (ERN) and feedback-locked negative feedback (FRN) (Ullsperger et al., 2014).

Feedback-related negativity (FRN) is another EEG component related to performance evaluation but specifically to feedback processing. FRN is a negative deflection over the frontocentral scalp occurring 200-300 ms after feedback display and is usually greater on negative than on positive feedback (reviewed by Ullsperger et al., 2014). Some researchers argue that FRN refers to a positive deflection but is less positive following negative compared to positive feedback (Cohen, Wilmes et al., 2011), or that FRN is obtained by subtracting signals following positive feedback from signals following negative feedback or the negative waveform generated by negative feedback alone (Nitta et al., 2017). Generally, FRN amplitude is higher when the difference between outcome and the expectation is larger and scales with the reward prediction error (Ullsperger et al., 2014). Therefore, FRN may represent a 'surprise' signal or the need for behavioural adjustments. As with ERN, ample evidence suggests that FRN is also generated by ACC (Ullsperger et al., 2014; Nitta et al., 2017).

Both ERN and FRN are related to performance evaluation, but ERN is usu-

ally response-locked, and it was mostly found in speeded time-reaction tasks where a correct response was unequivocally given, whereas FRN is tied to the feedback. ERN has also been suggested to reflect post-response conflict and FRN has been shown to reflect reward prediction error. Moreover, previous studies demonstrated that FRN decreases and ERN increases as learning progresses (Ullsperger et al., 2014). This finding can be explained by unique cognitive processes conveyed by the components: as learning progresses, the importance of feedback declines, whereas response error still elicits an alert. Although due to the hierarchical nature of our tasks, it was difficult for participants to realise a strategy error without explicit feedback, we focus on establishing the role of FRN in feedback processing and how the component reflects reward expectation, choice evaluation and decision adjustments.

1.3.2 Brain oscillations

In addition to event-related potentials, changes in brain oscillations also reflect key cognitive processes. One of our foci is on alpha-band oscillations, a dominant type of oscillations in the human brain with a mean frequency of about 10 Hz. Alpha-band oscillations has been believed to reflect attentional control of the knowledge system, including anticipatory and temporal attention (reviewed by Klimesch, 2012). Research on cognitive processes of alpha-band oscillations have been extensively conducted with oddball tasks, in which participants were asked to press a corresponding button as soon as a target or a non-target appeared (for example see Klimesch et al., 1998). When a cue is given prior to the target onset, it affects both pre- and post-stimulus activation levels, reflecting an association between alpha-band oscillations and anticipatory attention (Klimesch et al., 1992). Alpha-band was also found to decrease when anticipating the onset of a target and increase when anticipating a distractor on the contralateral hemifield (Worden et al., 2000). The alpha desynchronisation indexing visuospatial

attention bias effect was found in another Posner's cue-target paradigm in which an auditory cue was used, and the alpha lateralisation predicted performance (Thut et al., 2012). Another study with continuously presented bilateral flicker-stimulus showed enhanced alpha power in the contralateral parieto-occipital hemisphere relative to the ignored (distractor) side, suggesting increased alpha as a gating mechanism the brain exerts to actively suppress task-irrelevant stimuli (Kelly et al., 2006). In a random-dot kinematograms task, alpha power was found to be higher contralateral to the unattended side, and the alpha lateralisation correlated with visual discrimination performance (Händel et al., 2011). Additionally, it was revealed that not only increased alpha oscillations prior to a distractor predicted performance, but also the alpha could be modulated according to the strength of the distractor (Bonnefond and Jensen, 2012). In our metacognition task, we introduced a working memory delay period between the first and second set of stimuli presentation, and presented a warning signal 0.5 s before each stimulus onset. We expect to see changes in alpha-band oscillations during the working memory delay and the post-warning signal interval, which will reflect key attentional selection induced by task relevant stimuli.

Movement-related beta-frequency (15-29 Hz) changes have also been characterised in perceptual decision making tasks. Beta-band activity has been demonstrated to correlate with both movement-initiation and cancellation in a stop-signal task (J. Wessel, 2020). Participants were asked to respond as fast and accurately as possible to a left- or rightward arrow using their left or right index finger. On one-third of trials a stop-signal occurred at a short delay after the go-signal, in which case participants had to cancel the response. The study found decline in β-bursting at bilateral sensorimotor sites during movement initiation and increase in β-bursting at fronto-central sites during movement cancellation. In particular, the amount of β-bursting indicated successful or un-

successful movement cancellation. It suggests a connection between β-bursting and movement inhibition from motor cortex which is suppressed during movement initiation. Another study investigated decision dynamics and post-error adaptation conveyed by beta power using a flanker task, where participants had to report the orientation of a target arrow surrounded by either congruent (same orientation as target) or incongruent (opposite orientation to target) arrows (Fischer et al., 2018). The study found beta power lateralisation (BPL) to be a sensitive measure of response-threshold modulations. Specifically, the peak amplitudes of BPL was in accordance with increased decision boundary and decreased evidence accumulation rate on post-error trials.

Built upon the above-mentioned evidence, we sought to establish cognitive processes conveyed by ERP and brain oscillations in our experiments. In particular, we would focus on two levels of of cognition: basic and hierarchical components. Basic components refer to well-established EEG signals elicited by primary processing of visual stimuli, feedback or behavioural response, including lateralised visual evoked potentials, lateralised readiness potentials, P300, FRN, alpha desynchronization and beta power changes. We would further probe into neural correlates of task specific hierarchical cognitive processes, such as switch/stay decisions, changes in reward structures, built-up evidence across trials and choice confidence. To examine these advanced cognitive components, we would utilise computational models that capture task-related behavioural features, which will be introduced in the next section.

1.4 Computational Modelling

In the field of decision making, a tremendous amount of experiments have been conducted to understand cognitive processes and behaviour adaptations in various task scenarios and decision contexts. However, all observable behaviours

are orchestrated by a black box – the human brain. The exact mechanisms of human brain are not directly observable, even with state-of-art neuroimaging techniques, but can be approximated by models. Computational modelling is therefore increasingly used to probe into latent causes and mechanisms of behaviour that we cannot observe directly, as well as to get a simplified form of cognitive processes. Behavioural models aim to capture random and systematical variations in behaviour characteristics, including choice, accuracy and reaction time. Model parameters are usually fit to behavioural data or arbitrarily determined by task features. To build a model that best describes cognitive process, it is critical to first understand features of data and propose reasonable underlying mechanisms. Is it a probabilistic inference process? Is it a binary classification? Does it involve evidence accumulation within or across trials? How many variables are there to be considered and how complicated (i.e. number of parameters) the model ought to be? After building the basic model architecture, the next steps include fitting the model to data and interpreting the model parameters after it has been fit. Finally, when there are multiple alternative implementations of decision processes or many model variants are able to provide reasonable accounts, model selection is usually performed. Model selection methods that compare models' goodness of fit while penalising model complexity such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are often used to find out the optimal model that balances goodness of fit and model parsimony. In addition to behavioural data, there is a growing trend in linking neural evidence towards computational models to further understand the fundamental biological implementations of the decision process. With the development of neural recordings, it is increasingly possible to fit model parameters to neural data to identify core neural computations underpinning temporal dynamics, firing patterns and other traits of neural circuits

(reviewed by O'Connell et al., 2018). Here we discuss three specific models relevant to our studies: reinforcement learning, drift diffusion model and Bayesian Learner.

1.4.1 Reinforcement Learning

A key problem our tasks tackle is learning through experience and feedback within or built up across trials to adjust decision strategy and maximise reward gained throughout the tasks. A simple form of trial-and-error learning is categorised by reinforcement learning (RL) in computer science (R. S. Sutton and Barto, 1998). It posits the decision agent updates the decision value V for each stimulus s on trial $k+1$ with the reward obtained on last trial k. A learning rate parameter α, ranging from 0 to 1, determines the size of updating. The basic form of RL is shown below:

$$V_{k+1}(s_k) = (1-\alpha)V_k(s_k) + \alpha r_k \tag{1}$$

When a reward is absent, r_k becomes negative and $V(s_k)$ is reduced. With a larger learning rate the decision value updates more on the current feedback, and with a small learning rate it relies more on old beliefs. It has been argued that learning rate can be seen as a parameter controlling steepness of decay, with a larger learning rate resulting in reduced influence of historic experiences on the current step (Daw and Tobler, 2014). More advanced forms of RL include the idea of state (i.e. making predictions on all possible future reward in the presence of the stimulus) and whole decision strategy rather than a single action (i.e. choosing an action given a pool of all actions) as well as interactions between states and strategies. As we only plan to fit the basic form to our task, this chapter will not discuss these advanced forms for the sake of succinctness.

The biological implementations of RL are reflected by the link between pre-

diction error and dopamine, a neurotransmitter responsible for movement control, motivation and reward. The projections of dopamine neurons, largely located in the midbrain, have been found to broadcast a scalar signal similar to net prediction error as posited by RL (reviewed by Daw and Tobler, 2014). Specifically, the phasic firing rate of dopamine neurons elicited by excitatory or inhibitory events resonate with a reward prediction error. However, it was questioned whether the responses reflect the prediction error or the reward itself. An experiment with varying probabilities of reward following a predictive cue showed the firing rates of dopamine neurons in the monkey brain increased with reward expectancy, whereas activity evoked by the reward itself decreased. In other words, DA activity corresponded with reward uncertainty rather than the reward magnitude itself. (Fiorillo et al., 2003).

Human fMRI studies also revealed association between BOLD signals in the striatum with prediction errors, mirroring key findings on dopamine neurons in primates (reviewed by Daw and Tobler, 2014). In a sequential decision making task, participants were asked to avoid starvation in a virtual foraging task by making a number of consecutive foraging choices with probability outcomes. The study reported a mixed strategy combining heuristics and an optimal policy derived from a Markov decision process (a mathematical framework for partially random consecutive decisions, the form that many RL algorithms are stated in) best captured participants' behaviour, and BOLD signals in medial prefrontal cortex correlated with model predictions and choice uncertainties (Korn and Bach, 2003).

We plan to implement RL to our data in *Gold Rush* as we hypothesise people adapt their beliefs of reward structures using prediction errors obtained from trial to trial. However, there are a couple of things that differentiate our task from a simple RL. First of all, choices in *Gold Rush* are sequential, involving

two stages: patch choice and patch leaving. Neither of the choices are followed by a single reward vs. no-reward consequence. Instead, participants experience reward changes and adjust their estimation on the reward rate every second within a trial. To excel in the task, people need to project future reward gains against time spent, compute the average reward rate and compare with the environment. This complicated computing process can be modelled by several variations of RL such as the temporal difference model that projects future reward gains in timeslices of trials (Daw and Tobler, 2014) or models with rectified value functions. Another key question is whether people learn and update the reward rate as a whole or broken down into components (scaling factor and decay rate). We plan to tackle the question by building two types of RL and perform model selection. In addition, previous neural signatures of RL were predominately investigated by fMRI, whereas BOLD signal is a slow metabolic change and an ambiguous metric for dopamine changes. We aim to extend the findings with EEG to unveil the dynamic value updates and policy changes as predicted by RL.

1.4.2 Drift diffusion model

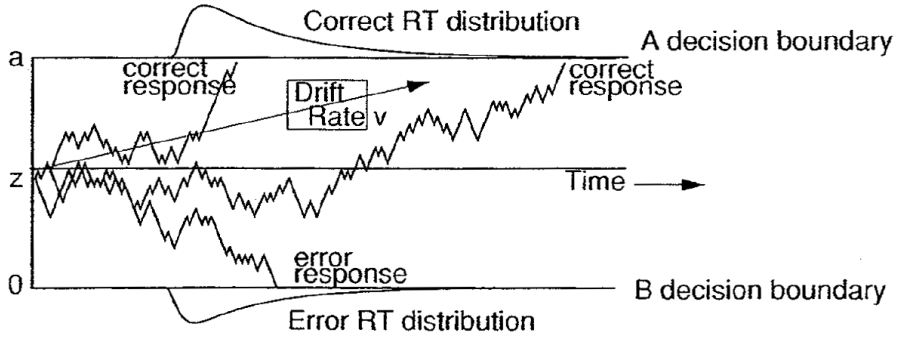


Figure 5: The drift diffusion model with three simulated decision paths categorised by drift rate v, boundary a and start point z. (Ratcliff and McKoon, 2008).

As both of our tasks involve a two-alternative forced choice (2AFC) stage (i.e. choosing between two patches and responding on left or right rotation), drift diffusion model (DDM) has come to our attention. DDM describes the evidence accumulation process in 2AFC tasks until the evidence hits a boundary favouring one of the choice options. A graphic illustration of DDM is presented by **Figure 5**, assuming the decision agent chooses between two options: the correct option **A** and the error option **B** (although in some studies there is no correct or error option, here we keep the notation for illustrative purpose). The evidence starts accumulating at the initial start point **z**. The speed of accumulation is captured by drift rate **v** and its sign determines the direction favoured by evidence. The boundaries **a** and **0** determines the threshold of each option, and once the accumulated evidence hits one of the boundaries, the agent is committed to that choice and the decision process ends. DDM captures three key behavioural traits of 2AFC decisions: accuracy, response time and response time distributions (Ratcliff and McKoon, 2008). Usually a larger drift rate shortens response time and boosts accuracy. Decreased decision boundaries cause speed-accuracy trade-offs, resulting in faster errors. Bias in start point indicates if the agent leans towards one of the options before any evidence is observed, and is usually modelled across trials. In addition to the three basic parameters, another parameter called non-decision time is used to capture any duration not directly relevant to the decision process, such as visual stimuli processing and action execution.

In Purcell and Kiani, 2016's hierarchical random dot kinematograms task, they modelled behaviour with a computational framework combining DDM and probabilistic inference. Specifically, the DDM part approximates the evidence accumulation process related to direction of dots movement, and the probabilistic inference part modelled switch/stay decision based on perceptual confid-

ence and switch evidence built up across trials. In Fischer et al., 2018's flanker task, DDM and beta power lateralisation (BPL) gave highly similar predictions on behaviour. On post-error trials, responses became slower and more accurate, in accordance with increased decision boundary simulated by DDM and increased peak amplitudes yielded by BPL. Moreover, flanker-driven deflection of BPL was reduced, so as the behaviour modelled by DDM, suggesting suppression of distractor information following errors. Overall, DDM and BPL together provided a reasonable account of post-error adjustments, increased decision thresholds and reduced attentional engagement in distractors, capturing post-error behavioural adaptations.

We consider DDM for the rotation judgement in the metacognition task. However, the model may not be a good candidate for the hierarchical structure of the task: participants were asked to select the target and report its rotation with a single response; in other words, it is a four-alternative rather than a two-alternative choice task, in which case accuracy and response time on the rotation might be influenced by the additional cognitive process of 'reward contingency'. Nevertheless, it is possible that rotation and switch/stay decisions are dissociable, allowing DDM to capture the perceptual decision.

1.4.3 Bayesian Learner

Both *Gold Rush* and metacognition tasks involve learning the reward structures and tracking the changes of the environment. The reversal of key reward parameters took place every 13-23 trials in *Gold Rush* and the reversal of reward contingency happened every 5-15 trials in metacognition. Thus, changes of the environment were in general predictable but not exactly predictable. The ability to estimate the probability of change based on observations across trials was critical in both tasks and the estimating process could be modelled by a Bayesian Learner. A Bayesian Learner is a learning model derived from the

Bayes' rule (e.g. O'Reilly, 2013):

$$p(\theta_i|x_{1:i}) \propto p(x_i|\theta_i)p(\theta_i) \tag{2}$$

...where $p(\theta_i|x_{1:i})$ represents the parameters of the environment given observations up to trial i, and $p(x_i|\theta_i)$ is the likelihood function, denoting the likelihood of obtaining the observation given the estimated environment, and $p(\theta_i)$ represents the prior probability of the environment regardless of observations. The Bayesian Learner can be used to update the observer's beliefs on the environment given observations. Changes in the environment will be reflected by new observations and eventually modify the observer's estimation on the parameters of the environment.

Another key question comes as how much historical information should be considered given a changing environment. If the environment changes fast, only recent experience is relevant to the current structure, and if the environment is relatively stable, the agent should take into account more historic information. It has been proposed that, just as reinforcement learning, the update of belief in Bayesian Learner should be controlled by a learning rate, the size of which should depend on volatility of the environment. A study examined humans' ability to adapt learning rates using a one-armed bandit task where participants chose between two coloured squares (Behrens et al., 2007). One of the squares yielded a certain chance of reward and the other square nothing. Two levels of volatility were applied: in a stable environment, probability of yield reward of each option was fixed; in a volatile environment, reward contingency reversed every 30 or 40 trials. The study demonstrated humans modulated the learning rate according to the volatility, as captured by a Bayesian Learner, and the learning rates correlated with BOLD signals of the ACC. This finding can also be interpreted by the role of ACC in exploration discussed earlier, that ACC governs forgetting

of old beliefs and exploring to learn the new rules. Neurochemical evidence emerged from pupilometry studies suggesting that noradrenaline levels represent the prior probability of change and the phasic noradrenaline may represent the posterior estimation on a change occurring (reviewed by O'Reilly, 2013).

In some more complicated human decision making studies, variations of Bayesian Learner have been made and compared against or combined with other learning models. For example, a study with a single spatial prediction task contrasted between RL and a continuous dynamic forward modelling, both combined with Bayesian estimation on possible 'landing points' of a robot that participants were asked to predict (O'Reilly et al., 2013). The results revealed that participants shifted between the strategies, and neural correlates underlying the two computational systems were dissociable, with a region in parietal cortex mediating the integration of the two models. In Purcell and Kiani, 2016's hierarchical RDM task, a computational framework combining DDM and probabilistic inference was built to capture the cognitive processes. The net direction of motion (left or right) was modelled by an expected accuracy parameter from DDM, taking into account both stimuli strength and duration. The expected accuracy combined with feedback were then integrated across trials in units of log-likelihood ratio to guide the switch strategy. We plan to fit RL and Bayesian Learner as well as a mixed model for our *Gold Rush* task, as we believe people update their beliefs on the reward structures across trials, but whether the updating follows a single trial-and-error process as predicted by RL or a probabilistic inference process as described by Bayesian Learner or a mixture of both is unknown. For the metacognition task, we hypothesise that the rotation judgement may be captured by DDM, whereas switch evidence accumulated across trials can be approximated by Bayesian estimation, with the possibility that the two levels of decision interact with each other.

1.5 Current Research

The thesis presents two projects studying hierarchical decision making with dynamic reward structures. The first project tests how participants in fact track continuous reward environments and adapt decision strategies in a virtual foraging scenario *Gold Rush*. More specifically, it examines whether people track continuous reward rates and build models that would allow them to anticipate structured changes very quickly in order to optimise patch leaving behaviour. The computational modelling integrates work showing how two brain systems that are typically dissociated with reward learning and motor control, could be dissociated in terms of two forms of computations: a reinforcement learning model of the environment obtained through experience with discrete events, and a Bayesian Learner updating probabilistic representation of the task structure via observations. Because of the need for a more continuous/highly resolved neural measurement, EEG is used to investigate patch choice and patch leaving foraging behaviour. Specifically, EEG signals are measured during the ecological foraging task with changing dynamics of rewards to examine how people learn to anticipate change in reward environments. Furthermore, different learning algorithms mentioned above are fitted to find out the best one to describe such anticipation and investigate the neural sequence of signals related to such construction of expectation and its violation. The study aims to tackle 1) how people adjust their estimates in response to the changing environment in ecological foraging, 2) how they invigorated extended behaviour within such environments and 3) how learning from complex and extended reward environments is retrieved when participants need to choose between patches.

The second project tackles how hierarchical decision processes are formulated and how people track changes in the environment with a conventional perceptual decision making task, where participants integrate evidence to update the two

aspects of their decision strategy (i.e. rotation and reward contingency) based on their own performance in the perceptual judgement as well as their experience across trials. A drift diffusion model is fitted to the perceptual part and a Bayesian Learner to the probabilistic representation of betting on the correct response. In addition, EEG is used to answer three questions about neural correlates: 1) if medial frontal ramping signals track the switching/reversal process specifically, 2) whether lateralization of alpha power between left and right visual cortex could be used as a measure of selective attention, whether this is predicted by frontal signals such as ramping signals and whether it predicts behavioural performance on the first level task (i.e. rotation) and 3) if response-elicited signals differentiate between switch/stay decisions and account for variations in reaction time.

The two studies will bridge research on conventional perceptual decisions and value-based decisions in complex environments. Both types of decisions require people to track and integrate evidence in a hierarchical manner, as both require within trial and between trial dynamics to be tracked. Another way in which the two studies complement each other is that while in the first study participants would always receive reward at a certain time in a patch and knew how to attribute reward to an environment, the second study tackles exactly this attribution problem when it is unclear whether an environment has gotten worse or performance. As reward outcomes are almost always performance dependent in the real world, being able to dissociate changes of environment from performance is essential for accurate learning. The studies aim to understand how people process structured changes in continuous reward environments, update their internal models to adapt to the changing environment in hierarchical decision making tasks, integrate different aspects of the environments (e.g. visual stimuli and feedback) to make a response as well as key neural signatures of

these decision elements.

2 Study 1: Foraging

Research interests in the field of decision making have spanned into foraging, which describes animals' choices in searching for food and has been extensively studied in ecology. In a foraging scenario, animals choose between exploiting the current environment and exploring a potentially richer environment. Instead of a single exploitation-exploration trade-off, animals continuously face the decision as energy level in the current environment declines. The optimal 'leaving to explore' time has been successfully predicted by marginal value theorem (MVT) across different organisms, that is, a predator should leave when the average reward rate of the current environment matches the average reward rate of the overall environment. Here, we extend the findings to human participants with a virtual foraging task, in which participants need to track reward changes, pick a patch to exploit and decide when to leave a patch. We show that people are able to pick a patch that maximises reward rate, leave a patch as predicted by MVT, track changes and integrate reward history into their decision. Model comparison shows that people infer reward rates for patch selection as predicted by a Bayesian Learner and learn from past experience across trials to decide when to leave as categorised by a simple Reinforcement Learning model. We also report neural representations of both basic stimuli processing and action execution, as well as high level cognition of integrating learned information into long term estimate and decision processes.

2.1 Introduction

Animals and humans seem to face a common foraging scenario: when resources are not uniformly distributed in the environment, the strategy for balancing the benefits of exploiting one patch and leaving for a richer one becomes essential for survival. For animals, the main goal is to maximise energy intake with minimum energy cost, whereas for humans, money is usually used as a vital commodity, but the underlying cost-benefit trade-off remains the same. From evolutionary perspectives, it has been argued that neurobiological mechanisms across organisms in the same paradigm may be captured by similar mathematical formulations (Adams et al., 2012). Marginal Value Theorem (MVT), which posits that the optimal leaving time in a patch is determined by the average energy intake rate within a patch and the average energy intake rate of the overall environment (Charnov, 1976), has successfully described widely observed patch leaving behaviour across organisms in ecology, from larva to humming birds (Stephens and Krebs, 1987). Previous studies in behavioural neuroscience on humans and primates provided further evidence for the optimal patch selection and patch leaving behaviour categorised by MVT, and have suggested that vmPFC and dACC might be the essential brain areas for foraging (Kolling et al., 2012; Hayden et al., 2011; Daw et al., 2006; Wilson et al., 2014; Bromberg-Martin and Hikosaka, 2011). However, previous research mainly studied foraging behaviour with relatively static reward structures (e.g. gambles with fixed values during an encounter), whereas in real life the resources may change dynamically due to varied ecological conditions of the environment. Furthermore, neural correlates of human participants were mainly provided by neuroimaging techniques, which averaged across several seconds of brain activities and may lose key insights in categorising fast changing brain signals reflecting quick learning and decision processes as well as their adaptions to a dynamic environment.

The current research aims to examine foraging behaviour on human participants with *Gold Rush*, a virtual foraging task with dynamic reward structures. We hypothesise that participants pick and leave a patch based on their estimates of the average reward rate of the patch and the overall environment as predicted by MVT, and they will be able to adjust their estimates according to changes. In addition, we build a series of learning models combining general linear models with reinforcement learning and Bayesian learner to describe the learning processes that categorise how participants figure out the reward rates, track changes and make comparisons.

Finally, we seek to establish neural representations of key decision elements with EEG decoding analyses. Previous research has shown it is possible to extract distinct, potentially overlapping neural representations that can predict performance on a trial-by-trial basis (Hubbard et al., 2019). A handful of literature used Hidden-Markov Models (HMM) to decode EEG signals and found neural signatures of the cascade of information processing. For example, Borst and Anderson, 2015 used hidden semi-Markov models to identify distinct processing stages of information processing in an associative recognition task. Another study with HMMs revealed task-dependent states representative of whole brain dynamic networks (Quinn et al., 2018). Vidaurre et al., 2019 decoded both early sensory areas and higher order areas involving information integration and decision making processes in a perceptual template-matching task with a novel version of HMM that allowed cross-trial temporal variability. An advantage of an HMM is that it attains temporal sensitivity that is not too small (e.g. in miliseconds) or too large (e.g. in tens of miliseconds), thus reducing dimensionality from miliseconds to several states while keeping high temporal resolution to pinpoint when and how information processing takes place. We therefore apply an HMM to probe the functional relevance of goal-related neural

representations with high temporal resolution. Specific hypotheses are set out in the corresponding results section.

We choose the decoding approach over classic ERP analyses because the latter aim at finding differences between conditions and are most often applied to tasks with distinct conditions, whereas our task requires processing of rich and hierarchical information simultaneously. It is of our interest to pin down when and what information is processed over and above observations. Note that we do not seek to align each state of HMM with a specific cognitive stage, as the causal interpretation of decoding models is ambiguous (Weichwald et al., 2015), and the cognitive processes may be interleaved in our task. The main aim of fitting HMM is to extract temporal dynamics of rich, potentially overlapping information processing involved in our task, from both low sensory processing or action execution stages to abstract, context-dependent cognitive processes.

2.2 Materials and Methods

2.2.1 Participants and Procedure

Thirty healthy participants (aged 18-40 years, right-handers) took part in the study. Participants were recruited from a student population at University of Oxford and all gave written consent using a form approved by the ethics committee. All participants finished two sessions on two separate days, leading to a dataset of 60 sessions in total. Each session took between 2.5 and 3.5 hours including the EEG setup. Participants received reimbursement for their time and performance upon completion of the experiment.

2.2.2 Behavioural Task

The behavioural task was programmed and delivered using Psychtoolbox in MATLAB. Visual stimuli were presented on a 24-inch (61.1-cm) monitor with a refresh rate of 60Hz.

Participants were instructed to play a game called Gold Rush with the goal of collecting as much reward as possible within limited time. The concept of time was converted to water supply in the game. Participants were given in total 60 bottles of water (presented by a water bar which indicates the water level in the current bottle plus a number printed on top of it showing how many unopened bottles are left) and were told that the game ended when water was depleted. Travelling to and staying in a patch consumed water and each bottle sustained for 1 minute so that participants experienced time cost in terms of water bar shrinking and the number of bottles reducing. During the task, participants chose between two patches, travelled to the patch and decided when to leave the patch, based on relevant reward structures and water consumption (i.e. travel time and patch digging time, linearly mapped to water consumption). Each trial began with two patches and their corresponding travelling distances

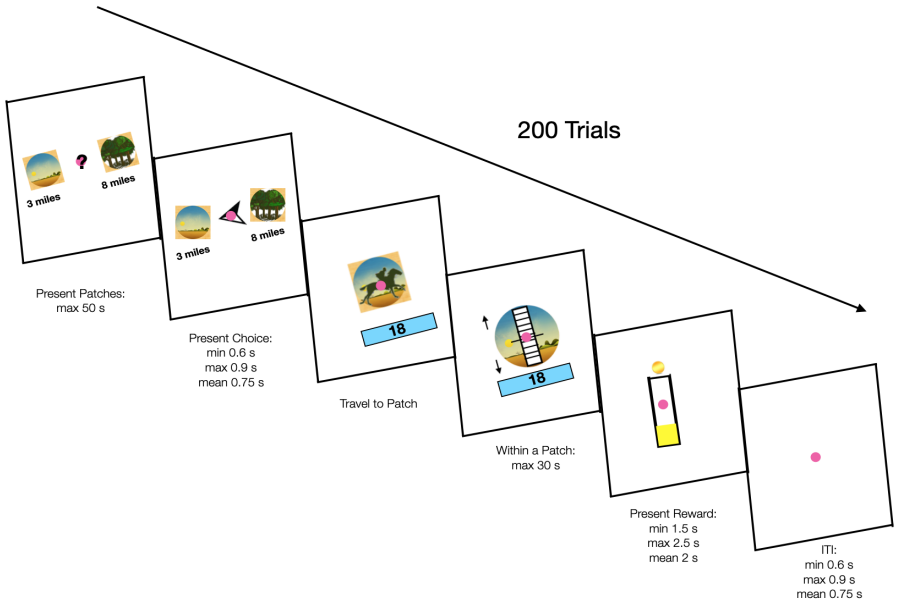


Figure 6: Timeline of the foraging task. Participants chose between two patches by weighing the reward structures and travel time. After selecting a patch, they saw an animation of travelling to the patch and water bar shrinking. Once entering the patch, participants started collecting gold and saw reward bar moving every second. Note the gold collecting process was automatic and no extra button press was required. Every several seconds (40% every 2 s) when a red dot appears on the screen participants may press SPACEBAR to leave. After leaving the patch, accumulated reward during this visit was shown, followed by a jittered ITI.

shown on the screen. Participants chose between the two patches by pressing the left or right arrow key within 50 seconds. The deadline was set to keep people engaged in the task and there was no water consumption during patch choice time. After making a response, the choice was presented on the screen for on average 750 milliseconds (jittered between 600 and 900 milliseconds). Then participants would see an animation of travelling to the chosen patch and the shrinking water bar. After that, participants entered the patch and started gathering reward. The instantaneous reward rate (i.e. the amount of reward collected per unit of time) was presented by a bar moving up or down along a

scale. Below the patch it was displayed the remaining water supply. While in a patch a red dot would appear on the centre of the screen every several seconds (every 2 seconds there was 40% chance for the red dot to appear) and when it appeared it would remain on the screen for 1 second and participants were able to leave the patch by pressing the SPACEBAR during the time. We added uncertainty to the leaving chance instead of allowing a free leaving decision to promote more proactive prediction, because if the participants did not know when they would be allowed to leave again, they needed to decide whether to take the chance they got based on their prediction of the patch's future reward state. If no leaving decision was made, participants would automatically leave the patch after 30 seconds. After leaving the patch, a reward bar would be displayed indicating the amount of reward gathered in the trial for on average 750 milliseconds (jittered between 600 and 900 milliseconds). The inter trial interval (ITI) was set to be on average 1.5 seconds (jittered between 1 and 2 seconds). Ater the ITI, the next trial started and the participants may choose to re-enter the same patch (with the starting reward value reset by the new trial's SF and DR) or enter the other one. To reduce eye movements for the sake of EEG analysis, a purple dot was presented as fixation at the screen centre throughout the experiment. A timeline of the task is depicted by **Figure 6**. Participants were given two sessions with 200 trials each, but the exact number of completed trials depended on the water consumption. On average, each participant completed 300 trials in total.

Instantaneous reward rate in each trial was delivered by an exponential decay as we believe it depicts the natural reward diminishing pattern. An exponential decay is captured by two factors: a scale factor that determines the overall magnitude of the values, and a decay rate that determines the speed of decline. The scale factor and decay rate over 200 trials varied at two levels with reversals

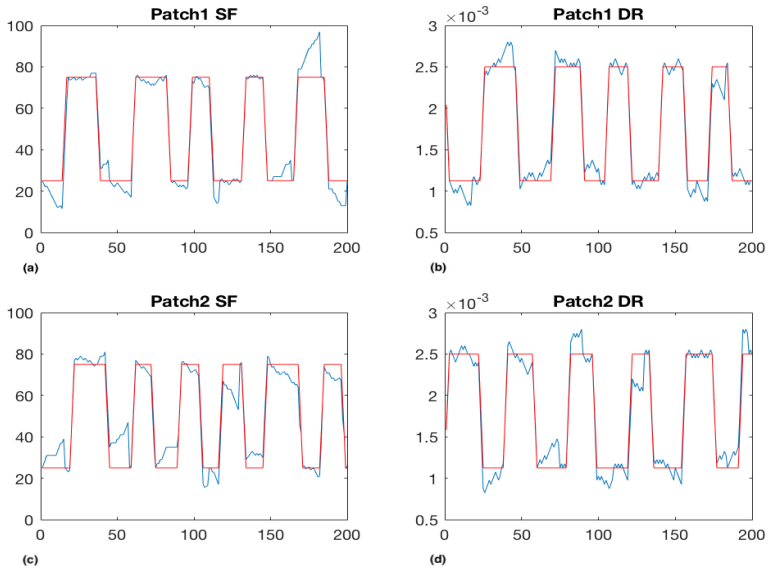


Figure 7: Schedule of scale factor and decay rate used in each patch from one participant for illustration (Patch 1: (a) and (b); Patch 2: (c) and (d)). The red line represents the noise free structure and the blue line depicts the real values. For each session a different schedule was used. X-axis: trials in a session; y-axis: values.

every 13-23 trials (for example see one participant's task schedule in **Figure 7**). We generated a different schedule for every session for each participant. A reversal means the variable changed from one level to the other and was achieved within 3 trials. The variables were combined with noise, which was generated by random walks with two separate drift rates: at each reversal the drift rate for the scale factor was randomly drawn from the sequence [-1, -0.5, 0, 0.5, 1] and for the decay rate from [-0.0002, -0.0001, 0, 0.0001, 0.0002]. We made sure that all the variables between the two patches within each participant were decorrelated (correlation coefficient < 0.5; **Figure 8**).

The scale factors and decay rates were fed into a simple exponential decay

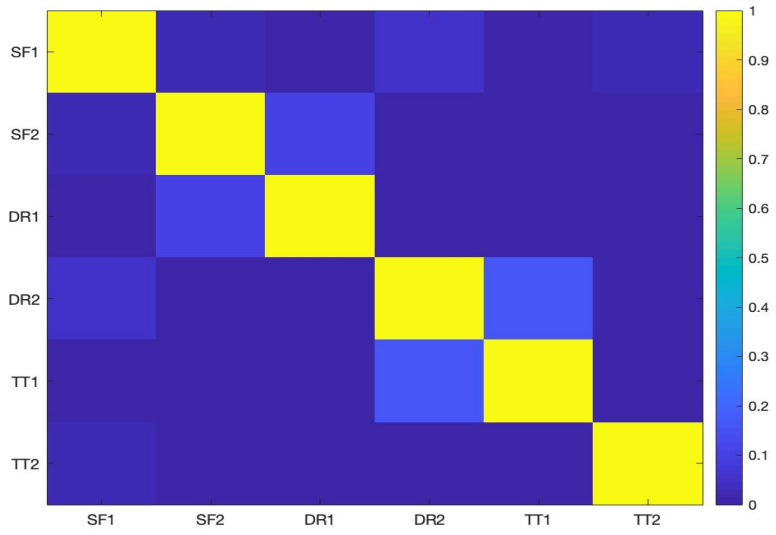


Figure 8: Correlation coefficients between the variables of two patches within each participant (SF1 = scale factor of patch 1; SF2 = scale factor of patch 2; DR1 = decay rate of patch 1; DR2 = decay rate of patch 2; TT1 = travelling time of patch 1; TT2 = travelling time of patch 2). All correlation coefficients were below 0.5.

equation to generate the final reward trajectories:

$$R = SF \times e^{-DR \times Timestep} \tag{3}$$

The maximum duration of each patch (= 30 seconds) was broken down into 1800 timesteps, with 60 timesteps per second. The reward trajectories were then combined with random noise and jittered by 1 second (i.e. 60 timesteps; **Figure 9**).

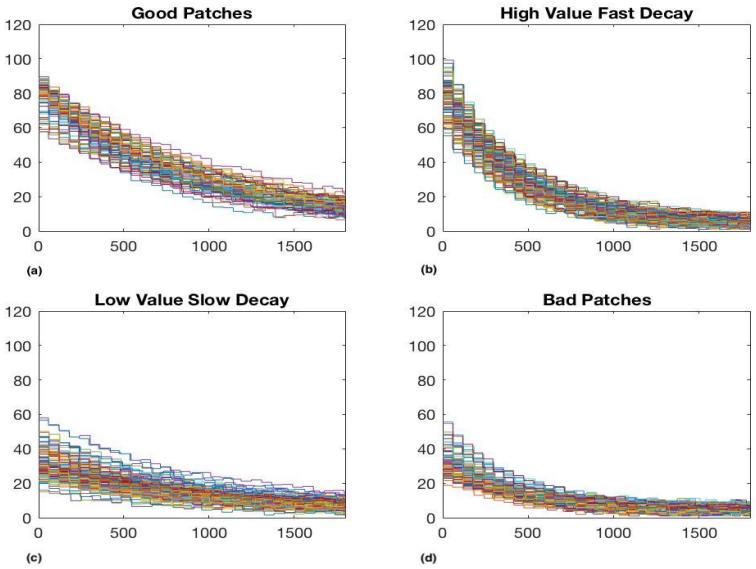


Figure 9: Reward trajectories used in the foraging task: (a) Patches associated with high scale factor and low decay rate (i.e. good patches); (b) Patches with high scale factor and high decay rate (i.e. mediocre patches); (c) Patches with low scale factor and low decay rate (i.e. mediocre patches); (d) Patches with low scale factor and high decay rate (i.e. bad patches). X-axis: timepoints (60 timepoints = 1 second); Y-axis: values in arbitrary unit.

2.3 Behavioural Analyses

In this study, we were interested in how people learn about simple reward structures within an environment. Additionally, we wanted to understand how people utilised the information for choosing between such environments and when to leave them. We hypothesised that people learned the overall value of an environment or the individual components. In our case, an environment was well characterised by a "scale factor", i.e. level or reward at the beginning and "decay rate", i.e. how fast the reward went down. To characterise the learning process we build a Bayesian learning model that tracks the most likely scaling factor and decay rate, given all observations in a trial. This is a useful solution to the learning problem because such models can easily implement learning within and between trials without a need for additional assumptions as the model can simply assign a probability of a specific reward event occurring given the underlying reward function (characterised by a scaling factor and decay rate). For alternative reinforcement learning approaches one would have to decide on whether to reinforce on an environment's overall structure at the end of the trial or somehow use each reward event in a complicated manner to inform value estimates. However, participants might also instead simply reinforce on their reward experiences directly (e.g. accumulated reward in a trial/environment), without considering those structural features directly. Additionally, although tracking the decay rate and scaling factor might be enough to compute the maximal average reward rate for each environment, leaving decisions may require different considerations in addition to a good approximation of the environment's value. In fact, because our reward environments were monotonically decreasing, the most important feature for making good leaving decisions is comparing the current reward level and its recent rate of change with their expectation of how good the next reward environments will be. To

understand exactly how participants integrate the features of the environment to choose between them and what information they use to decide when to leave, we did two separate model comparison based on logistic regression analyses, one for each type of behaviour (environment selection and leaving decisions).

2.3.1 Reinforcement Learning

We considered that participants might utilise simple learned values of the environments to guide their selection and leaving decisions. To test this, we built a simple reinforcement learning (RL) model using the basic Rescorla Wagner updating rule (A. R. Wagner and Rescorla, 1972):

$$Value(t+1) = Value(t) + \alpha \times RPE(t) \tag{4}$$

...where α denotes the learning rate, which was fit at group level (i.e. one learning rate for all participants). As the models were very simple with only one free parameter we used an equally simple grid search. Specifically, we created a vector of learning rates from 0.01 to 1 (with an increment of 0.01) and selected the optimal learning rate minimising the negative likelihood (see the model comparison section) when we fitted the resulting value regressor to the behaviour. The reward prediction error (RPE) was calculated as below:

$$RPE(t) = Reward(t) - Value(t) \tag{5}$$

...where "value" refers to multiple variables as it represents the specific feature that the learner is trying to associate with an environment. Specifically, we built models that learned one of the following: the overall value (total reward collected/accumulated in the visit; abbreviated as $AccRR$), average reward rate ($AvgRR$) or average reward rate within a patch ($AvgRRIN$). The difference

between $AvgRR$ and $AvgRRIN$ is the former takes into account travel time while the later does not, as specified by equation 6 and 7:

$$AvgRR = AccRR/(LT + TT) \tag{6}$$

$$AvgRRIN = AccRR/LT \tag{7}$$

...where LT represents the leaving time (time spent in the patch) and TT represents the travel time. As travel time was randomly drawn in each trial and thus could not be learned across encounters of the same patches, we examined if participants actually took a) time into account at all and b) if so considered travelling time as important when learning across trials.

2.3.2 Bayesian Learner

To make good decisions between environments participants had to learn about their reward structure. One way to do this is by learning the overall value of an environment option or by learning about its separate components (SF and DR). The RL model we implemented assumed people updated the value of a patch based on past reward experiences directly. Alternatively, the value could be inferred from one's knowledge about the reward structure i.e. the components of the environment that determine the reward over time. In our task those components were a scaling factor and a decay rate. To test whether participants might use those components to choose between patches we built a Bayesian learner that inferred the underlying scaling factor and decay rate by comparing each reward event to functions generated using a grid of possible scaling factors and decay rates. As an environment's reward values were determined by a simple exponential decay, for each possible pair of SF and DR a reward trajectory was

calculated as below:

$$V = SF \times e^{-DR \times t} \tag{8}$$

Where t denotes the time point at which an observation was made. Whenever a reward event was observed, the likelihood for such an observation was calculated for such an event using scaling factor and decay rate. Because we had added some Gaussian noise to the reward values they were unlikely to fall exactly on the functions' predictions and we therefore assumed a normal distribution around each point on the function instead. Posterior probabilities were obtained by simply multiplying the trial's likelihood with the previous prior according to Bayes rule.

$$p(Factor|Data) = p(Factor) \times p(Data|Factor) \tag{9}$$

Where the prior probability of a factor p(Factor), that is SF, DR and TP (transition probability), was initialised by a flat prior in the first trial. In subsequent trials, the last trial's posterior is combined with the transition function below to form the next trial's prior. $p(Data|Factor)$ is the computed likelihood of an event and $p(Factor|Data)$ is the posterior of an event.

Posteriors were taken as next events prior within a trial but as the environment could have changed between trials, the posterior probability from one trial only partially formed the prior belief in the next trial. In other words, we assumed that the posterior was only partially used, with the rest being determined by a transition function (i.e. the belief the environment was unchanged). As we assume participants had no particular knowledge of where change would occur to we just used an uninformative, flat prior. Mathematically, the transition probability determined the fraction of a trial's prior that was a flat rather

than last trial's posterior, as the agent was assumed to not know where the values would transition to:

$$p(Factor) = k \times U + (1 - k) \times (Factor|Data) \tag{10}$$

The parameter k denotes the transition probability and U represents the flat prior. Note that while SF and DR were used at each time point when computing likelihoods, the TP only came into play between trials to determine how much of the previous trial's posterior was used. Thus, the complete grid had the following variables:

Variable Name	Denotation	Vector Space
Scale Factor	SF	4 : 2 : 106
Decay Rate	DR	0.0001 : (0.0075-0.0001)/50 : 0.0075
Transition Probability	TP	0.04 : 0.02 : 0.24

Table 1: Estimated variables in Bayesian model.

There are two common ways of extracting a singular "expectation" from each factor using our Bayesian model. (1) After marginalising (i.e. summing up all other dimensions) one can average all possible values in the probability distribution by multiplying values with their respective probability; (2) Alternatively, one can marginalise and then take the state with the highest probability as representative. Which one is better depends on the data and task at hand. **Figure 10** displays the estimated SF and DR for two sessions from one participant, plotted together with the ground truth and updated magnitude at each time. Compared to Bayesian max, Bayesian mean provided estimation closer to the ground truth. This might be because the means integrate over the whole belief space rather than taking a singular value or because our grids were relatively sparse for computational reasons. Whichever the reasons in subsequent analyses we chose Bayesian means as the model prediction due to their better

tracking of the underlying ground truth. **Figure 11** shows an illustrative figure of transition probability from one participant.

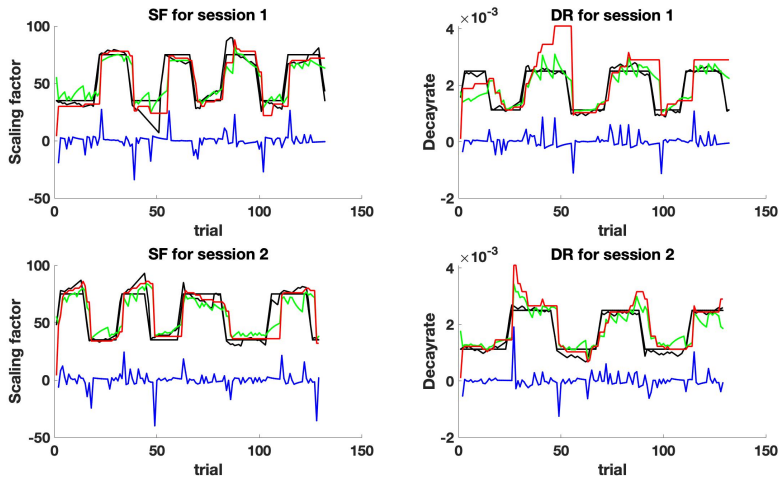


Figure 10: Ground truth and model prediction of SF and DR in two sessions illustrated by one subject. Plotted here are the noise-free variable values (straight black line: *no Noise*), actual variable values with Gaussian noise (jittered black line:*with Noise*), model predicted values by averaging all estimations (green line: *Bayesian mean*), model predicted values by picking the maximal posterior probability (red line: *Bayesian max*), the amount of update since last trial, with a peak or a dip occurring when a real change happened (blue line: *update since last value*).

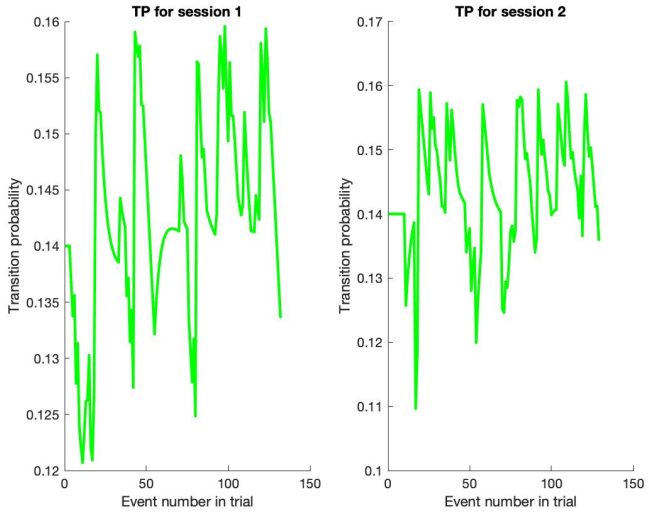


Figure 11: Transition probability illustrated by one subject. Two sessions were plotted separately.

2.3.3 Multiple regression analyses of patch selection and leaving decisions

To understand how participants did our task we ran a series of logistic multiple regressions to understand participants' patch choices and leave choices. Regression results were evaluated by looking at the beta weights as well as the negative log likelihoods (negLL) and Bayesian information criterion (BIC) (see next section). We had one candidate model for each type of decision that we were particularly interested in, which we will explain in detail to illustrate the general approach.

Patch choice

First, we looked at the patch choices to test whether they were more influenced by overall reward structure or the individual components of the structure.

Therefore, we built the following logistic regression to predict patch choice:

$$\begin{aligned} y(pc) = \beta_0 + \beta_1 \times TT_A + \beta_2 \times TT_B + \beta_3 \times SF_A \\ + \beta_4 \times SF_B + \beta_5 \times DR_A + \beta_6 \times DR_B \end{aligned} \tag{11}$$

...where y is participants choices of patch B over A, TT denotes the travel time which was presented to participants, and both SF and DR were derived from the Bayesian model. However, instead of directly using the SF and DR for making choices, participants could have derived the optimal possible maximal possible average reward rate ($MaxAvgRR$) given SF and DR and take under consideration together with travelling time. Thus, we ran the following regression as well:

$$\begin{aligned} y(pc) = beta_0 + \beta_1 \times MaxAvgRR_A + \beta_2 \times MaxAvgRR_B \\ + \beta_3 \times TT_A + \beta_4 \times TT_B \end{aligned} \tag{12}$$

While $MaxAvgRR$ denotes the maximum average reward rate computed with the SF and DR derived from the Bayesian model and shown travelling time, we also included the travelling time again, as participants could be affected by this information above what is expected from the $MaxAvgRR$ term.

Leaving decisions

We wanted to test whether leave choices would not only depend on the current reward level but also on the amount of reward change in the patch as well as the recent reward experiences across trials. Both of those features are important for making leaving decisions because the local change within a trial gives a good heuristic for reward levels collapsing and the history of reward experiences across trials gives a sense of reward richness and therefore could guide predictions about patch value for the next trials. We predict that both when the reward

level has been collapsing quickly in a trial or when the past patch experiences were particularly rich in reward, participants should leave earlier, as they may re-enter the rich patch and collect more reward using the same amount of time (as the patch would be replenished and the reward rate was relatively high in the beginning). Thus we built the following regression models to test this behaviourally:

$$\begin{aligned} y(lt) = \beta_0 + \beta_1 \times RewardLvL + \beta_2 \times RChange(R_x - R_{x-1}) \\ + \beta_3 \times AvgRinHistory(_{t-1}\ to\ _{t-5}\ Avg) \end{aligned} \tag{13}$$

...where y(lt) represents leaving at the given probe, $RewardLvL$ denotes the current reward level, $RChange$ computes the reward change since last reward event x, and $AvgRinHistory$ calculates the average reward rate ignoring travel time across past five trials. Note that for computing $AvgRinHistory$, a learning rate was fit at group level, as a parameter controlling the decay of past information.

To illustrate the elements of the past reward history we also broke $AvgRinHistory$ down into its constituent parts and also included the past travelling times (TT):

$$\begin{aligned} y(lt) = \beta_0 + \beta_1 \times RewardLvL + \beta_2 \times RChange(R_x - R_{x-1}) + \beta_3 \times AccR_{t-1} \\ + \beta_4 \times AccR_{t-2} + \beta_5 \times AccR_{t-3} + \beta_6 \times AccR_{t-4} \\ + \beta_7 \times LT_{t-1} + \beta_8 \times LT_{t-2} + \beta_9 \times LT_{t-3} \\ + \beta_{10} \times LT_{t-4} + \beta_{11} \times TT_{t-1} + \beta_{12} \times TT_{t-2} \end{aligned} \tag{14}$$

Accumulated reward in past four trials and travel time in past two trials were taken into account here, as we assume participants were aware that TT was randomly drawn in each trial and remote TT in history was not helpful in

projecting the future.

2.3.4 Model Comparison

We used two model selection metrics: the raw negative log likelihoods (negLL), where the numbers of parameters were identical, this number could be used to compare models directly. However to get specific cut-offs for better models and for when there were different numbers of parameters between models we also computed the Bayesian Information Criterion (BIC). For each model, we calculated negLL and BIC as follows:

$$negLL = -sum(log(p(Data))) \tag{15}$$

Given the current regression model.

$$BIC = 2 \times negLL + log(n) \times k \tag{16}$$

Where k equals the number of free parameters, and n the sample size or number of observations.

Another, popular metric AIC, is equal to the BIC we computed, except that it does not penalise model complexity as a function of the log number of observations (log(n)).

2.4 Behavioural Results

In the behavioural task, we are interested in two types of decision: participants chose which patch to travel to (patch choices) and they decided when to leave the patch (leave choices). Therefore, we first looked at the behavioural patterns associated with the two types of choices.

2.4.1 Basic Breakdown

Patch choices

Participants picked the patch with higher scaling factor, lower decay rate and the combination that yielded higher reward rate, as indicated by (**Figure 12**). **Figure 13** shows that participants picked the patch with higher maximum average reward rate.

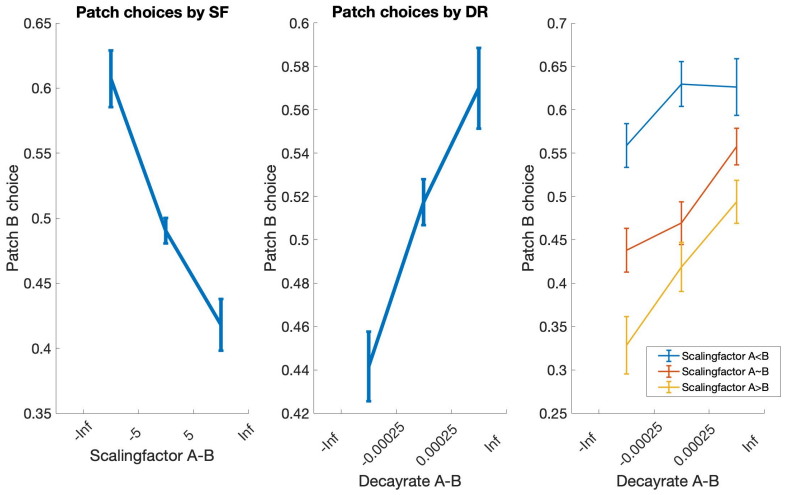


Figure 12: Proportion of choosing patch B given (1) scale factor patch A minus patch B; (2) decay rate patch A minus patch B; (3) split by both decay rate (x-axis) and scale factor (lines).

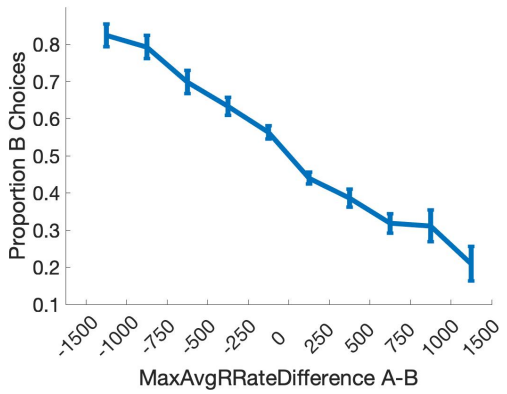


Figure 13: Proportion of choosing patch B given difference in maximum average reward rate between patch A and B. As patch A became advantageous, the proportion of choosing B over A decreased.

Leave choices

Most participants left the patch actively with a few exceptions. A cut-off threshold of 70 trials was applied (shown as a black line in **Figure 14**). Participant 5 and 20 showed a strong tendency to overstay in a patch as they left fewer than 70 trials throughout the two sessions. Despite of their overstaying tendency, the two participants were still included in the rest of analyses. **Figure 15** shows that on average 60% of the trials participants pressed to leave. Individual plots show that a few subjects almost never pressed to leave while a few left almost all in the first 15 seconds (right panel of **Figure 15**). Taken together it seems in a large proportion of trials participants pressed to leave and further analyses on the leave choices will reflect important factors contributing to the decision.

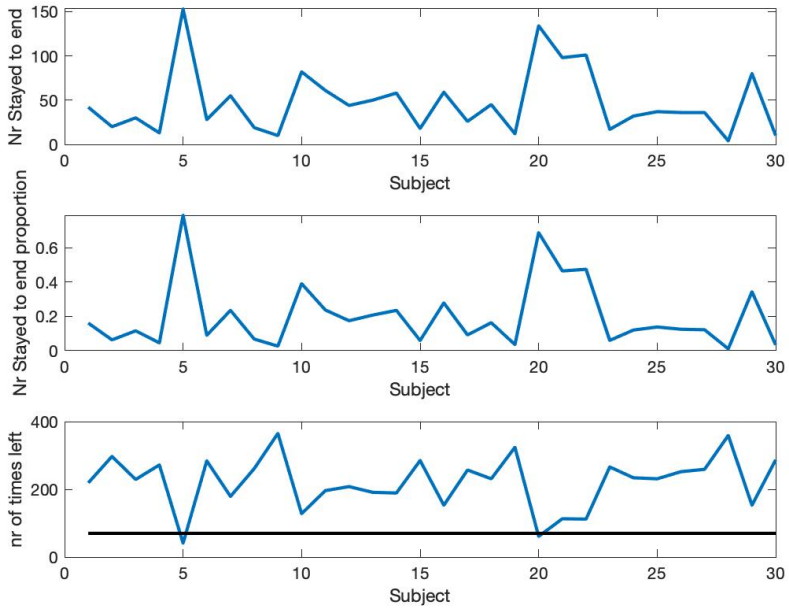


Figure 14: Participant-wise leave choices. X-axis: individual subjects for a total of 30. Y-axis: number of trials that stayed to the end (maximum patch stay for 30 s) (top); number of trials in proportion (middle); number of trials that pressed leave (bottom). The black line displays a cut-off threshold of 70 trials. Twenty-eight participants pressed to leave in at least 70 trials, whereas participant 5 and 20 tended to overstay.

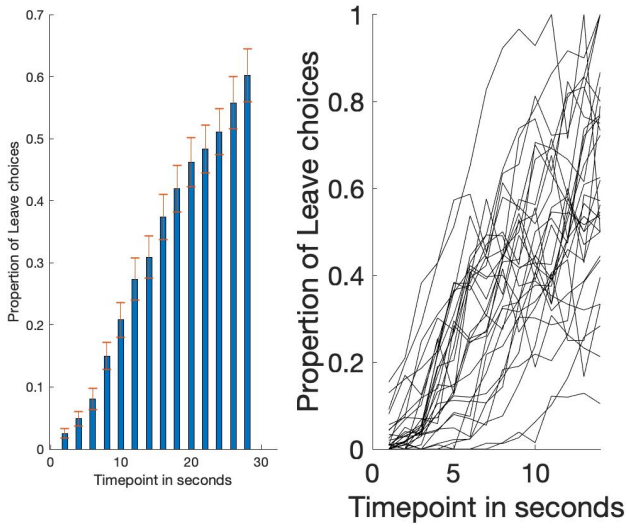


Figure 15: Proportion of accumulative leave choices as a function of time. The left panel plots the average across all participants and the right panel each participant separately (black lines). On average participants pressed to leave in 60% of the trials. Most participants left by 15 s in more than 50% of the trials, except a few left all the time and a few almost never left by 15 s.

2.4.2 Regression Results

For the patch selections we tested whether participants used simple heuristics that compared SF and DR and travelling times directly or combined all three factors into one singular value estimate. Alternatively, they might also just learn the overall values of a patch directly, ignoring SF and DR but instead reinforcing on the reward experiences and time. As it is not self-evident which version of "overall" value participants will compute, we tested three possible versions that represented the overall "quality" of a patch: (1) raw accumulated reward for a trial, ignoring the time it took to get it; (2) the overall average reward rate (accumulated reward rate divided by the sum of the time spent in the patch and travel time); (3) average reward rate in a patch (accumulated reward rate divided by time spent in the patch; this harvesting rate ignores the time spent

travelling). Regarding the integrated value estimate based on Scaling factor and decay rate we computed a maximum average reward rate based on the average reward curve by leaving for the presented travelling time (see **Figure 16** for illustration).

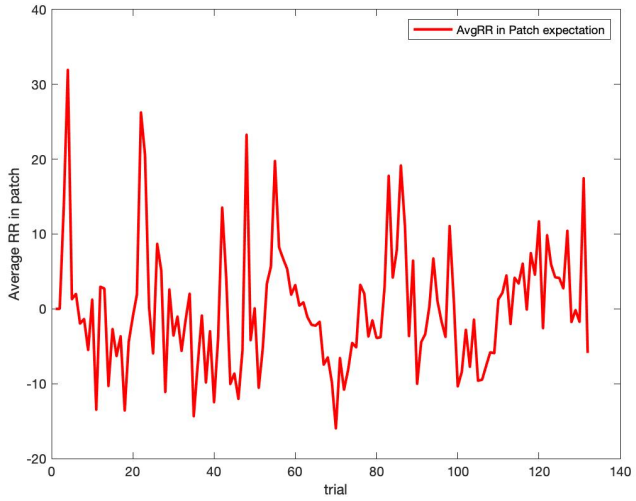


Figure 16: Expected maximum average reward rate given SF, DR and TT of the chosen patch. Presented here are data from one session from one subject for illustration.

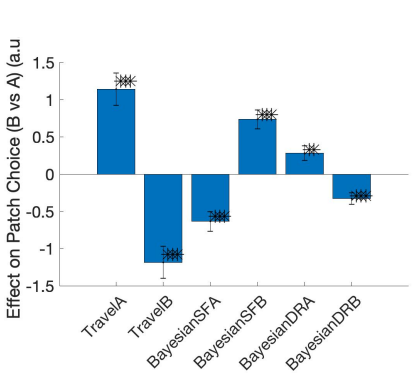
Last, but not least, we also used scaling factor and decay rate directly to inform choices together with the travelling time. We built all these logistic regressions to predict participants' choices to compare their ability to explain participants choices using model fit metrics.

For the patch leaving decisions, we tested whether participants took factors into account that went beyond what they saw on the screen and what they were experiencing in this trial (e.g. current reward level and derivative (difference between this and last reward sample)). Marginal Value theorem (MVT) predicts that participants should leave when the current instantaneous reward rate is at or below their expected long term average reward rate outside the patch,

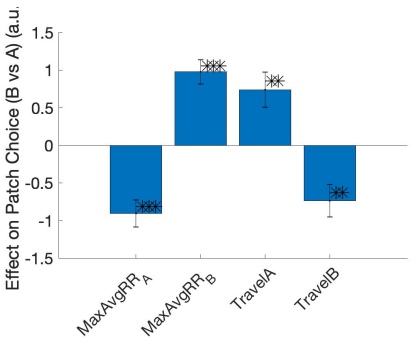
assuming that they are in a monotonically decaying patch. However, it leaves open how participants might build such expectations: (1) they might use the average reward rate in the past few trials including both travelling time and leaving time to determine the average; (2) they might ignore travel times when computing the average, as it was randomly sampled on every trial and therefore uninformative about the current richness of an environment; (3) they might simply track the aggregate or overall accumulated reward in the past few trials, irrespective of the time it took to collect the reward. Additionally, participants might use faster time scale predictions of the change of reward within a trial (e.g. in the form of computing the first derivative), which we also included and tested behaviourally. In other words, they could use the derivative to detect whether reward might soon collapse to an unacceptable level and leave preemptively.

2.4.2.1 GLM for patch choice

The Regression results show that Bayesian derived components and the corresponding maximum average reward rate captured participants' choices (**Figure 17**). Overall, people picked the patch with higher estimated SF, lower DR and higher expected return (maximum reward rate). We computed negative log likelihood (negLL) and BIC for each regression model (**Figure 18**, with statistics displayed by **Table 2**). Although there is no consensus in what level of difference in negLL and BIC is considered as the criterion of a good model, Raftery, 1995 suggested that >10 difference in BIC indicates very strong evidence in favour of the model with lower BIC. While using the Bayesian estimated SF and DR directly had an equally low negLL as using the Bayesian derived maximum average reward rate, the latter had a lower BIC because it required fewer parameters to fit equally well. Either way, it seemed that a Bayesian model that tracked the maximal average reward rate seemed to capture the learning process well.



(a) Bayesian components



(b) Bayesian derived max average reward rate

Figure 17: Beta weights of different factors (travel time and estimated SF and DR from the Bayesian model) predicting choosing patch B vs patch A. (a) model with components from Bayesian model as regressors; (b) model with derived maximum average reward rate (= accumulated reward at each time point / travel time) from Bayesian model. Factors in favour of patch B positively predicted choosing B over A, and factors against B had the opposite effect. Asterisks indicate statistical significance (t test against zero (df = 29); *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$).

Regressor	t score df = 29	P-value
Model (1)		
$Travel_A$	5.2000	0.0000
$Travel_B$	-5.5478	0.0000
$BayesianSF_A$	-4.7969	0.0000
$BayesianSF_B$	5.8202	0.0000
$BayesianDR_A$	2.8416	0.0081
$BayesianDR_B$	-4.2655	0.0002
Model (2)		
$MaxAvgRR_A$	-5.0325	0.0000
$MaxAvgRR_B$	6.0091	0.0000
$Travel_A$	3.1654	0.0036
$Travel_B$	-3.4310	0.0018

Table 2: Statistics of variables in the regression model for patch choice (t test against zero). (1) GLM with Bayesian components; (2) GLM with Bayesian derived maxim average reward rate and travel time.

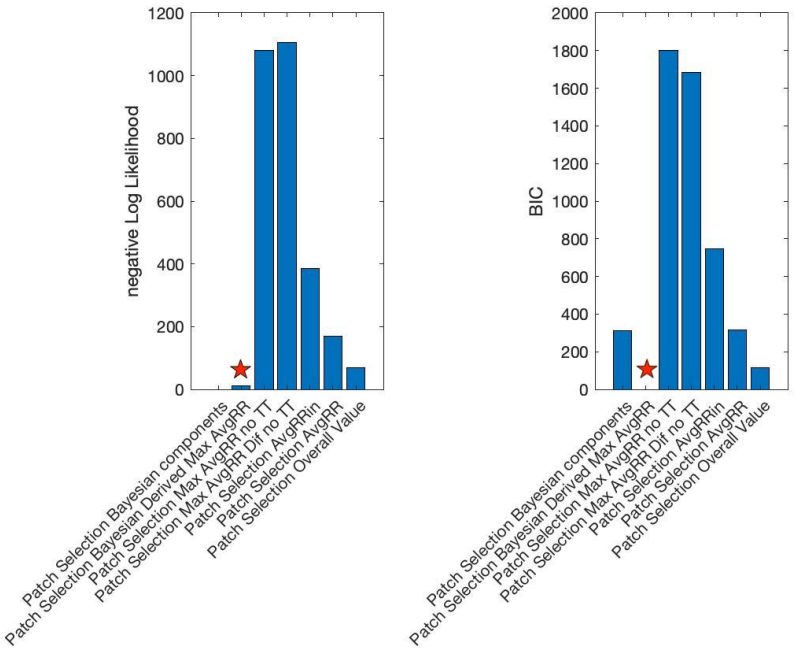


Figure 18: Model selection for patch choice: negative log likelihood plotted (left) and BIC (right) plotted against using different models. Lower negLL and BIC scores indicate better performance. Taken together, model with Bayesian derived maximal average reward rate and travel time was selected as the winning model (marked).

2.4.2.2 GLM for leave choice

First we wanted to test whether there was any evidence for participants taking the previous trial's reward history into account when making leaving decisions (**Figure 19**). We found that either using only the last trial's average reward rate in the patch or the average of the last 5 trials was better than not using the history at all.

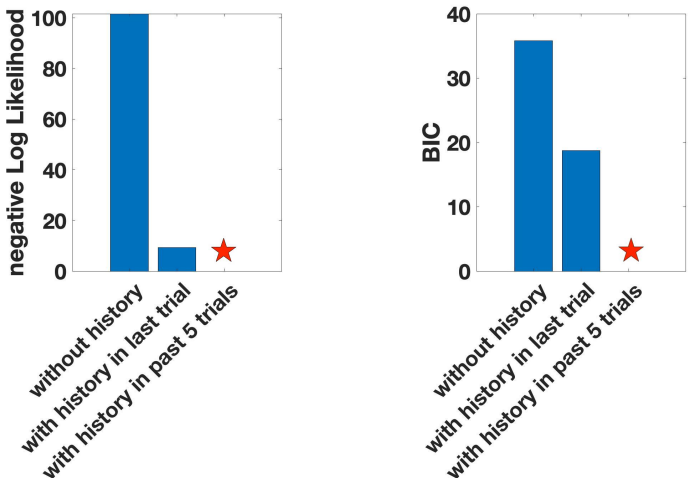


Figure 19: Model comparison for leave choice: (1) model without history; (2) model with history in the last trial; (3) model with history in up to past five trials. Taken together, model with average reward rate across past five trials was selected (marked).

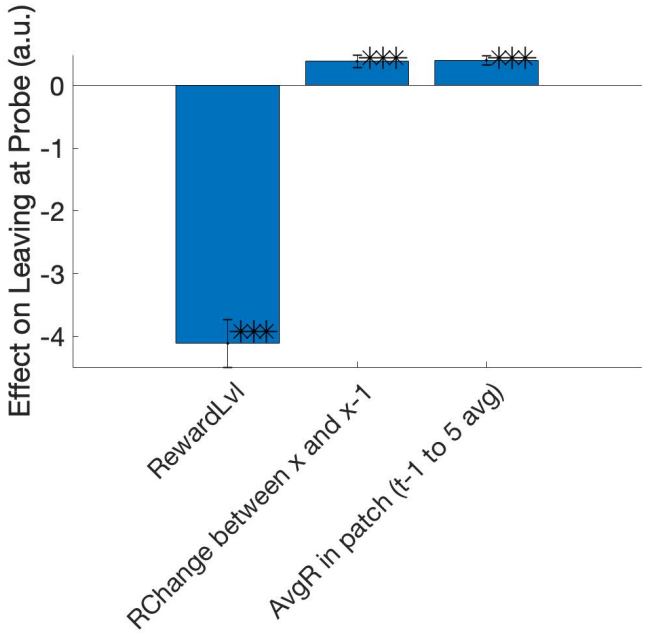


Figure 20: Beta weights of different factors predicting leaving decisions at each leaving probe. RewardLvL: current reward level; RChanget-t-1: the amount of reward decaying since last reward event; AvgR in patch: average reward rate associate with the patch in past five trials. Asterisks indicate statistical significance (t test against zero (df = 29); *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$).

Then we wanted to look at other factors driving leaving decisions (**Figure 20**). As we expected the largest driver of leaving was the current reward level. Beyond this, somewhat surprisingly, participants were slightly more likely to leave if the negative change of reward from one sample to the next was smallest, suggesting participants might be overstaying in more rapidly collapsing patches.

Again the results indicate that participants were more likely to leave i.e. left earlier, if the average reward history in the previously experienced patches was consistently high (**Figure 20**). To further illustrate the elements contributing to the effect of past experience on leaving, we ran another GLM that broke down

the reward rate associated with the past experiences into accumulated reward rate and leaving time in the past four trials. We also included the travel time in the past two trials separately to test whether there was an additional effect of this time factor (**Figure 21**). The results illustrate what our model comparison formally confirmed, i.e. that past experience with both accumulated reward rate and leaving time influenced leaving decisions, but previous travel time as a random factor did not have any effect. Relevant statistics are presented in **Table 3**.

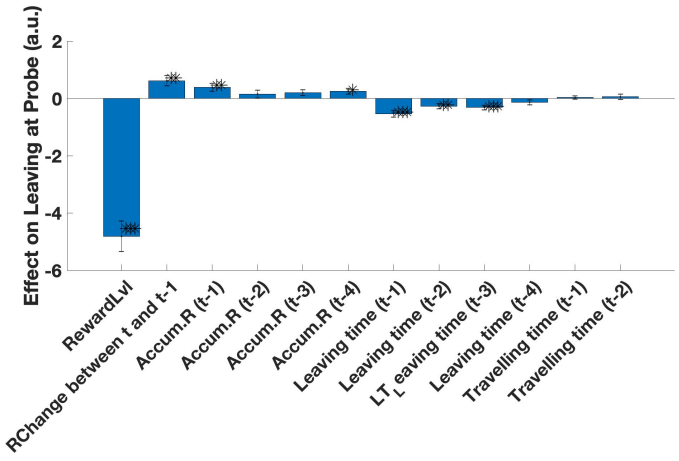


Figure 21: Regression results with average reward rate within a patch broken down into accumulated reward rate (up to four trials in the past), leaving time (four trials in the past) and travel time (two trials in the past) associated with the same patch. Asterisks indicate statistical significance (t test against zero (df = 29); *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$). Besides the current reward level, reward change, accumulated reward and leaving time in the past all contribute to leaving.

Regressor	t score df = 29	P-value
Model (1)		
RewardLvL	-10.7908	0.0000
RChange between x and x-1	3.7002	0.0009
AvgR in patch (t-1 to 5 avg)	5.6393	0.0000
Model (2)		
RewardLvL	-9.0952	0.0000
RChange between t and t-1	3.3816	0.0021
AccumR (t-1)	2.8606	0.0078
AccumR (t-2)	1.1122	0.2752
AccumR (t-3)	1.9899	0.0561
AccumR (t-4)	2.6139	0.0140
Leaving time (t-1)	-4.5530	0.0001
Leaving time (t-2)	-2.8667	0.0076
Leaving time (t-3)	-3.7612	0.0008
Leaving time (t-4)	-1.5066	0.1427
Travelling time (t-1)	0.6258	0.5363
Travelling time (t-2)	0.6324	0.5321

Table 3: Statistics of variables in the regression model for leave choice (t test against zero). (1) model with reward change and history; (2) model with average reward rate broken down.

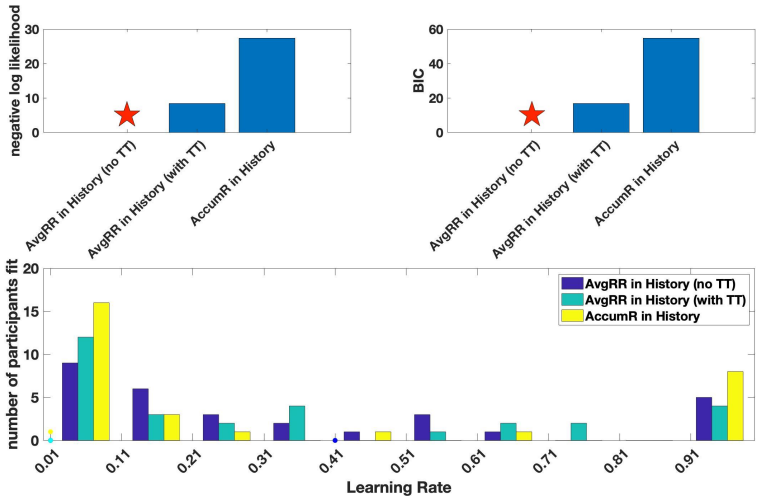


Figure 22: Model selection for leave choice: negative log likelihood (top left) and BIC (top right) plotted against different models with the winning model marked, and learning rates fit to individual participants (bottom). Best group-wise learning rates fitted in each model are shown by coloured dots.

We compared several models that were proposed to account for leave choice. **Figure 22 and 19** show the outcomes of model comparison. The model using average reward rate derived from a reinforcement learning model ignoring travel time was selected as the best candidate by both negLL and BIC. Past experience in the previous five trials mattered when deciding to leave. The notion that travel time as a random factor was ignored in leaving decision was supported by both GLM (non-significant as a predictor) and model comparison results (the model without travel time won out). The use of five previous trials was semi-arbitrary but confirmed by the fact that for the broken down GLM (**Figure 21**), beta weights decayed to zero in the past fourth and fifth trials. This decay shows that participants learned past experience in a way similar to the exponential function described using A. R. Wagner and Rescorla, 1972's alpha learning rule. Thus, we also fit such learning models (i.e. effect of past trials on the current choice decaying as an exponential function, with more distant trials casting lower influence) to the leaving behaviour.

Fitting such a learning model also allowed us to get one single value for the reward history (*AvgRRinHist*) to use for the EEG analyses. To have more comparable and robust regressors for this, we used the group fit learning rate, as some participants' learning rates did not fit very well, presumably because the history effect was not as strong as in standard Reinforcement learning tasks. For the models that used other types of overall value learning, getting proper learning rate was even harder, presumably because those values were not good descriptions of participants' behaviour to start with.

2.4.3 Summary of Behavioural Results

Taken together, participants tried to pick the best patch, using the estimated maximum reward rate (*AvgMaxRR*) to evaluate each option. A Bayesian model that tracked SF and DR underlying the reward structure also seemed to capture

learning processes well. However, we cannot conclude whether people chose a patch based on the overall estimate on the patch quality or based on the underlying components. The leaving decision was based on participants' current reward experience, as well as past recent reward experience in a trial and across trials. Specifically, people left early when the previously experienced patches were consistently good, possibly because it implied an overall rich environment and sticking to one patch yielded relatively high opportunity cost. It was also found that up to five past trials were taken into account in leaving decisions. Such decay in past information was useful as the reward structure of the environment regularly changed and recency filtering is good in environments subject to change (Behrens et al., 2007). Given such decay in the importance of past information, we computed average reward rate in past five trials as a single value for the reward history for subsequent EEG analyses. This together with other important factors (e.g. *AvgMaxRR, LeavingTime*) were used as essential behaviour features to analyse EEG signals, which will be covered in the next chapter.

2.5 EEG

2.5.1 Recording and Pre-processing

Electroencephalographic signals were recorded with active electrodes from 60 scalp electrodes mounted equidistantly on the standard 10-20 system elastic map (EasyCap). All electrodes were referenced to the right mastoid electrode and re-referenced offline to the average of two mastoids. Continuous EEG was recorded using a SynAmps RT 64-channel Amplifier (1000 Hz sampling rate). Offline EEG analysis was performed using a MATLAB toolbox developed by the Donders Institute in Nijmegen (FieldTrip). The data were epoched from -1 to 2 s for all trial events. The data were then band-pass-filtered at [1 35] Hz. Artefacts were detected and rejected using a correlation-based independent component analysis (ICA) approach provided by FieldTrip. Before running ICA, trials containing artefacts were removed by manual artefact rejection (with 'ft_rejectvisual' function of FieldTrip that visualises trial variances and allows the user to remove trials and electrodes directly from the GUI) to avoid contaminating component training. Artefact components were rejected if their correlation coefficients with the EOG electrodes exceeded 0.7. Muscle artefacts were detected using z-transforming on eight electrodes ('PO7', 'PO8', 'POZ', 'PO4', 'PO3', 'O1', 'OZ', 'O2') near the neck which were most affected by muscle movements with a z-score cutoff at 20. After muscle filtering a second visual inspection was applied to the data. We excluded 9 participants whose data contained an excess of artefacts (more than 20% trials rejected) from the subsequent analyses, yielding 21 participants in total.

2.5.2 Decoding Analyses Overview

An important characteristic of our task was the need to process and integrate multiple sources of information, such as reward components, travel time,

past reward history and expectation on reward structure changes. The active processing of the task related information in the brain usually happens relatively quickly (i.e. within several hundreds milliseconds). With high temporal resolution of EEG it was still challenging to understand how relevant information segments were processed and represented, in separate or combined, to guide goal-directed behaviour (patch choice and leave choice). Recent work has shown it is possible to decode EEG signals to track neural signatures with relatively rich information (Hubbard et al., 2019). Specifically, decoding models are trained with a subset of data (i.e. training set) to predict task relevant regressors with brain signals and evaluate the prediction with a test set (i.e. a subset of data that are not used in training). Cross-validation (i.e. hold out a subset of training data to measure goodness-of-fit and adjust parameters) is usually performed to improve generalisation.

We therefore aimed to use a decoding approach to understand the neural dynamics underlying choosing between the two patches and leaving a chosen one. Particularly, we would like to track the temporal dynamics of the following events: value signals (overall value of a patch vs reward components), reward changes, aggregated reward obtained at the end of a trial and decision signals (patch choice and leave choice). Moreover, we hope to answer questions related to implicit calculation of patch values and decision variables: do we see signals related to participants' future behaviour or recent reward history? Do we have value signals or just simpler action signals? Did participants track the current reward and reward change when viewing passive reward events (i.e. the reward events without pressed leaving or probes)? Did they keep the last trial's reward history in mind even they did not have to make a choice (i.e. when there is no leave decision to be made)? Do we see any Bayesian update signals when processing the overall value of the environment?

2.5.2.1 Standard Decoding Approach

Standard decoding is usually performed as a regression model which predicts stimulus features using brain signals. Let t index time point, X contain EEG data (trials by channels) and Y contain stimulus features (trials by stimulus or behaviour features), the regression model can be formulated as (summarised by Vidaurre et al., 2019):

$$Y_t = X_t v_t + \epsilon_t \tag{17}$$

...where ϵ_t is normally distributed noise and v_t denotes the decoding weights (channels by stimuli). Solving for the equation it is obtained:

$$v_t = (X_t' X_t)^{-1} X_t' Y_t \tag{18}$$

In this way, the stimulus and behaviour features are decoded from brain signals with regression and matrix inversion. Note that the standard decoding model runs at each time point across all trials, thus restricting the temporal variability of the model. As it has been shown that model capturing temporal variability outperformed the standard model (Vidaurre et al., 2019) and given that the complexity of our task may elicit cross-trial temporal variations, we decided not to use a standard decoding model, but a temporally unconstrained decoding analysis, a special type of HMM model that allows temporal variability suggested by Vidaurre et al., 2019 (to be introduced below).

2.5.2.2 Hidden Markov Model (HMM)

Hidden Markov models are a type of generative models that describe observable data by a set of internal factors (i.e. sequences). Each factor is called a 'state' and its corresponding observable data is called a 'symbol' (Yoon, 2009). The states are invisible (i.e. hidden) and each state determines the probability distribution of the symbol. States form a *Markov chain*, in which the future

state depends only on the current state. HMM has been widely applied to modelling biological sequences, such as protein and DNA sequences, in reading out the hidden functionality of a particular sequence (reviewed by Yoon, 2009). HMM has been proposed to model the dynamics in brain signals as the data are presented in a format similar to Markovian processes and the observable brain signals are hypothesised to be driven by hidden cognitive processes (Borst and Anderson, 2015). Specifically, the model estimates the state distributions, the probability of each state given data at each time point (i.e. *state probability*), as well as the probability of one state transitioning to another (i.e. *transition probability*).

2.5.2.3 Temporally Unconstrained Decoding Analysis (TUDA)

Instead of a standard HMM model, we used a model similar to the TUDA model introduced by Vidaurre et al., 2019 that does not assume synchronicity between trials and allows the decoder to have specific time courses (i.e. varied onset and offset time) for each trial. Essentially, the model estimates the probability of model k being active at each time point t differently in each trial. In doing so, the dimensionality of the decoder has been reduced from T to k, whereas the estimation needs to be performed at each time point in each trial. To simplify the fitting process, we fixed the number of states k to four in this task. However, variations of state number up to 10 did not seem to significantly change the results, but increased the run-time considerably.

The decoding was performed using the HMM-MAR toolbox developed by Diego Vidaurre and colleagues. We extracted a time window around three types of events: -0.5 to 0 s around patch choice, leave choice, 0 to 0.8 s around accumulated reward presentation and 0 to 0.8 s around reward events (**Figure 23**). The features consisted of signal changes at each time point, downsampled to 25 Hz across all trials. To reduce computational complexity, we performed PCA

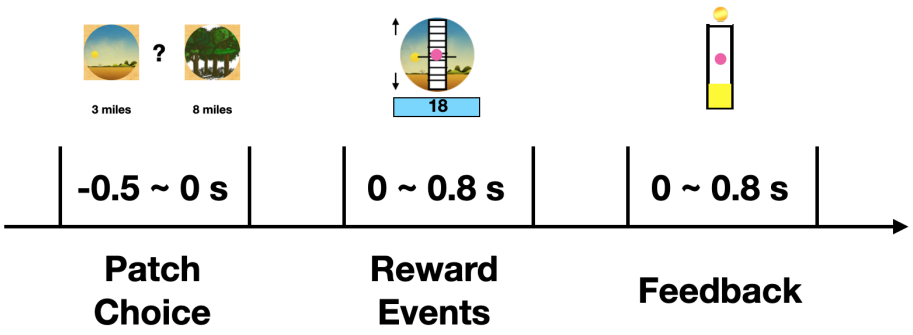


Figure 23: Task related events analysed by the decoder together with the epoched timings: patch choice (button pressed for choosing a patch), reward events (every time the reward changed) and feedback (display of the accumulated reward after leaving the patch).

and picked the first 15 components, obtaining an input dataset with dimensionality *15 components* $\times$ *time points* $\times$ *trials*. Tikhonov regularisation was used to penalise model complexity.

We analysed each epoch/time period separately. We divided the data into a training and test set. For each state from the TUDA model, we fitted a GLM predicting the values of task regressors (such as accumulated reward) from electrode values. To obtain a predicted value for the regressor value at each time point, we took the beta map for the most likely state at that time point, and compared it with the electrode-wise data at that time point. We then evaluated the predicted against actual regressor values at each time point by obtaining the Pearson correlation across trials (for illustration see **Figure 24**). For each of the events we performed two folds cross validation procedure where the dataset was randomly split into two groups and the algorithm ran with half of the data held out and the other half as the training set. With cross-validation the model kept the evaluation scores and adjusted the parameters. The prediction accuracy was measured with explained variance in behaviour features by neural data. Lastly, statistical significance test was performed by

comparing prediction accuracy against null model using permutation with 50 iterations. Data from two sessions were collapsed together, with an additional regressor controlling for the session effect.

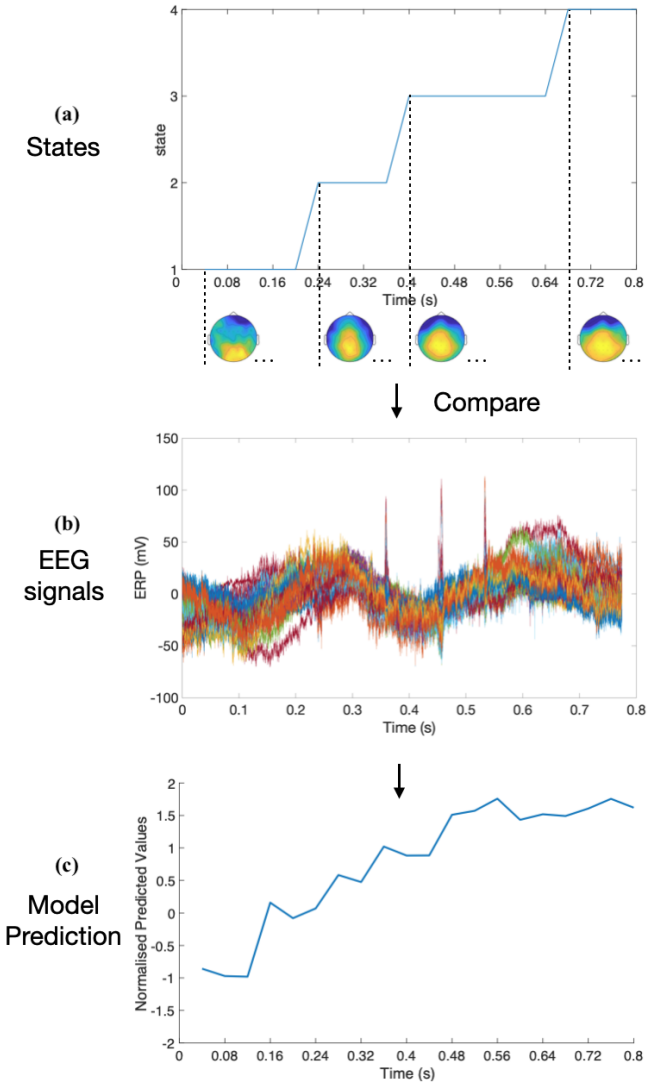


Figure 24: Illustration of how the TUDA decoder works in a specific hypothetical trial for the regressor *AccumulatedReward*. (a) Timecourse and topographies of the four states extracted from the TUDA model. The model itself was obtained from the training trials and was then used to make prediction for test trials. (b) Electrode-wise EEG signals in the test trial (plotted here are the raw signals but in practicality were downsampled to 250 Hz). (c) Time-by-time model prediction of *AccumulatedReward* for this trial obtained by comparing the states' topographies (a) with EEG signals (b). Note in actuality regressors related to feedback were predicted all at once, but here we present one regressor for illustration. Decoding accuracy was then measured by Pearson correlation between model prediction and the real values.

2.5.3 EEG Results

A TUDA model with four states was trained to investigate the decodability of neural signals for task relevant regressors around the following events: patch choice, reward events and feedback (i.e. display of accumulated reward during the visit). Given possible different cognitive processes underlying reward events when participants decided to leave the patch or not, the reward events were modelled in two ways: one including all events (referred as the *all reward events* model); the other excluding events when participants pressed to leave (referred as the *passive reward events* model). The analyses focus on two aspects: (1) the duration of states indicating the temporal distribution of the model; (2) the states' relation to behaviour showing when and what information can be decoded from the brain signals. In particular, we aim to examine the following hypotheses:

- The representational states derived from the model are able to predict variability in task related regressors (e.g. patch values, RTs, choices) over time, even without any knowledge to the task structures or pre-specified time windows (i.e. a completely unsupervised learning).

- As patch choice and leaving decisions both involve state-change consideration (Kolling and O'Reilly, 2018), and the behavioural analyses revealed reward history effects on the leaving decisions (**Figure 20**). We therefore predict that the value history effects (which represent the quality of the overall environment or 'state') can be decoded from the brain signals around patch choice and leaving decisions.

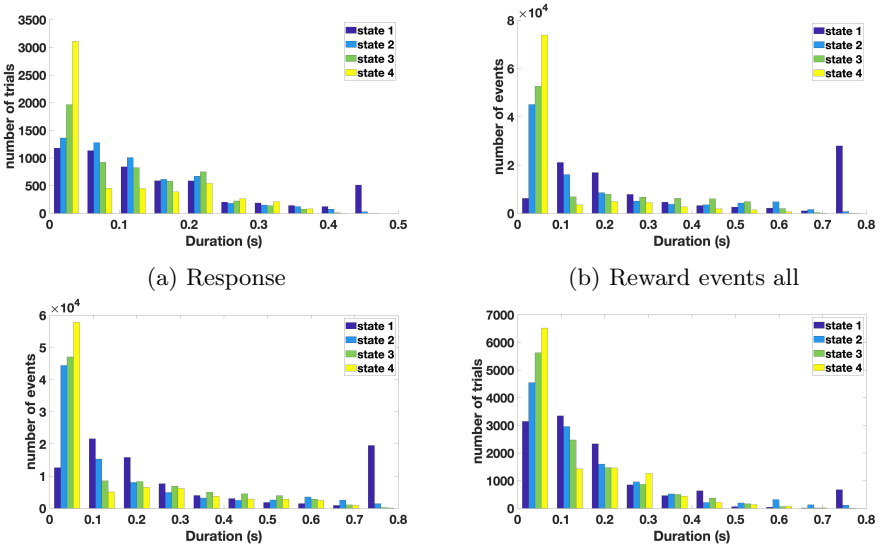
- As the behavioural analyses revealed value change effect on the leaving decision (**Figure 20**), we hypothesise that value change effects (i.e. 'reward changing trend' in **Table 5**) will be found in brain signals viewing reward

changes, indicating the value changes are used as a general monitoring.

- It has been found that firing rates in monkeys' dACC neurons increased in anticipation of smaller rewards and predicted monkeys' behaviour of abandoning the current patch for a new one (Hayden et al., 2011). We hypothesise that brain signals around feedback, which is pivotal in performance evaluation and decision adjustments, will be able to predict the next trial's behaviour (i.e. the next trial's switch and leaving time can be decoded from signals around the current trial's feedback).

Note that the following analyses will not address the spatial extents due to unreliable spatial interpretation of the decoding weights of the TUDA model (see Vidaurre et al., 2019).

2.5.3.1 Duration of states



(a) Response

(b) Reward events all

(c) Reward events derivative

(d) Accumulated reward

Figure 25: Duration of states plotted as a histogram for each event. In response-locked analyses, trials without active leave were removed.

We trained a TUDA decoder with four states (each a linear function mapping data to stimulus or behaviour) running at each time point across trials. All trials were assumed to start in state 1, and transitioned from state 1 to state 2, from state 2 to state 3 and then from state 3 to state 4. **Figure 25** presents the histograms of duration of states, revealing large temporal variability. Earlier states tended to last longer. Note that the states were inferred without prior knowledge of task timings (i.e. a completely unsupervised learning), and states 1-4 were trained within each event, not across four events.

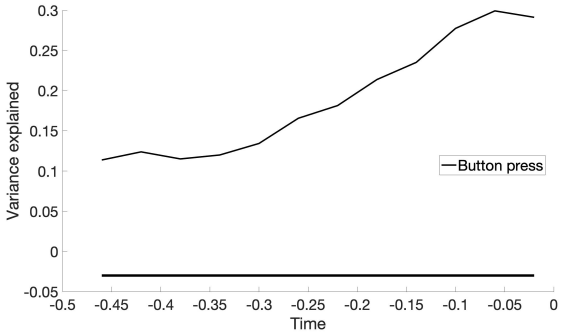
2.5.3.2 States' relation to regressors

Next we will present regressors the model predicted for each event, along with the decoding results for significantly predicted regressors, as well as the average occupancy showing the lifetimes and the temporal characteristics of the states. Note we ran the decoder with multiple regressors at once (all of the ones listed in the following tables) but only plotted the interesting ones for simplicity.

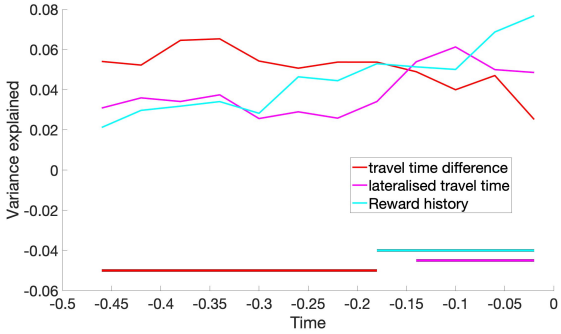
Patch selection and action selection

Behavioural analyses revealed that people picked the patch with higher average reward rate derived from a Bayesian Learner. We were interested in how the learning and evaluation of patch values as well as an actual response were decoded from EEG signals. We examined the decodability of EEG signals from -0.5 to 0 s prior to a response of patch choice with regressors listed in **Table 4**. The decodability was measured by explained variance in regressors by the neural data (R^2; cross-validated Pearson correlation between regressors and model prediction).

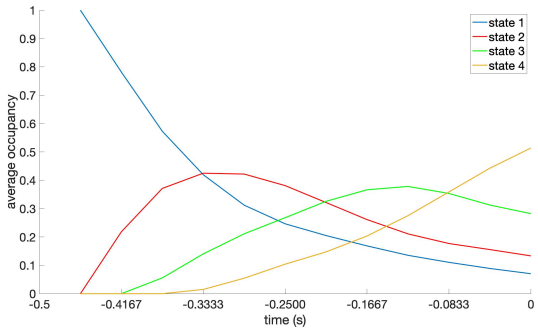
Figure 26 (a) and (b) shows the model successfully predicted response itself, travel time difference, average reward rate in history and lateralised travel time. Specifically, the neural signals accounting for chosen travel time occurred



(a) Response - right hand button press



(b) Response - travel time and history



(c) Response - states occupancy

Figure 26: Decoding results for patch choice. Model accuracy measured as adjusted cross-validated variance explained by response related regressors (R^2), plotted in separate for (a) the right hand button press, (b) travel time and reward history. Coloured bars at the bottom indicate the significant time period for regressors returned from permutation tests (i.e. when model accuracy > chance level accuracy). Prediction on the button press was significant all the time, whereas relatively early on chosen-unchosen travel time and late on lateralised travel time and reward history. (c) Average occupancy assigned to each decoding state as a function of time around patch choice. The event started exclusively in state 1, followed by 2, 3 and 4, with one state being dominant at a time.

Regressor	Description
Lateralised reward rate difference	lateralised maximum reward rate difference between chosen and unchosen patch derived from Bayesian learner
Reward rate difference	maximum reward rate difference between chosen and unchosen patch derived from Bayesian learner
Travel time difference	travel time difference between chosen and unchosen patch
Lateralised travel time	lateralised travel time difference
Reward history	average reward rate ignoring travel time derived from RL
LeavingTime	leaving time of the chosen patch
Button press	right hand button press

Table 4: List of patch choice related regressors examined by TUDA.

early, whereas for average reward rate and travel time related button press relatively late, nearly right before the action. Note the offset of chosen travel time matched the onset of average reward rate and travel time related button press, indicating participants first integrated the visually presented travelling time into abstract choice signals and then converted it into lateralised travel time (i.e. figured out what button to press based on their preference). Reward history was decoded from signals from -0.2 s until the response, implying choice vigor was influenced by historic information.

Average occupancy (percentage of trials assigned to each state) is shown by **Figure 26 (c)**. The peak of state 1 was found quite early (˜-0.45 s). The peak of state 2 occurred later (˜-0.33 s), followed by state 3 (˜-0.15 s) and 4 (˜0 s).

Reward events and patch leaving

Every time a reward changed in a patch was considered as a reward event. In order to evaluate reward events properly, participants had to take into account their context, such as the current reward rate, time spent in the patch,

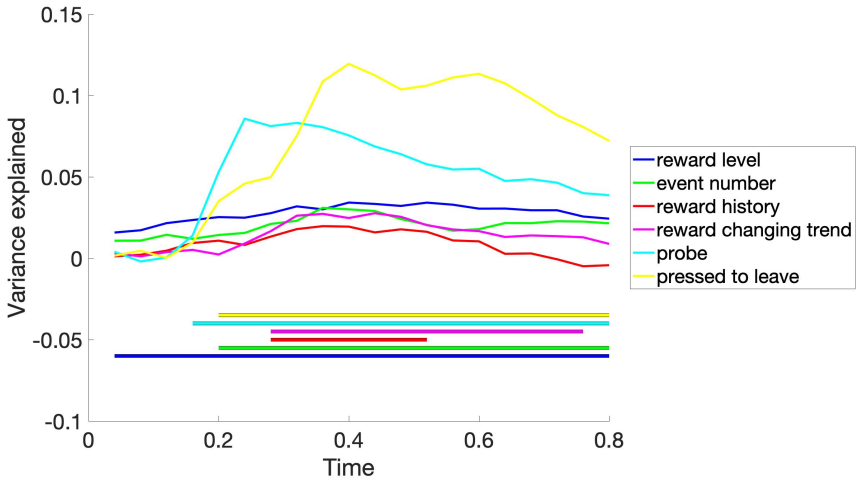
reward history and the local change from last reward event to current. We then examined the decodability of neural signals from 0 to 0.8 s around reward events and ran control analyses excluding leave decision themselves. To dissociate processing of reward events from patch leaving, we built two models using regressors specified in **Table 5**, with one including all events (the *all reward events* model) as well as regressors controlling for leave choices and probes, and the other model without the events when a probe was present or when participants pressed to leave (the *passive reward events* model).

Figure 27 (a) shows significant clusters of the experienced reward were found

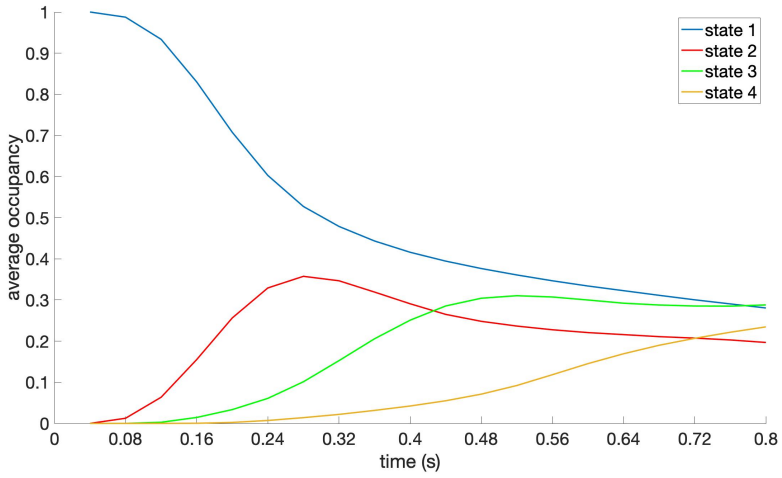
Regressor	Description
Reward level	experienced reward rate (i.e. current reward level)
Event number	number of reward events have been observed
Reward history	average reward rate ignoring travel time derived from RL
Reward changing trend	derivative of experienced reward at the probe which reflects the trend of reward changes and the decay speed
Probe	whether a probe occurred (i.e. an option of leaving was given)
Pressed to leave	whether pressed leave at the reward event

Table 5: List of reward events related regressors examined by TUDA.

from quite early (˜0 s), and event number and leaving choices (i.e. see probes or leave at the current reward event) from ˜0.2 s onward, as well as the reward changing trend from ˜0.3 s. Interestingly, when including actual choices, past reward rate was significant during 0.3 - 0.5 s, but no longer significant without leave choices (**Figure 28 (a)**). Also, the decision relevant signal was stronger when actual choices were included. However, it cannot be dissociated whether the difference was caused by merely having more data when including choices or whether history related signals were indeed stronger in choices. Average occupancy looks similar in both models (**Figure 27 (b)** and **Figure 28 (b)**),



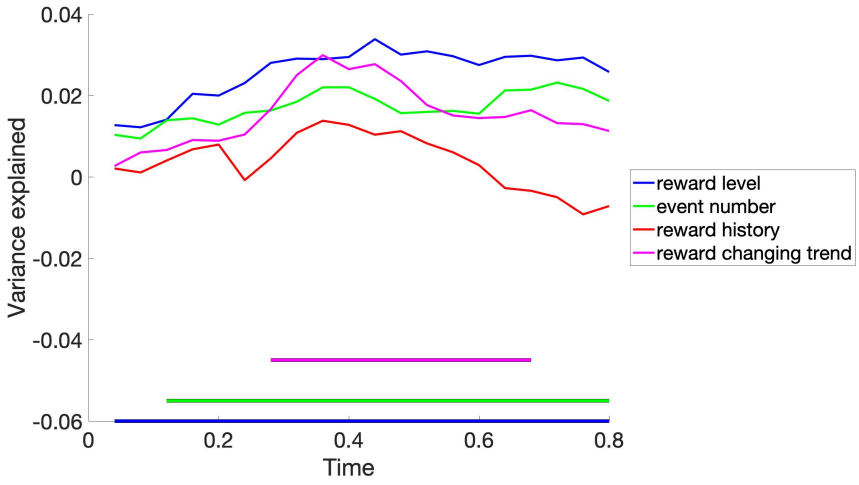
(a) All reward events - decoding accuracy



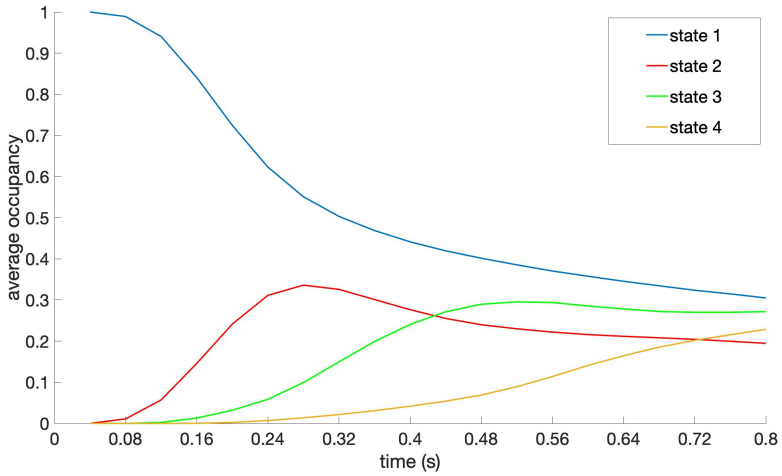
(b) All reward events - states occupancy

Figure 27: Decoding results for the reward events including leave choices. (a) Model accuracy measured as adjusted cross-validated variance explained by reward events related regressors (R^2). Coloured bars at the bottom indicate the significant time period of regressors returned from permutation tests (i.e. when model accuracy > chance level accuracy). Signals significantly predicted leaving when leave choices were included. Reward level, event number, probes and leaving were significantly predicted almost all the way. Reward history was significantly decoded from 0.3 s onwards and reward changing trend from 0.3 to 0.5 s. (b) Average occupancy assigned to each decoding state as a function of time around reward events. The peak of the other three states occurred at 0.28 s, 0.52 s and 0.8 s, respectively, but none of them exceeded the average occupancy taken by state 1.

where state 1 seemed to be dominant throughout the event but the other three states added up to account for more than 50% trials after 0.3 s.



(a) Passive reward events - decoding accuracy



(b) Passive reward events - states occupancy

Figure 28: Decoding results for the reward events excluding leaving. (a) Model accuracy measured as adjusted cross-validated variance explained by reward events related regressors (R^2). Coloured bars at the bottom indicate the significant time period of regressors returned from permutation tests (i.e. when model accuracy > chance level accuracy). Signals significantly predicted leaving when leave choices were included. Reward level and event number were still significantly predicted all the way, whereas reward history became non-significant and reward changing trend became weaker when leave choices were excluded. (b) Average occupancy assigned to each decoding state as a function of time around reward events. The peak of the other three states occurred at 0.28 s, 0.52 s and 0.8 s, respectively, but none of them exceeded the average occupancy taken by state 1.

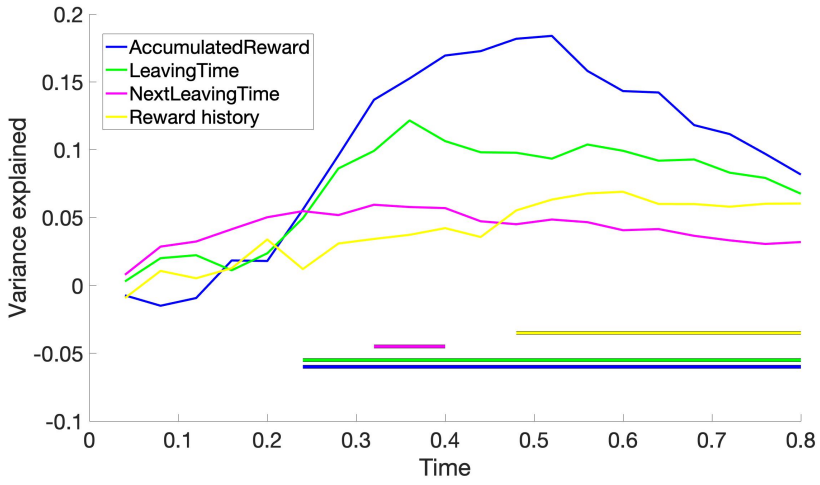
Accumulated reward and decision evaluation

To examine how the value signals looked like when the total reward was delivered as people evaluated the decision and planned for the next trials, we decoded neural signals from 0 to 0.8 s around the reward display. We posited neural signals around feedback would be related to feedback itself, overall stay in the patch, reward history, time (both travel time and time spent in the patch), whether to switch and how long to stay in the next trial (**Table 6**).

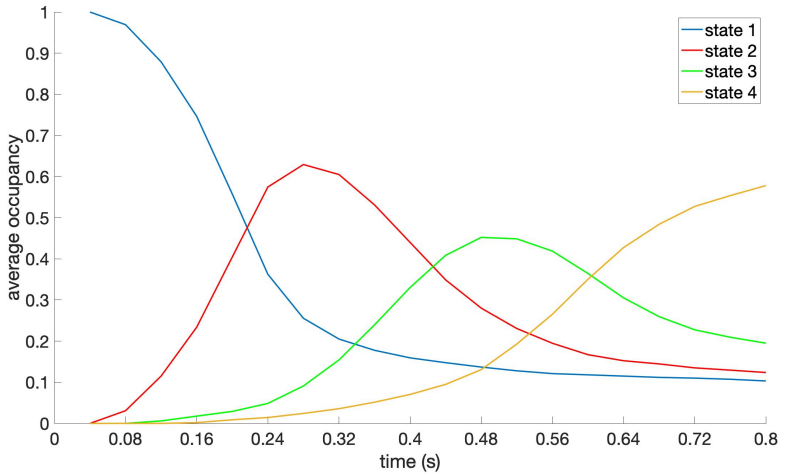
Regressor	Description
AccumulatedReward	overall reward collected in the visit
LeavingTime	overall stay in the patch
TravellingTime	travel time of the chosen patch
NextLeavingTime	leaving time in the next trial
NextSwitch	whether the other patch is chosen in the next trial
Reward history	average reward rate ignoring travel time derived from RL

Table 6: List of feedback related regressors examined by TUDA.

Figure 29 (a) shows the neural signals significantly predicted accumulated reward, leaving time, leaving time in the next trial and historic average reward rate. Specifically, accumulated reward and leaving time stayed consistently significant from 0.2 s onward, whereas average reward rate in history occurred from 0.5 s. Prediction on the next leaving time was only transiently significant (0.3 - 0.4 s) at a time period when a classic P300 is usually found. Average occupancy of states (**Figure 29 (b)**) shows an early peak of state 1 (˜0.08 s), followed by state 2 (˜0.3 s), 3 (˜0.5 s) and 4 (˜0.75 s).



(a) Feedback - decoding accuracy



(b) Feedback - states occupancy

Figure 29: Decoding results for feedback. (a) Model accuracy measured as adjusted cross-validated variance explained by feedback related regressors (R^2). Coloured bars at the bottom indicate the significant time period of regressors returned from permutation tests (i.e. when model accuracy > chance level accuracy). Accumulated reward and leaving time were significant from 0.2 s onwards, and reward history from 0.5 s, whereas leaving time in the next trial was significant between 0.3 and 0.4 s. (b) Average occupancy assigned to each decoding state as a function of time around feedback display. Average occupancy assigned to each decoding state as a function of time around feedback display.

2.5.4 EEG Discussion

The *Gold Rush* task investigates how people estimate reward structures, track changes and adapt decision strategies in a virtual foraging scenario. Behavioural analyses suggest decisions involved in this task were context-dependent and hierarchical (i.e. multiple levels of inferences and information processing), integrating patch values, reward changes and quality of the overall environment. Therefore, we sought to characterise neural representations of abstract, task relevant features by decoding EEG signals with the TUDA method, a special type of HMM model that was used to reveal hidden states underlying sequences of data with temporal variations across trials (Vidaurre et al., 2019).

We decoded travel time, followed by reward history and button press from signals around the patch choice, suggesting people integrated travel time into choice value, compared against the overall environment (derived from reward history) and translated the choice value into a specific button press. However, the regressor "maximal reward rate derived from Bayesian" (i.e. the regressor that categorised each patch's maximal reward rate computed from DR and SF derived from the Bayesian model) was not decoded from brain signals, whereas the behavioural results which showed the regressor most explained patch choice were not reflected by strong EEG signals at that point. A plausible explanation is that the average reward rate, as an estimate of the overall quality of a patch, was integrated from experience and computed before patch presentation, rather than a temporary variable suddenly came across when people saw the patches, and therefore the signals were overshadowed by the most explicit information participants saw during patch presentation – the travel time.

When including leave choices, we decoded reward level, event number, reward history, reward derivative, probes and leave choices from signals around reward events. These suggest reward events were not evaluated alone, but trans-

lated into a comparison between the current environment and the overall environment and eventually to the leaving decision. The results are aligned with previous findings that dACC monitors state changes and decides to switch or adjust behavioural strategies in foraging scenarios (for example, Meder et al., 2017, Kolling et al., 2012; reviewed by Kolling and O'Reilly, 2018). When taking out leaving, only reward derivative was decoded in addition to reward level and event number, while traces of reward history were no longer found. It seems reward derivative as an estimate of the speed of reward decay and the quality of the current environment was considered when people processed reward events, whereas reward history was mostly taken into account when participants considered leaving. However, our results do not rule out the possibility that analyses including leaving simply contained more data and yielded on average stronger signals. Although we did not establish neural signatures with ERP analyses (as this is not the research purpose of this study and it is difficult to unequivocally read out the interpretation of recordings from a single electrode given rich information processing involved), the results that leaving decisions seemed to be made with comparing the current environment and the overall environment are consistent with previous findings showing that rewards are evaluated in relation to expectations regarding FRN (a negative deflection occurring from the frontocentral scalp around 200-300 ms post feedback; more negative after negative than positive feedback) (Chase et al., 2011; Bellebaum et al., 2010; Holroyd and Krigolson, 2007;) (for review, see Holroyd and Coles, 2002; Ullsperger et al., 2014; Cohen, Wilmes et al., 2011), which has been suggested to scale with the reward prediction error, and P300 (a positive component emerging from the centro-parietal area around 300-400 ms post stimulus presentation) (initially found by S. Sutton et al., 1965), which has been suggested to encode uncertainty or surprise (Duncan-Johnson and Donchin, 1977), in context updating

(Donchin and Coles, 1988) and response initiation (Nieuwenhuis et al., 2005). However, there is no consensus on the exact functional differences between the two components. Future research may address (1) if the formulation of leaving decision by comparing between the option at hand and the overall environment relates to P300 (as its previously suggested role in updating context and adjusting behaviour) and (2) if FRN scales with the prediction error derived from the RL model used in our scenario.

We decoded accumulated reward and time spent from signals around feedback, suggesting relevant signals were represented and potentially integrated immediately into one signal taking reward amount and time spent into account, as an overall performance evaluation. Note that travel time was not decoded at the point of feedback delivery, showing the average reward rate was computed based on only in patch time not the travel time and indicating travel time is not a simple passage of time signal but instead considered as the relevant time only when contextualising reward (i.e. during patch choice). As predicted, we decoded the next trial's leaving time from the feedback-locked signals, indicating the reward modulated the overall commitment in a patch. The results resemble previous findings that the firing rates of dACC neurons in primates' brain seem to be associated with reward monitoring and behavioural adjustment (the firing rates increased as anticipating smaller rewards and switching to a new patch) in a foraging scenario (Hayden et al., 2011), and the reward representations in human dACC predicted switching behaviour as a function of the varying reward rates (Wittmann et al., 2016). Our results extend the finding by pinpointing the exact timing of such behavioural adjustment might take place (i.e. 0.3-0.4 s post feedback). Furthermore, the time period when the next trial's leaving time was decoded roughly matched the same time period when classic P300 was normally found (S. Sutton et al., 1965). As P300 has been believed to re-

flect motivation-guided information processing (Nieuwenhuis et al., 2005), our results suggest that participants might form a commitment of when to leave or change vigour at the feedback for the next trial. In addition, reward history was decoded from 0.5 to 0.8 s, indicating participants may first compute the time discounted value of a patch and then integrate it into the long term estimate several hundred milliseconds later.

The TUDA model estimated the exact timing of the states (i.e. at which time points each state activated). The four states revealed a plausible sequence of active representations of task related information behind observed behaviour, such as the choice values over and above the chosen action. Note that decoding analysis does not serve a purpose as post hoc explanations, for the decoding models in general do not support unambiguous causal statements (Weichwald et al., 2015), and therefore we do not interpret the functional significance of each state in separate, but focus on the model's predictive power in performance, which shows what information can be read out from the brain signals (i.e. temporally overlapping decoding indicates the underlying neural activity multiplexes coding of the factors). We are particularly interested in establishing the source of performance differences with simultaneously evolving neural representations. We show that EEG decoding analysis can separate effects of low level stimulus/response from those of abstract, context-level task representations (the fact that performance can be decoded from EEG signals showcases the effects can be separated). It is admitted that an arbitrary number of states was chosen. However, there is no 'correct' number of states, and what matters is to identify a number that describes the dataset in a plausible manner (as discussed in Quinn et al., 2018). In addition, as it is not straightforward to causally pinpoint spatial sources from decoding weights (as mentioned in Vidaurre et al., 2019 and a further discussion in Weichwald et al., 2015), we do not seek to estab-

lish spatial signatures but rather focus on the temporal dynamics in our study. Future research may resort to encoding models (as suggested by Vidaurre et al., 2019 and Myers et al., 2015) or source reconstruction methods to examine the spatial extents of the regions involved.

Taken together, the decoding results revealed not only low-level processes, such as button press in patch choice, reward level and event number in reward events, as well as accumulated reward in feedback, but also some abstract, high-level features. Reward history seemed to play a role in all three events, highlighting the importance of a high-level representation of the overall environment throughout the task. The results complement previous findings that dACC neurons integrate information over multiple actions, suggesting dACC is linked to the processing of reward history, integrating past information to formulate representations of the overall environment, and tracking environment changes as well as facilitating decision adjustments (Meder et al., 2017; Wittmann et al., 2016; reviewed by Kolling and O'Reilly, 2018). We speculate that the state monitoring of dACC took place the same time when reward history was decoded (for example, -0.18-0 s prior to response, 0.3-0.4 s post reward events and 0.5 s post feedback), and such temporal dynamics are highly context-dependent (i.e. specific to decisions).

Furthermore, the fact that people compared the current environment with the overall environment are in accordance with behavioural results, as well as the leaving model proposed by Marginal Value Theorem (i.e. a predator leaves when the marginal capture rate within the patch diminishes to the average capture rate of the overall environment; Charnov, 1976), and the exploration-exploitation trade-off found in previous studies with human participants and primates (Hayden et al., 2011; Daw et al., 2006; Kolling et al., 2012). This finding also fits in the cognitive processes underlying *state-change decisions*

proposed by Kolling and O'Reilly, 2018 that when people leave the current state, they compare the expected values of the entire environment with the average reward rate within the current state based on past experiences. Neural representation of reward derivatives corroborates the previously found evidence that leaving the current environment is guided by a good-bad evaluation of the current environment based on reward trends, a process associated with dACC activity (Wittmann et al., 2016). We used EEG based, timepoint-by-timepoint decoding analyses to show neural signatures of both low-level stimulus/response and high-level abstract task representations. Our analyses mainly focused on relating temporal dynamics of neural signatures to decision processes, and future research with fine source localisation techniques may further tackle the mechanisms of specific brain areas involved in the cognitive processes suggested by our study, especially relating to previously found role of vmPFC and ACC in evaluating and searching alternatives (Kolling et al., 2012), as well as the role of dmPFC, including preSMA, in tracking changes and deciding to explore (Kolling and O'Reilly, 2018).

2.6 Conclusion

The current research examined humans foraging decisions with a virtual foraging task *Gold Rush*. The results have demonstrated people picked and left a patch that maximised average reward rate as predicted by MVT, weighing reward rates within a patch against the overall environment, and people were able to track changes and adjust their estimates accordingly. Computational models suggested that people chose a patch with higher average reward rate, and the reward history was integrated to formulate an overall evaluation of the environment. Travel time was used to contextualise reward during patch selection, but not integrated into the overall quality of the patch. We found neural representations of reward history in patch choice, leaving decision, and feedback processing, highlighting the importance of 'learning from the past' in foraging scenario. We also found neural signatures of long term commitment of a patch modulated by current outcomes. Both behavioural models and decoding analyses showed travel time was not used to determine leaving or to compute average reward rate for performance evaluation, but considered as relevant only to contextualise the reward during patch choice. Taken together, the study shows foraging decisions are not guided by a simple good-bad judgement or a simple calculation of expected values, but by high-level cognition integrating past information, inferences of reward structures, track of changes as well as predictions for future.

3 Study 2: Metacognition

Metacognition describes a type of hierarchical perceptual decision making involving two levels of information processing: a low level perceiving and categorising physical properties of the world, and a high level allocating attentional resource to preferential stimuli. This study probes into cognitive processes in metacognition with a hierarchical visual perception task, in which participants choose to report the rotation of a target patch based on their visual perception and inference of the reward contingency. Here, we report people are able to pick the target patch according to choice confidence and integrated experience across trials. In addition, the hierarchical information processing follows simple probabilistic inference described by a Bayesian Learner. We also report a series of neural signatures underlying key cognitive processes from EEG signals. We show enhanced selective attention and anticipatory attention related to accuracy, cognitive effort and switch readiness in the task. We report links between response readiness potential and RT, as well as between beta desynchronization and response threshold setting process. We suggest the connection between feedback related negativity and reward, switching and task difficulty might fit into a learning-unlearning framework. Combining EEG signals and Bayesian Learner, we propose neural evidence for hierarchical decision processes integrating choice confidence and learned experience.

3.1 Introduction

Perceptual decision making is not uncommon in daily life. Is that mushroom poisonous? How fast is that car moving? Where does that noise come from? Which way is it to the bathroom? This type of decision includes integrating sensory inputs, such as the sights, the sounds and the smells, into categorical judgements (Gold and Heekeren, 2014). Perceptual decision making has been studied by two broad classes of methods: the processing of sensory evidence (i.e. accuracy increased as signal strength, given individual sensitivity and threshold) modelled by *psychophysics*; the categorical judgement (e.g. *'is the signal really there given the perceived evidence'*) modelled by probabilistic inference (also called *ideal observer theory*) (Gold and Heekeren, 2014). However, real life decisions are not always a simple sensory perception. Purcell and Kiani, 2016 proposed a hierarchy of interacting decision processes: a high-level strategy guides overall information processing and expects outcomes, and a low-level choice processes stimuli and executes actions.

Neural correlates have been established with a combination of perceptual tasks and single cell recording on monkeys and rodents, revealing a link between choice confidence and LIP, as well as between motion direction and MT/V5, LIP, SC, FEF and dlPFC (reviewed by Gold and Shadlen, 2007). In the human brain, dlPFC and ventral temporal cortex activities together predicted performance in a face/house recognition task (Heekeren et al., 2004). Neuroimaging studies on humans also suggested a wide range of brain areas, such as pgACC (Bang and Fleming, 2018), striatum (Hebart et al., 2016), medial prefrontal cortex (Lebreton et al., 2015; Martino et al., 2013), dorsal anterior cingulate cortex (Fleming et al., 2012), and rostro-lateral prefrontal cortex (Martino et al., 2013; Fleming et al., 2012) related to choice confidence. Our study does not aim to contrast previously established brain areas in relation to choice confidence. We

do not explicitly measure confidence in our task, but seek to investigate how the reliability of perceptual evidence, which is a key component of choice confidence and reversely proportional to task difficulty, influences participants' performance evaluation and decision adjustments. To distinguish such confidence from the explicitly measured choice confidence in previous studies, we refer it as 'accuracy expectation' and 'reward expectation', depending on if it relates to task difficulty itself or both difficulty and reward contingency. We hypothesise that accuracy expectation declines linearly with increasing task difficulty, whereas reward expectation takes into account both difficulty and tracked reward structures.

Several EEG components have been suggested to relate to key stages of perceptual decision making. Alpha oscillations (˜10 Hz) have been suggested to reflect attentional control, especially the anticipatory attention (Klimesch, 2012). Specifically, alpha power increases in anticipating an distractor and decreases in anticipating the target on the contralateral hemisphere relative to the attended side, and the modulation of alpha power correlated with performance (Worden et al., 2000; Thut et al., 2012; Kelly et al., 2006; Händel et al., 2011; Bonnefond and Jensen, 2012). Another EEG component that has been found to indicate visual attentional control is the visual-evoked potential (VEP), a type of electrical potential recorded over the visual cortex following visual stimulus (reviewed by Creel, 2019). The component has been suggested to link to selective visual attention devoted to visual stimulus (Hillyard et al., 1998; Störmer et al., 2009). A cued visual attention task identified VEP lateralisation in parieto-occipital channels at an early time period (80-120 ms post-stimulus) and a late period (180-300 ms post-stimulus) in both overt and covert orienting conditions, whereas correlation with the lateralisation and visual sensitivity was found only in the covert attention condition, but not in the overt one (Hilo-Merkovich et al.,

2018). An earlier study with a similar procedure except using a unilateral cue suggested attentional modulations during the 150-200 ms and 250-330 ms time periods, but not in an early period 100-130 ms (Eimer et al., 2006).

We are also interested in the response related components, especially brain signals reflecting how multiple sources of information (e.g. difficulty, stay/switch) are integrated and converted to a response. Lateralised readiness potential, a negative potential observed in the motor area contralateral to the responding hand has been identified to reflect response activation processes (reviewed by Kappenman and Luck, 2011). A study combining drift diffusion model with a neural network model found LRP correlates with the drift rates, indicating an early evidence accumulation process, and that LRP is linked to the threshold layer, suggesting a late, ballistic response activation ('motor') process (van Vugt et al., 2014). In addition, beta oscillations (15-29 Hz) have been found to correlate with movement-initiation and cancellation (J. Wessel, 2020), as well as response threshold (Fischer et al., 2018), and might be associated with task difficulty, accuracy and assertiveness of response.

Feedback processing is particularly important in this task. We seek to identify neural signatures of how participants adjust their strategy based on task difficulty, their own performance as well as a history of feedback. Previous studies have identified two EEG components in relation to feedback processing. A negative deflection, feedback-related negativity (FRN) occurring 200-300 ms post feedback, has been found to reflect good-bad evaluation of choices, reward prediction errors and learning (reviewed by Ullsperger et al., 2014; Cohen, Wilmes et al., 2011). P300, a positive component occurring 300-400 ms post feedback around the centro-parietal area has been suggested to reflect the processing of motivationally significant events (first identified by S. Sutton et al., 1965; reviewed by Nieuwenhuis et al., 2005), as captured by the context

updating theory (Donchin and Coles, 1988) or the stimulus evaluation theory (Duncan-Johnson, 1981).

The current research examines hierarchical perceptual decision making with a visual perception task. Participants are instructed to select a gabor patch out of two and report its rotation direction between two presentations. One of the patches has a deterministic relationship between accuracy and reward whereas the other patch yields random outcome. Reward contingency switches between the two patches after a certain number of trials. Participants can use reward history to figure out which is the target patch – but the unusual feature of our task is that reward history should be weighted by participants' evaluation of their own performance, based on task difficulty (which will be related to perceptual confidence). Because some trials are harder than others (smaller rotation), reward omission on these trials is less valuable evidence for a side reversal than it would be on an easy trial. We present results in terms of behaviour (accuracy and reaction times) and EEG signals. Additionally, two models are used: 1) a drift diffusion model to characterise behaviour (do individual decision components, decision boundary, drift rate and non-decision time, vary with task difficulty and intention to switch patches?), and 2) a Bayesian inference model is used to characterise confidence in the target patch – this is used in the analysis of EEG responses to feedback.

Specific hypotheses for each analysis are set out in the relevant results sections.

We also seek to reveal neural signatures underlying key cognitive stages in this task with EEG, such as the attentional control, response preparation and execution, as well as the error detection and performance evaluation (based on prior literature reviewed above). Especially, we will try to answer the following questions, which are set out in more detail, with predictions, at the start of the

relevant results sections:

- Can we establish lateralised visual-evoked potentials (lVEPs) around the stimulus presentations? Do lVEPs reflect perceptual accuracy and the cognitive effort devoted into the task? Are lVEPs modulated by switch?
- Can we establish alpha desynchronisation when participants anticipate the onset of stimulus and try to direct their spatial attention? Is alpha desynchronisation modulated by accuracy, difficulty and switch?
- Can we establish response readiness related signals (e.g. lateralised readiness potentials (LRPs))? Do LRPs reflect response speed and switch?
- Can we establish beta oscillations as indices of response preparation? Do beta oscillations reflect response speed, difficulty and switch?
- Can we read out from FRN and P300 components how people learn from feedback? Is feedback processing modulated by difficulty, accuracy, switch and prior expectations derived from Bayesian model?

3.2 Materials and Methods

3.2.1 Participants and Procedure

Thirty-six healthy right handed participants (aged 18-40 years) took part in the study. Participants were recruited from a student population at the University of Oxford and all gave written consent using a form approved by the ethics committee. Participants conducted a behavioural training at least one day before the experiment and were only invited to take part in the experiment had their performance showed they were able to track the reversals of reward contingency. All participants completed the total 360 trials of the experiment. Nine participants were excluded from analyses due to frequent eye movements (in over 25% trials their gaze deviated from 200 pixels on the x-axis from the screen centre), resulting in a final sample size of twenty-seven. Participants received reimbursement for their time and performance upon completion of the experiment.

3.2.2 Behavioural Task

Participants performed a task in which they adapted their decision strategy according to unpredictable changes in the environment (**Figure 31**). Specifically they viewed, in parallel, two patches on the left and right of the screen. The patches rotated and participants reported the rotation of *one* of the patches, with a button press. Crucially, only one of the patches was relevant to a reward. Correct response on the relevant patch led to a certain delivery of reward, whereas incorrect response led to no reward, and response on the irrelevant patch regardless of its correctness led to a 50% chance. Therefore, receiving a reward or not became the sole feedback and the decision strategy included both attending to the correct patch and choosing the correct rotation direction. Changes in environment (i.e. reward relevance of a patch) were reversed on average every

10 trials (randomly drawn from 5 to 15 trials), without any reversal in the first 5 trials (as shown by **Figure 30**), and the perceptual noise (i.e. the rotating degree) varied from trial to trial and thus so did participants' confidence that they were correct. Participants should take into account that confidence, plus how many other trials were recently unrewarded, to decide whether a reward omission was due to perceptual error, or an actual reversal in which was the relevant patch.

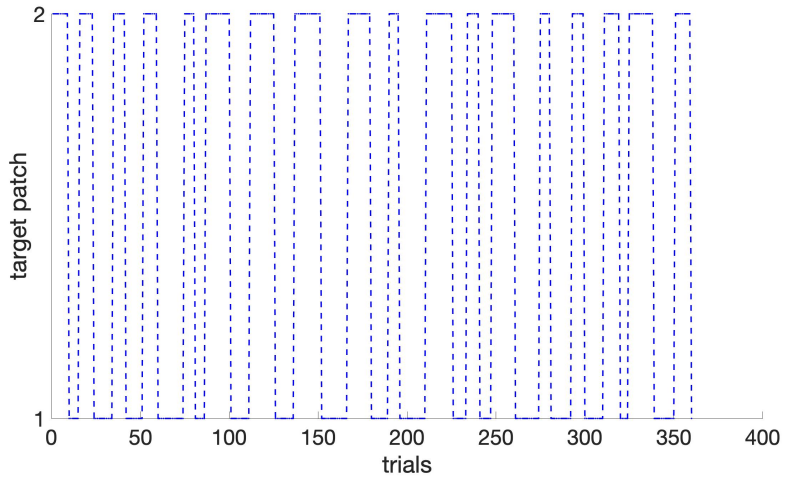


Figure 30: Distribution of reward relevance illustrated by one participant's task schedule. The reward relevance changed from one patch to the other every 5 to 15 trials (uniformly distributed with a mean of 10 trials).

The behavioural task was programmed and delivered using Psychtoolbox in MATLAB. Visual stimuli were presented on a 24-inch (61.1-cm) monitor with a refresh rate of 60Hz. Each trial began with a fixation point presented for 0.5s. After that the fixation point shrank and a pair of Gabor patches was presented for 0.5s. The patches were tilted either clockwise or counterclockwise with an angle randomly from 30 to 60°. Afterwards, an inter stimuli interval (ISI) was given in the form of a blank screen lasting between 3 to 5s. A second fixation point then appeared and lasted for 0.5s indicating the onset of the

second stimuli presentation. The pair of patches was presented again but rotated with an angle randomly drawn from [0, 3, 6, 9, 12]°. Zero degree means no rotation. Participants were allowed to make a response as soon as the second stimulus appeared by pressing a button with left index finger (left patch rotating counterclockwise), left middle finger (right patch rotating counterclockwise), right index finger (left patch rotating clockwise) or right middle finger (right patch rotating clockwise). Participants were instructed to respond on the patch they believed to be deterministic to the reward. The second stimulus pair remained on the screen until a response was made or for maximum 1.5s. After that, a jittered delay was given in the form of a blank screen from 0.3 to 0.8s. Participants then received a feedback of either reward or no reward, indicated by a blue or yellow dot in the centre. The mapping of colour and reward type was counterbalanced across subjects. The feedback was presented for 1s with patches remaining on the screen, followed by an inter trial interval (ITI) jittered between 3 to 5s. Each participant completed 3 sessions of 120 trials with self-paced break in between, resulting in 360 trials in total per subject. Across trials the reward contingency (i.e. which patch was deterministic to reward) reversed every n trials, with n randomly drawn from 5 to 15. An overview of the task structure is shown by **Figure 31**.

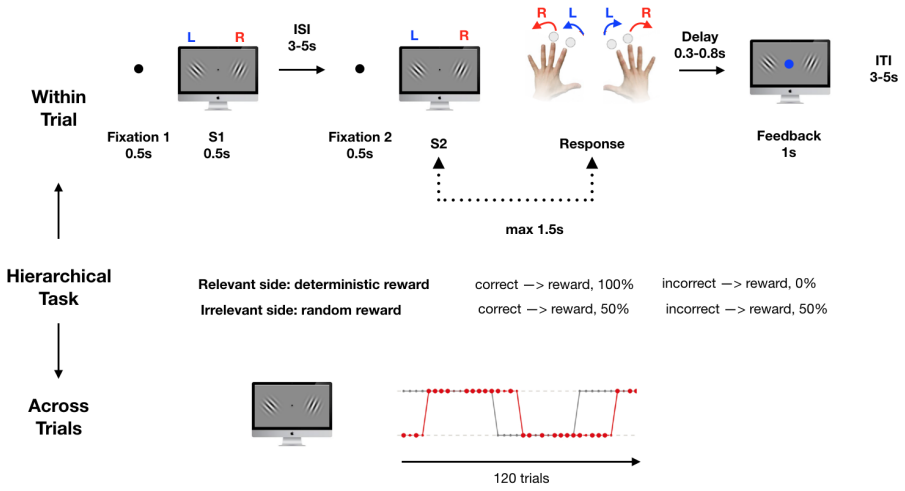


Figure 31: Overview of the behavioural task. Participants were required to track the rotation direction of the patch deterministic to reward. The rotating degree as a source of perceptual noise formulated the first level task and varied randomly from trial to trial. The reward relevance of the patch as an actual reversal formulated the second level task and was reversed every few trials across the total of 120 trials.

3.3 Behaviour

3.3.1 Behavioural Results

Effects of perceptual difficulty

To verify that participants were sensitive to our difficulty manipulation (differences in degree of rotation), we checked that accuracy increased as difficulty decreased. The behavioural results revealed that in general people were able to track the rotation and their performance improved as the difficulty decreased (**Figure 32(a)**) [logistic regression of probability of responding 'clockwise' on angle of rotation: group t statistic = 17.2169, p = 9.8251e-16]. Reaction time decreased as perceptual difficulty decreased **Figure 33**.

If participants took the perceptual difficulty into account when deciding whether a negative feedback indicated a reversal (rather than perceptual error), then given that an un-rewarded trial occurred, participants should be more likely to switch sides on the next trial if the unrewarded trial was easy. This was indeed the case (**Figure 32(b)**) [linear regression of probability of switching, given trial unrewarded, on absolute angle of rotation: group t statistic = 7.3977, p = 7.4306e-08], stronger in participants with better performance on the perceptual task (those with a steep sigmoid function in **Figure 32(a)**)but the results revealed no such effect (**Figure 32(c)**) [linear regression of beta value from analysis b on beta value from analysis a; group t statistic = 2.1508, p = 0.27706].

Cross-trial effects on switching

The decision to switch sides appears to build up over multiple trials. Firstly, people show some sensitivity to time (trials) since last switch: given that a trial was unrewarded, the probability that participants switched built up the more trials had elapsed since they last switched sides (**Figure 32(d)**) [t = 4.1979,

p = 0.00027857, linear regression]. In contrast, it was *not* the case that the probability of a ground truth reversal increased with number of trials since last behavioural reversal (blue line, **Figure 32(d)**), due to the distribution of reversal, combined with the fact that behavioural reversals didn't map exactly onto ground truth reversals but followed with a (variable) few trials' delay. Note that there was a ceiling at 15 trials, as a reversal happened at maximum every 15 trials.

Secondly, people integrated multiple trials' outcomes in the decision to switch sides. Participants were more likely to switch following an unrewarded trial, if recent trials had also been unrewarded. A logistic regression predicting the probability of switch, given that trial t was unrewarded, showed significant effects of reward omission at trials *t-1* and *t-2* but not further into the past [-4: t = 1.1922, p = 0.24396; -3: t = 3.7204, p = 0.001; -2: t = 10.0482, p = 1.918e-10, multiple logistic regression]. Previous trials had an effect on the probability of switch, with more recent trials having bigger impact (**Figure 32(e)**).

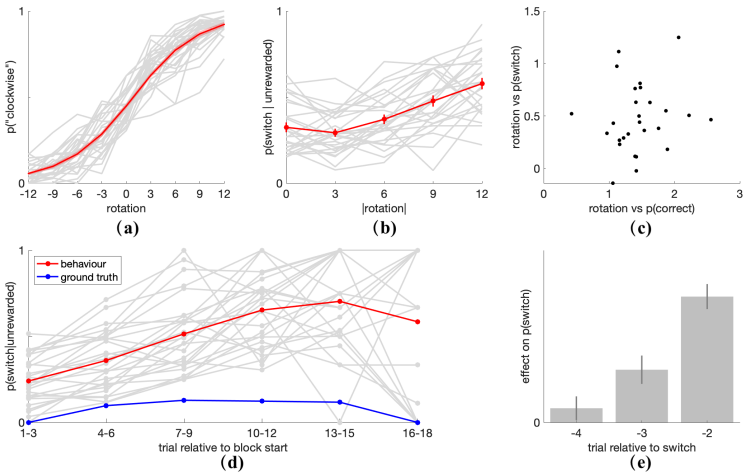


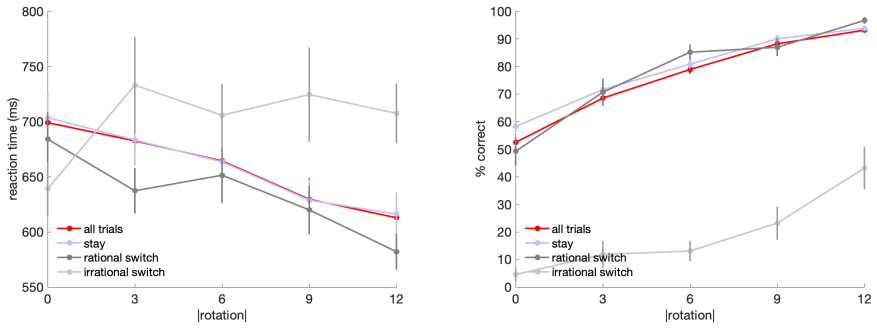
Figure 32: Behavioural results. (a) Probability of responding clockwise as the rotation angle varied from more anticlockwise to more clockwise. Individual data were plotted as grey lines and the mean was shown by the red line. (b) Probability of switch given unrewarded against different difficulty levels. (c) Correctness predicted by difficulty plotted against switch predicted by difficulty. Each dot represents the data point from one participant. (d) Probability of switch given unrewarded across numbers of trials relative to the block start. (e) Weights of previous 2, 3, and 4 trials on switch.

Behaviour around a switch point

We were particularly interested in behaviour before and after a switch decision. One hypothesis would be that switch was attributed to the reduced commitment to the current patch, which led to reduced concentration (reduced accuracy, and faster reaction times). On the other hand, switch may enhance concentration as people devoted more effort to the task when gathering and weighting evidence that suggested to leave the current patch (increased accuracy and slower reaction times). To investigate these bidirectional possibilities, we next examined behaviour in terms of reaction time and accuracy the trials on which participants switched sides. Additionally, we split side switches into rational and irrational switches. Rational switch was defined as switch when an actual re-

versal has happened (i.e. people switched from the ground-truth irrelevant side, to the ground-truth relevant side) and irrational switch otherwise (i.e. people switched to the irrelevant side). We hypothesised that people in general were more concentrated in the task when making rational switch, whereas irrational switch was probably due to poor concentration or action mistakes. Such a split allowed us to test our hypotheses by looking at RT and accuracy differences between different switch types.

Reaction time and accuracy displayed similar effects in stay and rational switch trials (**Figure 33**). Rational switch was defined as switch when an actual reversal had happened (people switched when they should) and the irrational switch otherwise. People responded faster and more accurately as difficulty decreased in stay and rational switch trials. In irrational switch trials, however, RT did not change as difficulty and in general was longer. Although responses in irrational switch trials were slightly slower compared to the other two, and responses in stay were also slightly slower than rational switch, the difference between the three was not significant ($F(2, 26) = 1.13$, $p = 0.3285$). Accuracy in irrational switch trials was significantly lower than the other two ($F(2, 26) = 181.96$, $p = 6.65871e\text{-}27$), and there was no difference between stay and rational switch (multiple comparison test: $p = 0.9020$). Accuracy slightly climbed in easy trials but the overall accuracy was much lower than in other trials, indicating irrational switch was probably driven by sampling bias and reduced concentration.



(a) reaction time

(b) Percentage of accuracy

Figure 33: Reaction time (a) and percentage of accuracy (b) in all, stay and switch trials. Switch trials were split by rational switch (an actual reversal had happened) and irrational switch (no reversal).

As switch evidence was accumulated across multiple trials **Figure 32(e)**, we examined how RTs and accuracy built up around a switch trial. Mean reaction time was approximately 660ms across all trials. Reaction time decreased one trial before a switch and increased one trial after (**Figure 34(a)**). Mean perceptual correctness was slightly above 0.8 across all trials. We defined a perceptually correct trial as one on which the participant reported the correct rotation direction, regardless of whether the relevant side was chosen - therefore some perceptually correct trials were unrewarded (if the irrelevant side was chosen). Perceptual correctness also decreased one trial before a switch but rose again and remained relatively stable after a switch (**Figure 34(b)**). Probability of reward was about 0.8 across all trials, plummeted to almost zero one trial before a switch and stayed at around 0.7 after (**Figure 34(c)**).

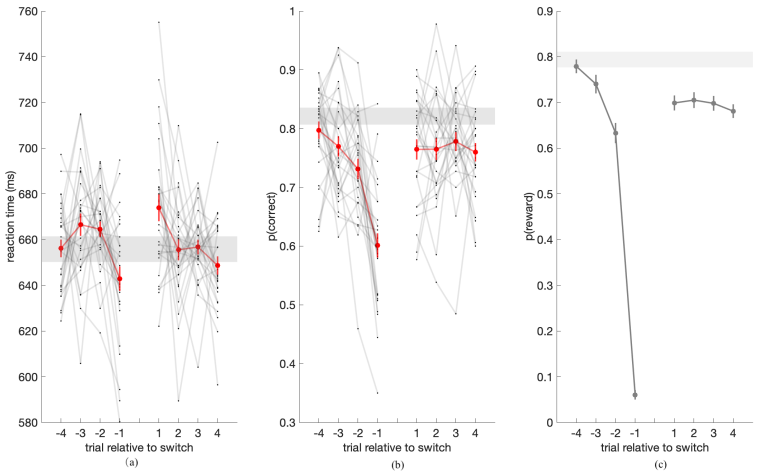


Figure 34: Behavioural results around all switch points. (a) Reaction time plotted against trial relative to switch. The red line shows mean RT of the corresponding trial and the shaded bar represents the baseline RT (averaged over all four trials before and four trials after a switch). (b) Probability of correctness plotted against trial relative to switch. (c) Probability of rewarded plotted against trial relative to switch.

As **Figure 33** suggests notable behavioural differences between rational and irrational switches, we then split switch trials as shown by **Figure 35 and Figure 36**. Reaction time was shorter before a rational switch (**Figure 35(a)**), but no such change before an irrational switch (**Figure 36(a)**), and responses slowed down after both rational and irrational switch. Perceptual correctness was slightly higher than average before a rational switch (**Figure 35(b)**), but plummeted almost to 0.1 before an irrational switch (**Figure 36(b)**). In both cases, accuracy remained relatively stable after a switch, with irrational switch slightly below average but slowly climbing to the mean. Probability of being rewarded massively dropped in both cases (**Figure 35(c) and Figure 36(c)**), and the reward probability matched accuracy after a switch (relatively stable after rational and rising up after irrational switch). Note that in the irrational switch case, trials started being unrewarded two trials before, indicating that

irrational switch was driven by the reward history at least a bit.

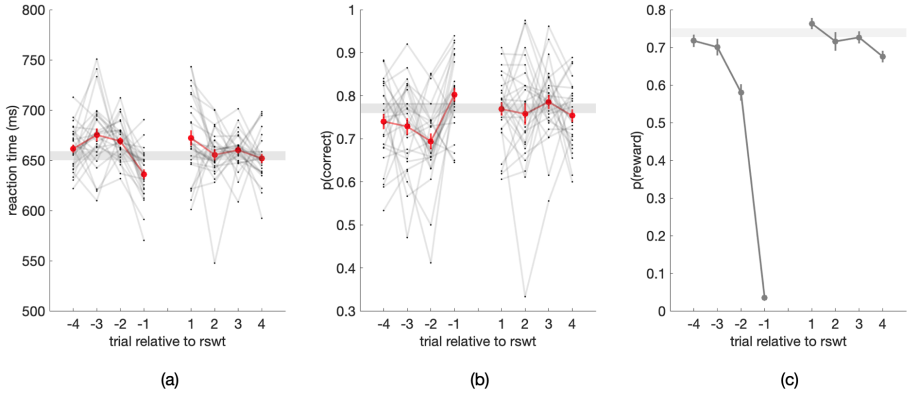


Figure 35: Behavioural results around rational switch points. Reaction time (a), chance of perceptual correctness (b) and chance of being rewarded (c) plotted relative to a rational switch.

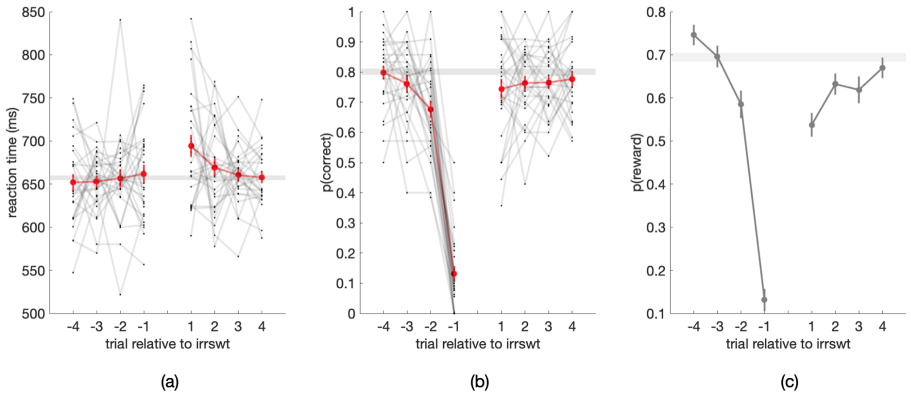


Figure 36: Behavioural results around irrational switch points. Reaction time (a), chance of perceptual correctness (b) and chance of being rewarded (c) plotted relative to an irrational switch.

To rule out the possibility that the longer RT in irrational switch trials was caused by outliers (i.e. abnormally slow trials that deviate from the average), we plotted distributions of RT in three types of response (**Figure 37**). All three types displayed similar shape, with rational switch trials having higher proportion of fast responses and irrational switch the opposite, but none of the

conditions revealed any abnormality. Therefore, the average RT shown before revealed some general response patterns.

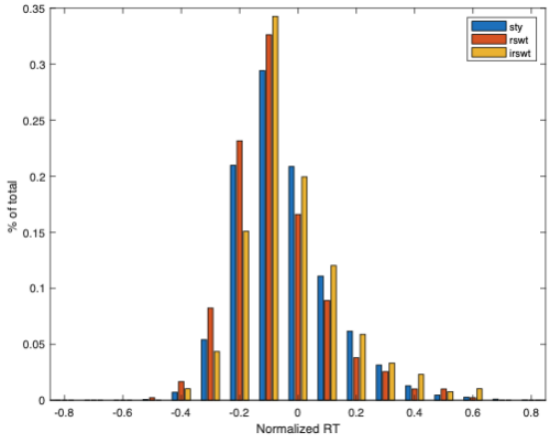
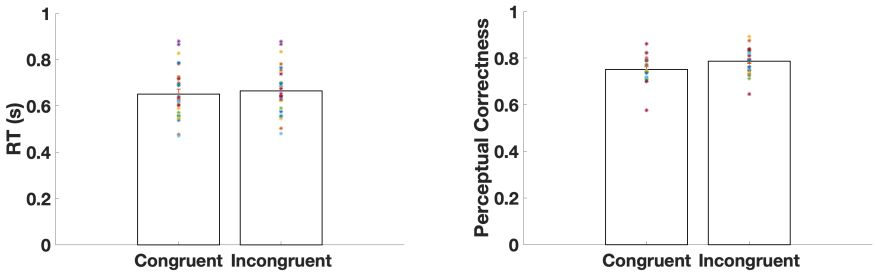


Figure 37: Distribution of response time in stay, rational switch and irrational switch trials. X-axis: normalized RT (mean RT subtracted; in seconds). Y-axis: proportion of trials out of total.

Simon effect

As the button layout did not match hand response with a specific patch (i.e. left hand responding to both patches rather than the left patch only), we hypothesized that the button congruency might influence the reaction time and accuracy. The results revealed that when the pressed button was incongruent (i.e. left hand responding to the right patch or vice versa), the response was slower ($t = -3.2011$, $p = 0.00359$) but more accurate ($t = -4.0936$, $p = 0.000366$) (**Figure 38**).



(a) RT (b) Accuracy

Figure 38: (a) Reaction time split by button congruency. (b) Perceptual correctness split by button conguency. Congruent button were defined as left hand responding to left patch or right hand responding to right patch, and the incongruent button were the contrary. Each asterisk was the data point from one participant, and the black square was the average across participants.

Stimuli congruence effect

Although participants were instructed to attend to one patch, their performance could be influenced by conflicting or congruent information from the two patches. We investigated the stimuli congruence effect with a GLM and it turned to be present in our task (**Figure 39**). Responses were more accurate given the same rotation direction and the same base angle direction, but the two effects did not add up. Therefore, congruent visual information boosted accuracy but more than one source of congruence may undermine the performance.

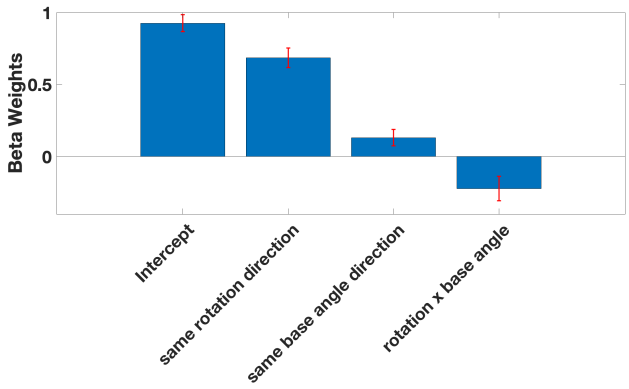


Figure 39: Effect of congruent direction of rotation (both clockwise or anti-clockwise), same base angle direction (both positive or negative at S1) and the interaction of the two on perceptual accuracy.

3.3.2 Summary of Behavioural Findings

Participants' performance in the task was generally consistent with our hypotheses: they were able to detect the rotation and achieved an average accuracy of 80% in the task (given that 20% trials did not rotate at all, the actual perceptual accuracy would be around 90%); the accuracy increased as difficulty decreased; they were able to infer the reward relevance and switched to the correct patch based on feedback and the evaluation on their own performance, combined with the previous experience. Based on the ground truth, we split switch trials into rational and irrational switch. The performance in stay and rational switch trials was excellent and improved as difficulty decreased, whereas in irrational switch trials the responses were inaccurate and slow regardless of difficulty. Further analyses showed that before a rational switch, both accuracy and reaction time were relatively better, but before an irrational switch responses were much more erroneous and slower. After either switch, responses slowed down a bit. These suggest that irrational switch was largely accounted by poor concentration on the task, whereas for a rational switch to be made,

participants usually paid more attention.

3.4 Drift Diffusion Model (DDM)

As decisions in our task involved sensory evidence accumulation, we decided to further characterise the behavioural effects using the Drift Diffusion Model. The purpose of fitting a DDM to data was to break down reaction time and accuracy further to know if the differences were attributed to decision components such as different decision boundaries, drift rates or non-decision time. We particularly wanted to understand the Simon effect (incongruent responses were slower but more accurate). The increase in accuracy suggests that the slowing on incompatible trials was not, as expected, due to an increase in non-decision time (which would not affect accuracy), but that the decision bound or drift rate must be affected.

3.4.1 DDM Methods

We built a drift diffusion model (DDM) to analyse the reaction time and accuracy data using the DMAT toolbox (Vandekerckhove and Tuerlinckx, 2008). The DMAT toolbox fits a conventional Ratcliff diffusion model to RT and accuracy, with seven free parameters. The decision boundary, denoted as a, determines the threshold when accumulated evidence reaches and a decision is made. The starting point of the evidence accumulation process z_0 is sampled from a uniform distribution with mean z and range s_z in each trial. The drift rate specifies the tendency of evidence to drift off towards one of the boundaries, and is sampled from a normal distribution with mean v and standard deviation η. Non-decision time τ sampled from a uniform distribution with mean T and range s_t, estimates the time needed to perform non-decision processes such as the execution of motor response (Vandekerckhove and Tuerlinckx, 2007).

In DMAT, parameters are represented by matrices and users can customise the design matrices based on assumptions of how parameters change in different

conditions. A typical parameter matrix can be represented by $\mathbf{P} = \{\mathbf{D}_a \times \mathbf{a}, \mathbf{D}_{T_er} \times \mathbf{T}_{er}, \mathbf{D}_\eta \times \eta, \mathbf{D}_z \times \mathbf{z}, \mathbf{D}_{sz} \times \mathbf{s}_z, \mathbf{D}_{st} \times \mathbf{S}_t, \mathbf{D}_v \times \mathbf{v}\}$, where $\mathbf{D}$ represents the design matrix for a specific parameter. If a design matrix contains only ones, then the parameter is fixed across all conditions. If in C conditions, the design matrix is a C × C identity matrix, then the parameter varies freely across those conditions. The rest $\mathbf{a}$, $\mathbf{T}_{er}$, η, $\mathbf{z}$, $\mathbf{s}_z$, $\mathbf{s}_t$ and $\mathbf{v}$ are the parameters over which the algorithm will optimise the fit to data (Vandekerckhove and Tuerlinckx, 2007).

Free parameters were fitted using a separate design matrix for each free parameter. For example, a set of drift rates could be fitted independently for each of several conditions, or fitted as a linear function of a task variable such as perceptual difficulty. The free parameters that can be fitted in the DMAT toolbox are described in **Table 7** below. We first used DDM to investigate the effect of button congruency on the evidence accumulation process. We fitted a DDM separately in congruent and incongruent trials and took difficulty as the manipulated condition within each model with all parameters fixed, except the drift rate, for which we assumed a negative linear relationship with difficulty (a vector [1, 0.8, 0.6, 0.4, 0.2] was used as the design matrix for drift rate). We then fitted DDM separately in stay, rational switch and irrational switch trials to test if the model could capture patterns of RT and accuracy found in different decision processes. As splitting switch trials significantly reduced the sample size, we selected only three intermediate difficulty levels to reduce uncertainty (as the easiest and the most difficult trials introduced higher bias in parameter estimation, causing failure in convergence). We again assumed a negative linear relationship between drift rate and difficulty ([0.8, 0.6, 0.4]) and fixed the rest of the parameters within each model. The DMAT toolbox then optimised each parameter using a negative multinomial log-likelihood function (MLF) with 5000 simulations.

Parameter Name	Denotation	Design Matrix
decision boundary	**a**	ones
mean of non-decision time	**T**	ones
standard deviation of drift rate	η	ones
mean of starting point	**z**	ones
range of starting point	$\mathbf{s}_z$	ones
range of non-decision time	$\mathbf{s}_t$	ones
mean of drift rate	**v**	[1, 0.8, 0.6, 0.4, 0.2] for five difficulty levels

Table 7: Parameters of DDM fitted in the DMAT toolbox.

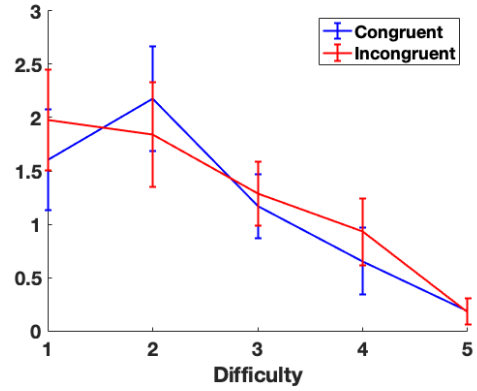
3.4.2 DDM Results

Simon effect

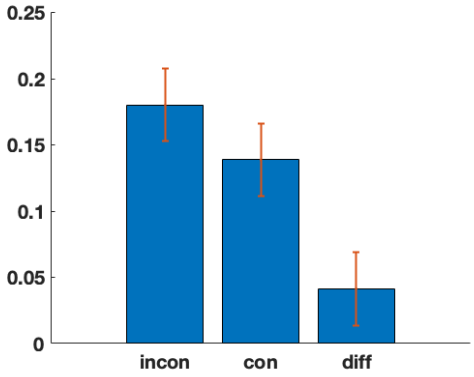
The fact that response-incongruent trials (e.g., respond with the left hand to the right patch) were slower and more accurate suggests an intriguing situation in which slowing the motor response (due to the Simon effect) allows longer evidence accumulation and hence, higher accuracy. We attempted to capture this process with the DDM.

We fitted a DDM separately to congruent and incongruent trials. Within each trial category, the drift rate was modelled to depend on perceptual difficulty **Table 7**, whereas the decision boundary and the non-decision time were constants (for each trial type). Consistent with our hypothesis, the decision boundary was higher on incongruent trials, however the difference was not statistically significant (z = 1.0571, p = 0.2905) **Figure 40**. There was no difference in non-decision time between congruent and incongruent trials (z = 1.0331, p = 0.3016). There was no difference in drift rate between congruent and incongruent trials (Wilcoxon signed rank test; z = 1.2585, p = 0.2082), but drift rate did decrease as difficulty increased (linear regression: congruent: t = -5.0345, p = 1.5267e-06; incongruent: t = -5.7472, p = 5.9061e-08). These results indicate the slower but more accurate responses in incongruent trials were probably partly

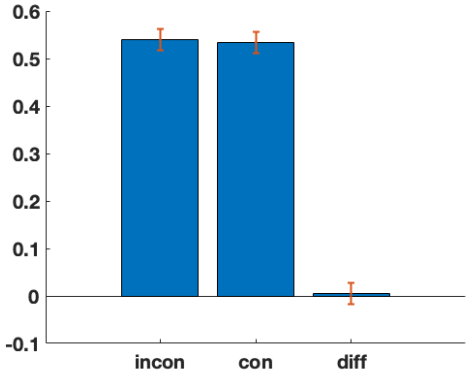
attributed to increased decision boundary, but the effect was not statistically significant, probably due to insufficient power.



(a) Drift rate



(b) Decision boundary



(c) Non-decision time

Figure 40: Estimated drift rate (a), decision boundary (b) and the non-decision time (c) from the DDM for the Simon effect. Parameters were estimated independently in congruent and incongruent trials. For decision boundary and the non-decision time, difference between congruent and incongruent trials was calculated and plotted here.

Switching effects

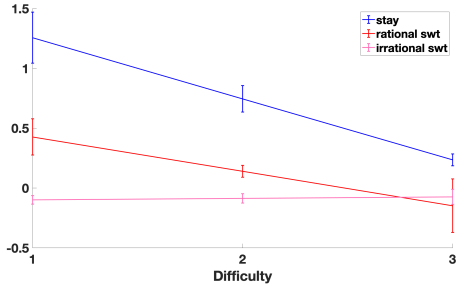
As analysis of RT and accuracy revealed differences among rational switch, irrational switch and stay, we fitted DDM separately in the three conditions to further dissociate the speed-accuracy trade-off underlying different decision processes. Analysis of the raw RTs and accuracy showed that RTs were faster on rational switch trials than other trial types, and accuracy was lower on irrational switch trials than other trial types **Figure 33**.

We limited the analysis to three intermediate difficult levels (rotation 9°, 6°and 3°), as there were ceiling effects on accuracy in the easiest case and there was no correct evidence to accumulate in the most difficult case. We assumed a linear relationship between difficulty and drift rate in all three conditions, but a constant decision boundary and non-decision time for each condition.

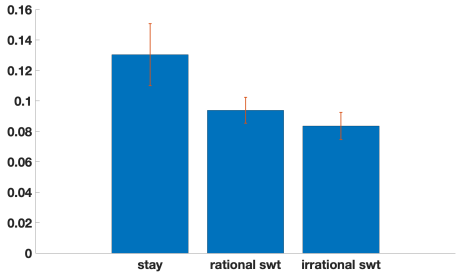
As **Figure 41** shows, the drift rate on stay trials was higher than rational switch ($z = 6.4056$, $p = 1.4976e\text{-}10$), but both decreased linearly as increasing difficulty and remained above zero. Drift rate of irrational switch was almost zero and stayed flat across difficulty in the irrational switch trials. Decision boundary was fixed across difficulty and was found higher in stay than rational switch, but the difference was not significant ($z = 1.5856$, $p = 0.1128$). Non-decision time was significantly higher in stay than rational switch ($z = 3.1232$, $p = 0.0018$). The increased drift rate in stay trials would tend to predict shorter RTs, but the increased non-decision time would predict longer RTs. The slightly longer RT found in stay than rational switch trials (**Figure 33**) was probably attributed to non-decision time. Furthermore, since non-decision time does not affect accuracy, this also explains no difference in accuracy between stay and rational switch trials.

Twelve out of the total twenty-seven subjects had so few irrational switch trials that the model failed to converge, resulting in a sample size of fifteen for

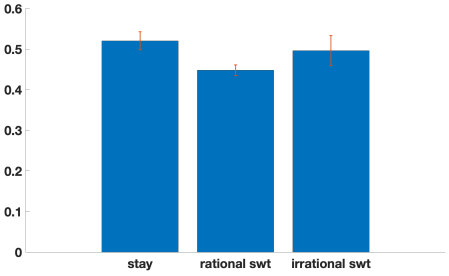
the irrational switch case. As the irrational switch trials were undersampled, statistical analysis of its parameters was not carried out, although its low drift rate and high non-decision time might well explain the 'slow errors' found in RT and accuracy.



(a) Drift rate



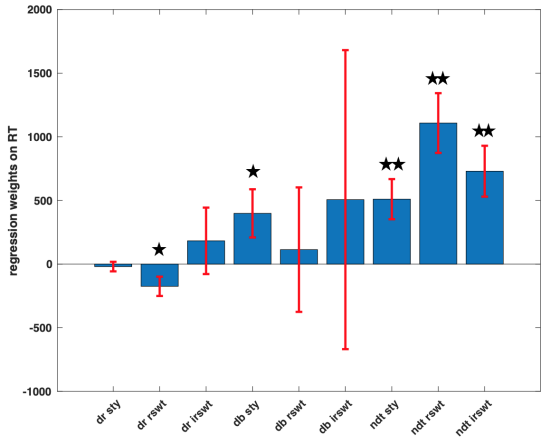
(b) Decision boundary



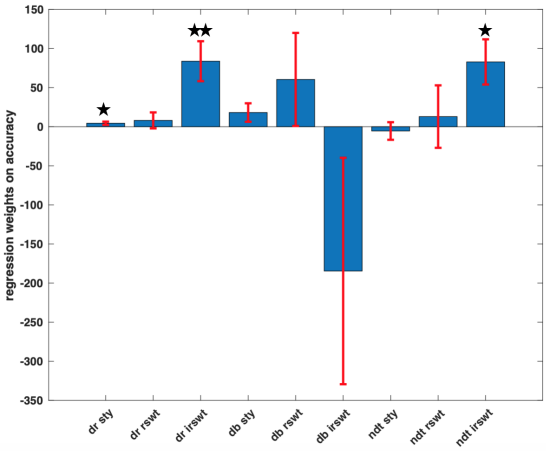
(c) Non-decision time

Figure 41: Estimated drift rate (a), decision boundary (b) and the non-decision time (c) from the DDM for trials with intermediate difficulty levels (rotation 9°, 6°and 3°). Parameters were estimated independently in rational switch, irrational switch and stay trials. Drift rates were modelled as linearly increasing across difficulty.

We then tested which model parameters were most strongly associated with RT and accuracy observed in behaviour by running GLM between each parameter and behavioural trait. As **Figure 42** reveals, RT in stay was most reflected by decision boundary ($p = 0.0455$) and non-decision time ($p = 0.0034$), whereas its accuracy was associated with drift rate ($p = 0.0424$). In rational switch, RT was reflected by drift rate ($p = 0.0299$) and non-decision time ($p = 7.64e\text{-}05$), whereas its accuracy was not related to any parameter. In irrational switch, RT was strongly associated with non-decision time ($p = 0.00295$), and accuracy was reflected by both drift rate ($p = 0.00594$) and non-decision time ($p = 0.0132$).



(a) RT



(b) Accuracy

Figure 42: Regression weights of drift rate, decision boundary and non-decision time from DDM on (a) RT and (b) accuracy in stay, rational switch and irrational switch trials. Parameters significantly different from zero are marked by asterisks (one asterisk, $p < 0.05$; two asterisks, $p < 0.01$).

3.4.3 DDM Discussion

As responses in incongruent trials were slower but more accurate than in congruent trials, we hypothesised that this may be associated with higher decision boundary. The DDM output confirmed this hypothesis, although the difference

was not significant. Incongruent trials also showed slightly higher drift rate and non-decision time. These results suggest that button incongruency might promote the cognitive effort devoted into the task. People seemed to be more cautious and concentrate to overcome the button incongruency, and that incongruency did not halt the decision by taking higher motor execution time. Overall, however, DDM could not provide a plausible explanation for Simon effect observed in our task.

The model gave nearly zero drift rates for irrational switch trials, whereas the low drift rates were significantly related to accuracy. Possibly as the accuracy in irrational switch was generally low (approx. 12%), tiny changes in drift rates may have big impact on accuracy. The non-decision time made the responses slower and more accurate, whereas the classical DDM regards the non-decision time as part of the observed RT and it should not be associated with accuracy. Moreover, the positive effect of drift rate on RT and negative effect of boundary on the accuracy contradict our hypotheses. Given the low sample size of irrational trials (the model converged on 15 out of 27 subjects), the model fitting is unreliable in this case and should be interpreted with caution. Drift rates of both stay and rational switch made the decision faster and more accurate, whereas the boundary led to slower but more accurate decisions, and the non-decision time made decisions slower but was irrelevant to accuracy. These findings suggest the slightly longer RT in stay than rational switch trials were probably explained increased non-decision time. Note that in rational switch trials, a significant relationship was found only between parameters and RT, whereas between parameters and accuracy the effect was qualitatively sensible but not statistically significant. It is possible that rational switch trials were generally highly accurate (approx. 80% were perceptually correct), the effect was at ceiling and the model fitting was largely based on RT. Another noteworthy

finding is that although drift rates of stay were significantly higher than rational switch, the former were not significantly more accurate or faster than the latter, contradicting with what the model would predict in behaviour. Taken together, the model gave a plausible description within each decision type, but parameter comparison between different decisions was not reliable. Therefore, we decided not to carry out further processing of DDM outcomes, such as relating EEG signals to the parameters.

3.5 Bayesian Model (BM)

Participants always had to infer the target patch based on a history of feedback. Therefore their belief strength about which was the target patch should depend on the precise set of trials seen (rewarded/ unrewarded X difficulty). We built a Bayesian model to capture what an ideal observer would infer about the target patch, and their belief strength, on a trial by trial basis. The purpose of this was that we expected the response to feedback (as captured in the EEG) to depend on belief strength, because beliefs (of stay) are more changeable by feedback when beliefs about the target patch are weak (as weak beliefs indicate trials might have been unrewarded for a while and the suspicion of switch raised; therefore a single feedback was taken into account more carefully when participants were pondering over 'leaving').

3.5.1 BM Methods

For each patch, we assume the observer would estimate the posterior probability of 'this is the correct patch given the received feedback', based on the prior belief ('probability of this is the correct patch') and the likelihood ('probability of receiving the current feedback given the correct patch'). Note that the likelihood would entirely depend on perceptual accuracy for the chosen patch, and for the unchosen patch the likelihood would just be 0.5. According to the Bayes' rule, we constructed the following equation for the chosen patch:

$$post_c = p(perceptually\ correct) \times prior \tag{19}$$

For the unchosen patch, we used the following equation:

$$post_{u_c} = 0.5 \times prior \tag{20}$$

The posterior probability would serve as the prior for the next trial, discounted by a leakage parameter k:

$$prior = (1 - k) \times post + k \tag{21}$$

Where k represents a flat prior equivalent to the average chance of an actual reversal (= 0.1 as the actual reversal was drawn randomly from 5 to 15 trials and should happen on average every 10 trials).

3.5.2 BM Results

In general, model prediction was aligned with the ground truth changes as well as participants' behaviour (**Figure 43**). We then evaluated the model performance with logistic regressions (**Figure 44**), and found no general bias in ground truth (t test on the intercept: t = -0.0852, p = 0.9328) or participants' choices (t test on the intercept: t = -0.6350, p = 0.5309), but the choice confidence was associated with the ground truth (t = 21.1908, p = 6.2901e-18) and both the prior belief (t = 21.3839, p = 5.0301e-18) and the ground truth (t = 6.2629, p = 1.2540e-06) were able to predict participants' choices. The results revealed that the model was able to track the ground truth and update the choice confidence accordingly and to a large extent predict participants' choices.

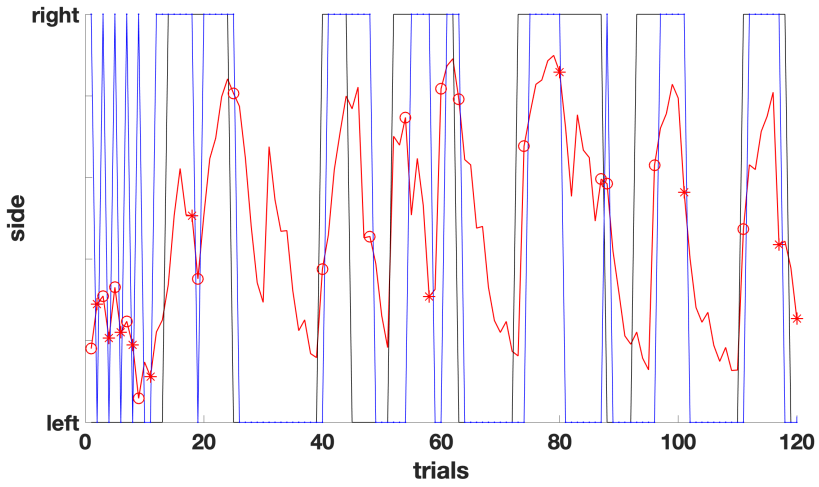
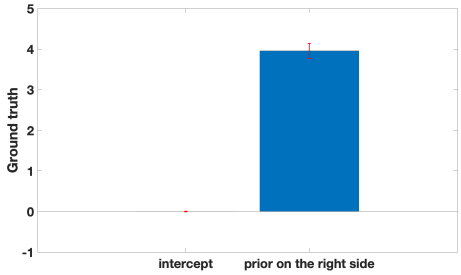
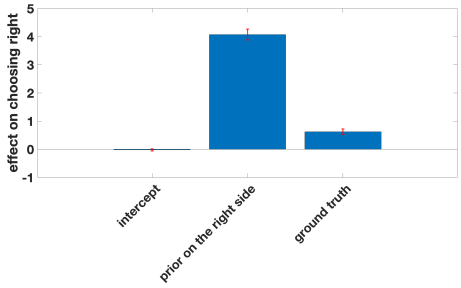


Figure 43: Schematic illustration of model prediction on one subject's data: choice confidence (posterior probability on the chosen side; red) plotted together with ground truth (black) and participants' chosen side (blue). red dot: rational switch; red asterisks: irrational switch.



(a) Ground truth



(b) Performance

Figure 44: GLM on model performance across all subjects: (a) prior belief predicting the correct side (ground truth) and (b) effect of prior belief on choosing the side (participants' behaviour).

3.5.3 BM Discussion

The problem facing participants during the task was to choose the patch relevant to reward and report its rotation. Because they did not know the reward relevance, they must estimate from the feedback, cross-trial experience as well as the assessment on their own performance (or choice confidence) based on the objective difficulty. We built a Bayesian inference model that inferred the reward relevance by taking into account choice confidence in each trial. We found both the ground truth (the actual reward relevance) and participants' choices were reflected by the Bayesian prediction. The Bayesian model captures the causal inference that we hypothesized participants used in our task very well. Since the Bayesian model provides an estimation on reward expectations,

we next investigated how feedback related neural signals were modulated by reward expectations using the prediction from Bayesian model (see the next chapter).

3.6 EEG

3.6.1 EEG Materials and Methods

Recording and Processing

Electroencephalographic signals were recorded with active electrodes from 61 scalp electrodes mounted equidistantly on the standard 10-20 system elastic map (EasyCap). All electrodes were referenced to the right mastoid electrode and re-referenced offline to the average of all electrodes. Continuous EEG was recorded using a SynAmps RT 64-channel Amplifier (1000 Hz sampling rate). Offline EEG analysis was performed using a MATLAB toolbox developed by the Donders Institute in Nijmegen (Fieldtrip). The data were epoched from -1 to 2s relative to the target onset except for S2 which was epoched from -3 to 2 s. The data were then band-pass-filtered at [1 35] Hz, baseline corrected to 0.2 s window before each event. Artefacts were detected and rejected using a correlation-based independent component analysis (ICA) approach. Before running ICA, trials containing artefacts were removed by visual inspection to avoid contaminating component training. Artefact components were identified if their correlation coefficients with the EOG electrodes exceeded 0.7. Muscle artefacts were detected using z-transforming on eight electrodes ('PO7', 'PO8', 'POZ', 'PO4', 'PO3', 'O1', 'OZ', 'O2') near the neck which were most affected by muscle movements with a z-score cutoff at 20. After muscle filtering a second visual inspection was applied to the data.

3.6.2 EEG Results

We analyzed EEG signals time locked to four timepoints: the onset of the first stimulus, the onset of the second stimulus, the response and the feedback. Specifically, we performed analyses on evoked potentials for all four events, and time-frequency components for stimulus presentation (S1 and S2) and response.

For most analyses, we used a general linear model (GLM) approach, either with parametric regressors (such as a regressor weighting trials by the five levels of difficulty) or using dummy regressors to capture different trial types. In particular, the following questions guided EEG analyses:

- Can we establish EEG correlates (e.g. visual-evoked potentials (VEP) and alpha oscillations) of attentional effects involved in visual information processing? Were visual selective attention conveyed by VEPs and anticipatory attention conveyed by alpha oscillations modulated by task related factors such as perceptual accuracy, difficulty and switch?

- How do the response related signals (e.g. lateralised readiness potentials (LRPs) and beta power) reflect behavioural changes, in terms of choice commitment and response speed, as well as the above mentioned factors?

- How do feedback related components (e.g. FRN and P300) reflect feedback processing in our task? Is the switch evidence that was found to build up across trials in behavioural data also reflected by those brain components? How is choice confidence (e.g. expected perceptual accuracy and expected reward contingency) derived from Bayesian model reflected by brain signals?

3.6.2.1 Time-locked to onset of stimuli (S1 and S2)

Lateralized visual-related potentials

In our task, participants chose to attend to one or other side of the screen. We therefore sought to establish EEG correlates of this differential attention effect. Lateralised visual-evoked potentials have been found to reflect selective visual attention (Eimer et al., 2006; Hilo-Merkovich et al., 2018). We therefore wondered if factors such as attended side, accuracy and switch would be reflected

by VEP fluctuations. To investigate the hypothesis that the visual related potentials (VEP) would be stronger on the contralateral hemisphere of the visual cortex to the attended side and the signals would be stronger as the cognitive effort increased, we time-locked to the denoised data from 0 to 500 ms around S2 and built a linear regression model as specified below to examine the effect of cognitive effort and the attended side (left vs. right patch) on the VEP:

$$\begin{aligned} VEP = \beta_1 \times attend_L + \beta_2 \times attend_R + \beta_3 \times WAM_L \\ +\beta_4 \times WAM_R + \beta_5 \times Zero_L + \beta_6 \times Zero_R \end{aligned} \tag{22}$$

Here $attend_L$ reflected the left patch was chosen (1=left, 0=right) and $attend_R$ reflected the right patch (0=left, 1=right) was chosen. The term WAM referred to the weighted accuracy measure, which varied from -4 to 3 as the difficulty level increased, whereas negative and positive values reflected perceptually incorrect and correct, respectively(**Figure 45**). In this way, WAM took into account both accuracy and difficulty and served as a good proxy for measuring the cognitive effort. As WAM became zero when the rotation degree was zero, we added two additional regressors $Zero_L$ and $Zero_R$ for those zero degree trials as co-regressors.

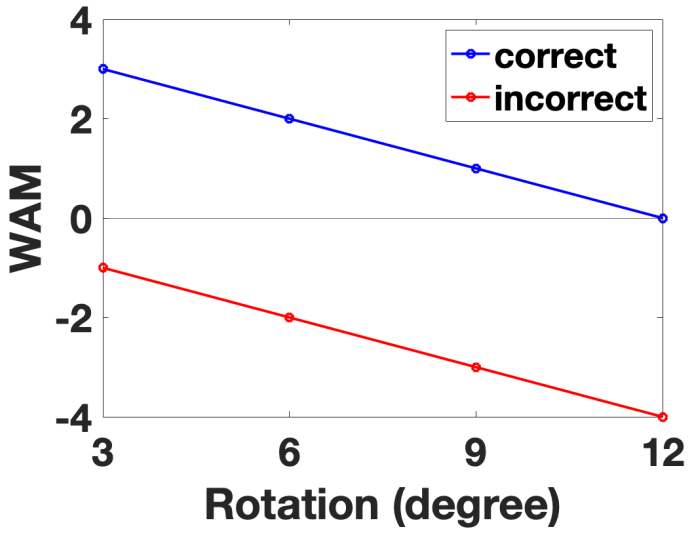
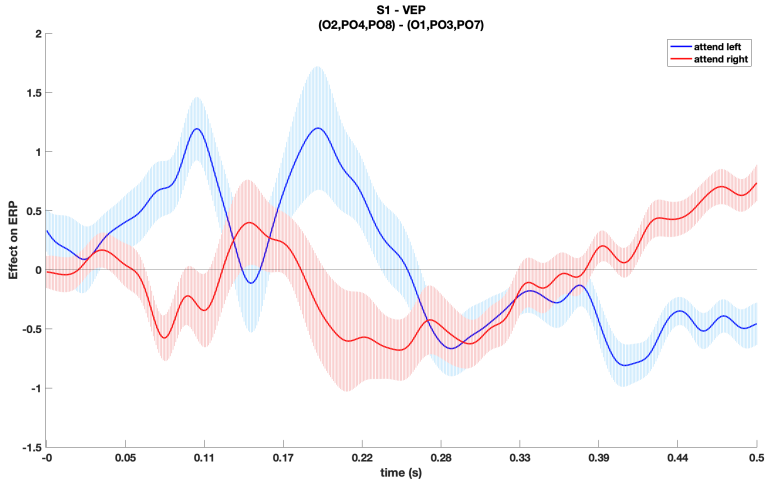
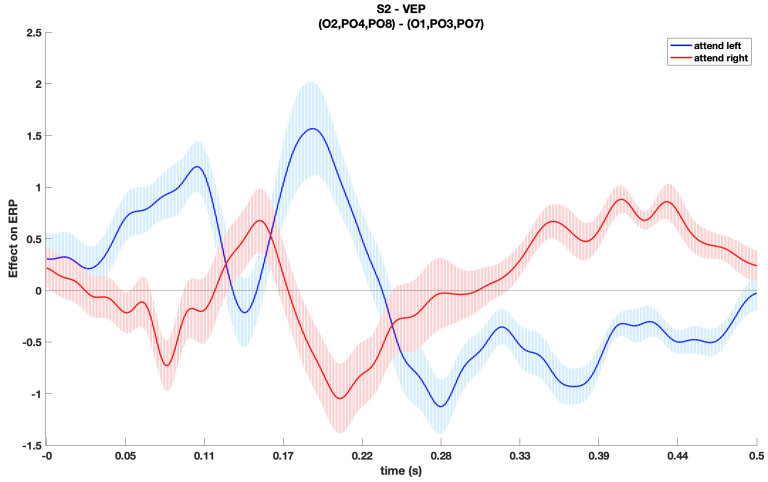


Figure 45: Weighted accuracy measure (WAM) plotted against five levels of difficulty, with perceptually correct in blue and perceptually incorrect in red.

We first asked whether selective attention to one or other side is reflected in the visual evoked potential. **Figure 46** shows lateralized visual evoked potentials (VEP) when participants attended to the left and right visual side after S1 and S2 onset. The results have shown opposite deflections in lateralized VEPs attending to contralateral and ipsilateral side, with the contralateral side displaying strong positive ERPs. The lateralized VEPs of the contra-ipsi contrast peaked between 170 and 220 ms for at S1 and S2 (S1: t = 4.8871, p = 4.5271e-05; S2: t = 5.9297, p = 2.9519e-06), as shown by **Figure 47**. The results show the existence of lateralised VEP as the signals differed from zero when attending to left or right side as well as significant opposite deflection between different attended sides. We then performed subsequent analyses on how lateralised VEP reflected other factors such as accuracy and WAM.



(a) S1



(b) S2

Figure 46: Lateralized Visual evoked potentials (averaged three electrodes from the right hemisphere (O2+PO4+PO8) minus averaged three electrodes from the left hemisphere (O1+PO3+PO7)) split by attending to the left (blue) and attending to the right (red) side, time-locked to S1 (top) and S2 (bottom). Signals were plotted from 0 to 500 ms post stimulus onset. Y-axis: effect of attended side on ERP (plotted here were the beta weights of 'attendL' and 'attendR' from GLM, in arbitrary units proportional to the actual potential). X-axis: time in seconds.

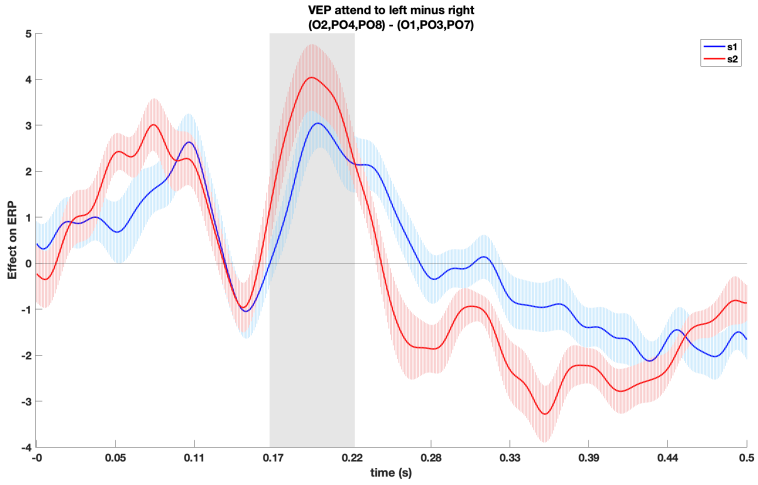
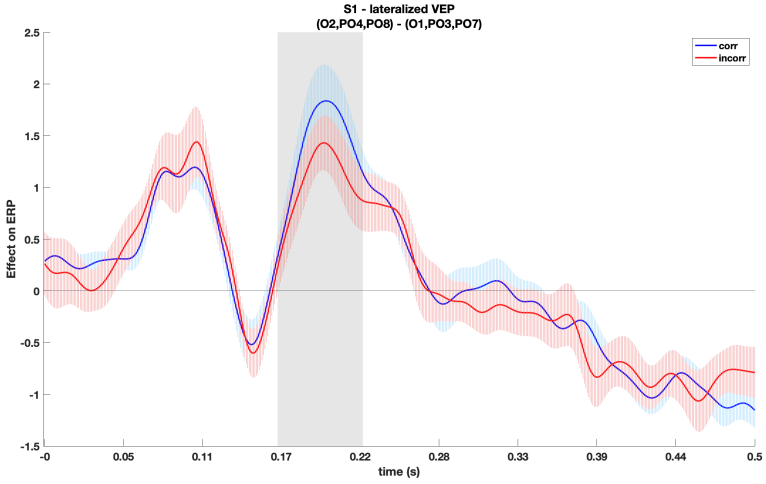


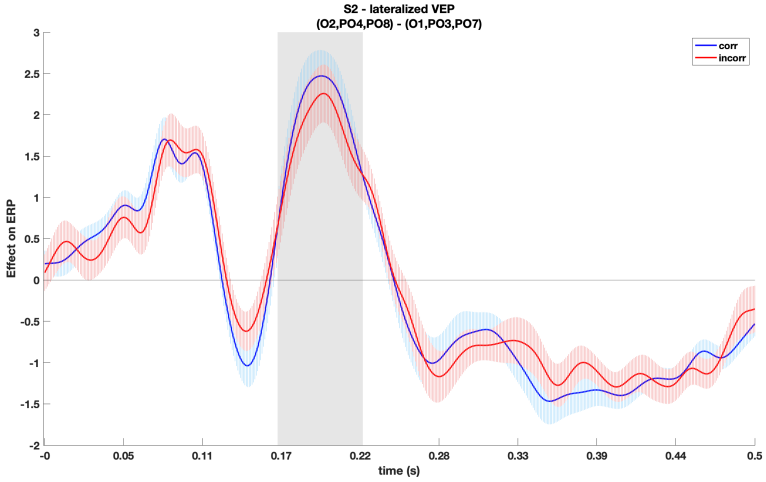
Figure 47: Lateralized VEPs (signals from the right hemisphere minus the left hemisphere when attending to the contralateral side), time-locked to S1 (blue) and S2 (red). Signals were plotted from 0 to 500 ms post stimulus onset. The grey shaded area illustrates the peak of lateralized VEP found between 170 ms and 220 ms. Y-axis: effect of attended side on ERP (plotted here were the difference in beta weights ('attendL' minus 'attendR') from GLM, in arbitrary units proportional to the actual potential). X-axis: time in seconds.

Lateralized VEP by perceptual accuracy

We hypothesised that if the lateralized VEP effect reflected selective attention, it should be stronger on trials where the response was perceptually correct. As shown by **Figure 48**, VEP was stronger in trials that were perceptually correct both at S1 and S2 (S1: $t = 2.9745$, $p = 0.0063$; S2: $t = 2.3165$, $p = 0.0287$).



(a) S1



(b) S2

Figure 48: Lateralized VEPs split by perceptually correct (blue) and incorrect (red) trials, time-locked to S1 (top) and S2 (bottom). Signals were plotted from 0 to 500 ms post stimulus onset. Y-axis: effect of accuracy on ERP (plotted here were the beta weights of 'correct' and 'incorrect' from GLM, in arbitrary units proportional to the actual potential). X-axis: time in seconds.

Lateralized VEP by weighted perceptual accuracy measures (WAM)

Since difficulty varied across trials, it was harder to be perceptually correct on some trials than others. When participants were perceptually correct on very difficult trials, this should reflect a higher level of attention (on average) than when they were correct on easy trials; and conversely when they were perceptually incorrect on easy trials, this should reflect a lower level of attention than when they were incorrect on difficult trials. As mentioned before, to capture this we defined a weighted accuracy measure (WAM) as follows: [-4, -3, -2, -1] are from easy to difficult for perceptually incorrect trials, and [0, 1, 2, 3] from easy to difficult for perceptually correct trials. The effect of WAM on the lateralized VEP are shown by **Figure 49**. Note that the zero degree trials were separately controlled as co-regressors, whereas only the other four difficulty levels were actually testified by this analysis. Although the effect was slightly noisy, the VEP lateralization indeed depended on WAM especially from 170 to 220 ms post stimulus presentation for both S1 and S2 ($t = 2.5496$, $p = 0.017$ for S1; $t = 2.8831$, $p = 0.0078$ for S2).

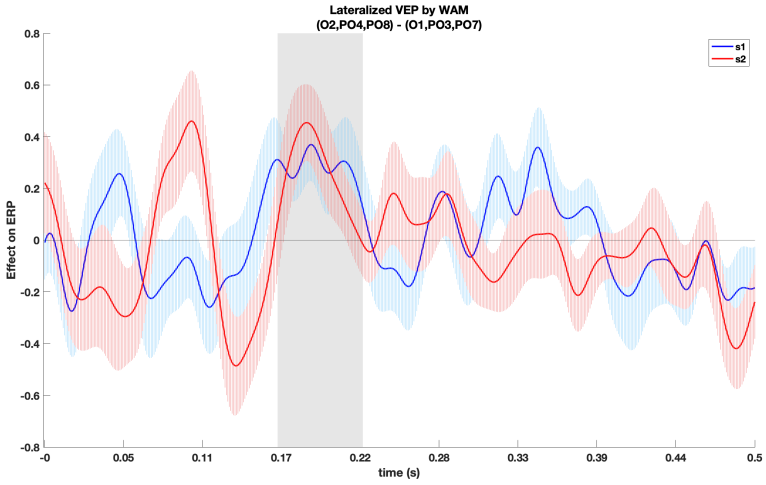


Figure 49: Effect of WAM on lateralized VEP time-locked to S1 and S2 from 0 to 500 ms post stimulus presentation. Y-axis: effect of WAM on ERP (plotted here were the beta weights of 'WAM' from GLM, in arbitrary units proportional to the actual potential). X-axis: time in seconds.

Attention modulated power changes

Next we wondered if anticipatory attention was present during the run-up to S2 period when participants saw nothing on the screen, tried to hold S1 in working memory and anticipated the upcoming S2 and whether attention was increased on trials with high concentration (high WAM). Previous studies established that the alpha desynchronisation reflected enhanced anticipatory attention (Worden et al., 2000; Thut et al., 2012; Kelly et al., 2006; Händel et al., 2011; Bonnefond and Jensen, 2012), so we expected that alpha desynchronisation prior to s2 would be higher on high-WAM trials. Anticipatory alpha modulation analysis was performed from -1 to 2 s around S1 and from -3 to 1 s around S2. We used a multi-taper time frequency transformation on the noise-free signals around events, downsampled to 250 Hz, with a hanning taper and limited the wavelets of interest from 2 to 40 Hz. We used a window length of 300 ms sliding in steps of 40 ms. To confirm the effect of attention on the alpha power, we performed

equation 22 and 23 on the collapsed power across the frequency range of 8-14 Hz.

Lateralized alpha power in the ISI: attention effects

Lateralized anticipatory alpha power is known to relate to selective attention (Worden et al., 2000; Thut et al., 2012; Kelly et al., 2006; Händel et al., 2011; Bonnefond and Jensen, 2012). We therefore investigated how lateralized alpha power during the ISI (in the run-up to S2) varied with the above measures (we performed a regression as specified by **Equation 22** on the alpha band). **Figure 50** shows the effect of WAM on the lateralized alpha power (attended relative to the unattended side on the contralateral hemisphere) and the lateralized alpha power. The results show alpha desynchronization on the attended side starting from -1.3 s before the S2 onset (t = -2.6246, p = 0.0143). Alpha power was not modulated by WAM (t = -0.0751, p = 0.9407).

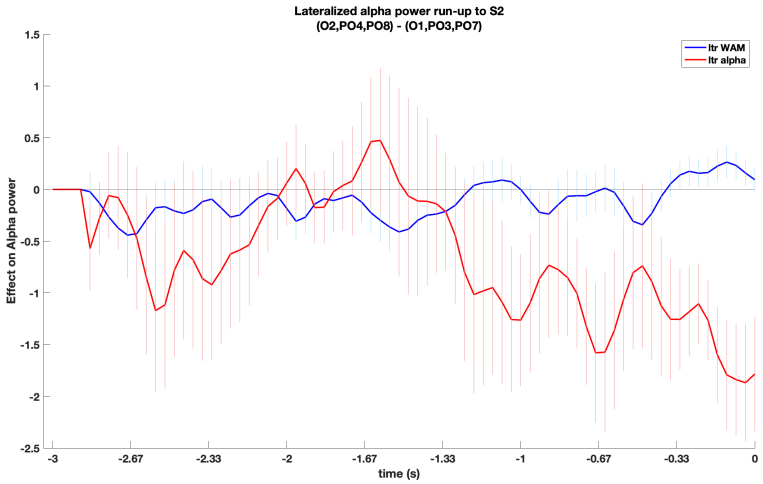
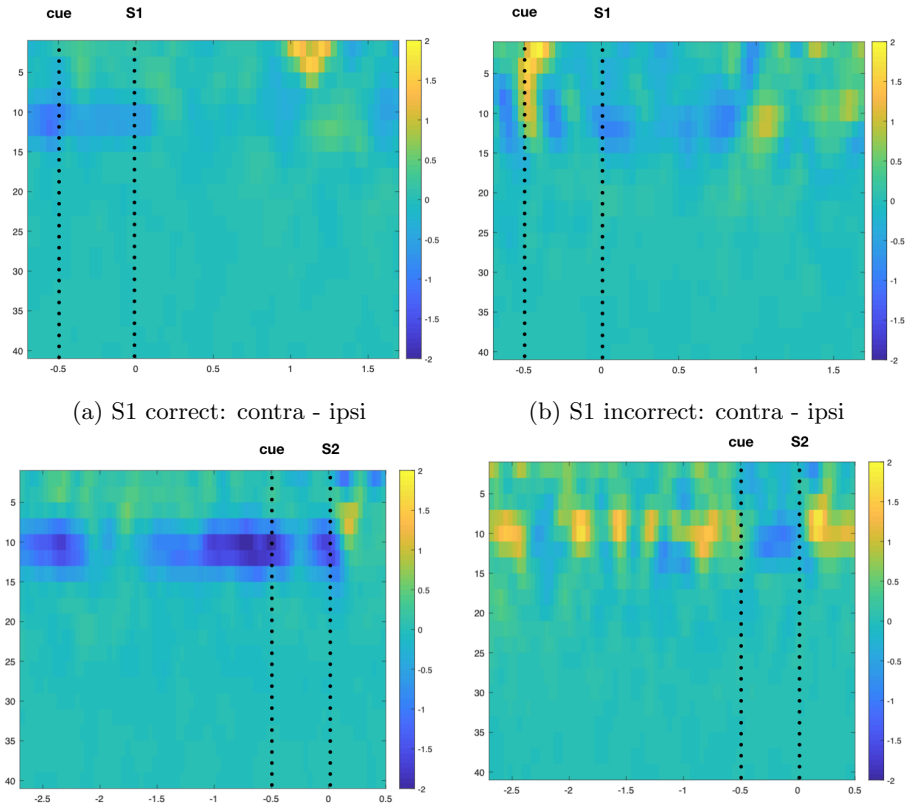


Figure 50: Lateralized alpha power (red) and lateralized effect of WAM (blue) on the alpha power run-up to S2 (-3 to 0 s). Y-axis: beta weights of 'WAM_L - WAM_R' and 'attendL - attendR' from GLM, in arbitrary units proportional to the actual potential/power. X-axis: time in seconds.

Next we wondered if enhanced anticipatory attention conveyed by alpha desynchronization found during ISI reflects perceptual accuracy. **Figure 51** shows power difference between contralateral and ipsilateral electrodes from the occipital cortex split into correct and incorrect trials time-locked to S1 and S2, respectively. It seems correct trials displayed stronger alpha desynchronization in the run-up to S2, compared to incorrect trials, but the difference was not significant across participants (t-test on averaged lateralised alpha band (8-14 Hz, 1 to 1.5 s post S1, correct vs. incorrect: t = -1.3603, p = 0.1854; -1.3 to -0.5 s prior to S2: t = -1.6142, p = 0.1185).



(a) S1 correct: contra - ipsi

(b) S1 incorrect: contra - ipsi

(c) S2 correct: contra - ipsi

(d) S2 incorrect: contra - ipsi

Figure 51: Difference in power between contralateral and ipsilateral electrodes around the occipital cortex relative to the attended side, time-locked to S1 perceptually correct (a), S1 perceptually incorrect (b), S2 perceptually correct (c) and S2 perceptually incorrect (d) trials. Two pairs of electrodes (PO3+PO7 vs. PO4+PO8) from the left and right hemisphere were selected. X-axis: time relative to stimulus onset (s). Y-axis: frequency (Hz).

Are attention effects stronger on (some) switch trials?

We next hypothesised that the decision type (switch/stay) could be reflected by attention to the visual stimuli. Therefore we built another regression model to

examine the effect of decision type as following:

$$\begin{aligned} VEP = {} & \beta_1 \times attend_L stay + \beta_2 \times attend_R stay + \beta_3 \times WAM_L stay \\ & + \beta_4 \times WAM_R stay + \beta_5 \times Zero_L stay + \beta_6 \times Zero_R stay \\ & + \beta_7 \times attend_L rswt + \beta_8 \times attend_R rswt + \beta_9 \times WAM_L rswt \\ & + \beta_{10} \times WAM_R rswt + \beta_{11} \times Zero_L rswt + \beta_{12} \times Zero_R rswt \\ & + \beta_{13} \times attend_L irrswt + \beta_{14} \times attend_R irrswt + \beta_{15} \times WAM_L irrswt \\ & + \beta_{16} \times WAM_R irrswt + \beta_{17} \times Zero_L irrswt + \beta_{18} \times Zero_R irrswt \end{aligned} \tag{23}$$

Here we split the switch trials into rational (rswt) and irrational (irrswt) switch. The regressors '$Zero_L$' and '$Zero_R$' controlled the trials with zero rotation, and 'WAM_L' and 'WAM_R' included WAM in trials with the other four difficulty levels. In each regressor, trials that did not belong to the specified condition were coded as zero (for example, '$attend_R irrswt$' captured only irrational switch and attending to the right side trials, whereas all other trials, such as stay, rational switch and attending to left trials were coded as zero). This regression broke down attended side, WAM and difficulty into stay and switch trials, thus allowing us to investigate the effects separately.

Attention effects were absent in irrational switch trials

In our behavioural analysis, we noted a category of trials (irrational switch trials) on which participants appeared to be in a low attention state. On these trials reaction times were slow and did not depend on difficulty. We therefore tested whether the attention effects on stimulus processing were reduced in irrational switch trials. Results revealed that the VEP lateralization (**Figure 52**) was not present in irrational switch trials (averaged lateralized VEP 170-220 ms: S1: t = 0.6586, p = 0.5160; S2: t = 0.9501, p = 0.3508).

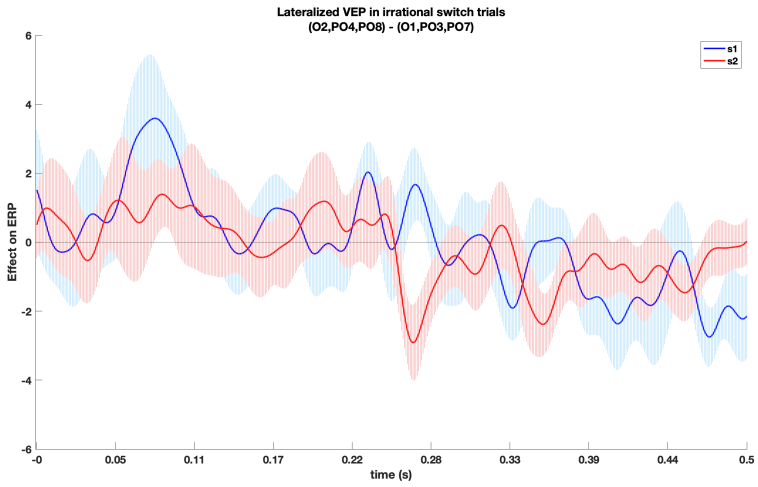
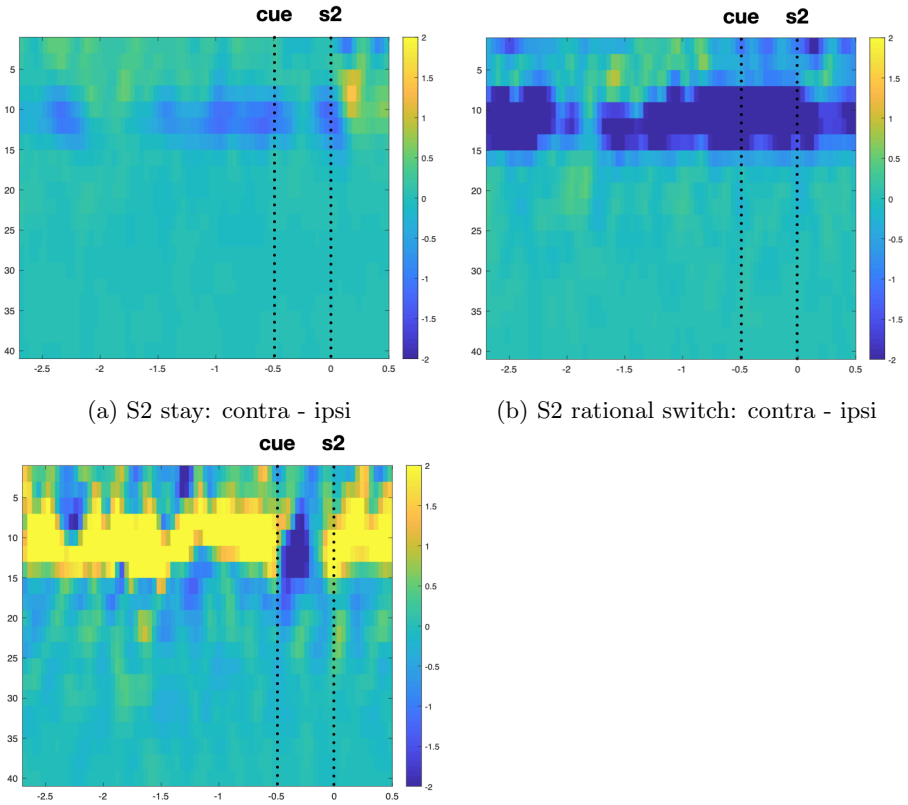


Figure 52: Lateralized VEP in irrational switch trials time-locked to S1 and S2 from 0 to 500 ms post stimulus presentation. Y-axis: beta weights of 'attendL - attendR' from GLM, in arbitrary units proportional to the actual potential. X-axis: time in seconds.

We next examined whether the attention effects were reflected in alpha modulation. We picked two pairs of electrodes from the left and right hemisphere around the parietal-occipital area and calculated the difference between attended and unattended side. Alpha desynchronization (decreased alpha power from the attended relative to the unattended side) was found both in stay and rational switch trials before S2 onset (**Figure 53**), but rational switch showed a much stronger effect. Opposite effect (increased alpha power from the attended side) was found in the irrational switch trials, suggesting lack of enhanced attention or even reduced attention in these trials. However, there was no significant difference between the three conditions (ANOVA: $F(2, 26) = 2.96$, $p = 0.0577$) or any of the pairs (stay vs. rational switch: $t = 1.6501$, $p = 0.1110$; stay vs. irrational switch: -1.2753, $p = 0.2135$; rational switch vs. irrational switch: -1.8142, $p = 0.0812$) across participants.



(a) S2 stay: contra - ipsi

(b) S2 rational switch: contra - ipsi

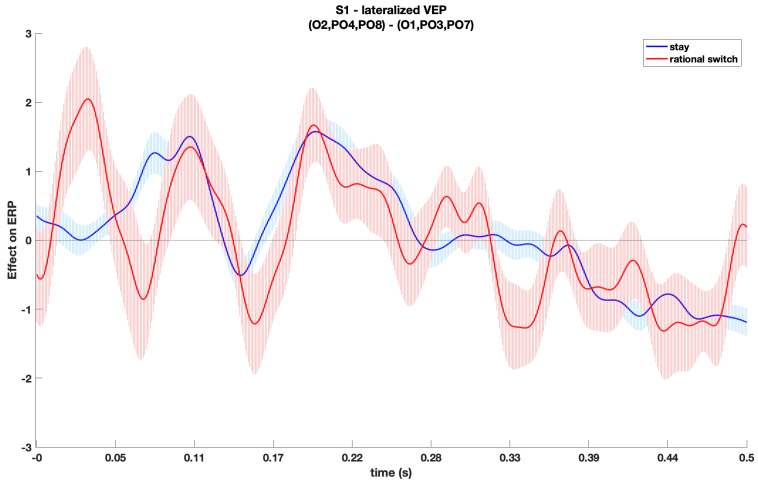
(c) S2 irrational switch: contra - ipsi

Figure 53: Difference in power between contralateral and ipsilateral electrodes around the occipital cortex relative to the attended side, time-locked to S2 stay (a), S2 rational switch (b), and S2 irrational switch (c) trials. Two pairs of electrodes (PO3+PO7 vs. PO4+PO8) from the left and right hemisphere were selected. X-axis: time relative to stimulus onset (s). Y-axis: frequency (Hz).

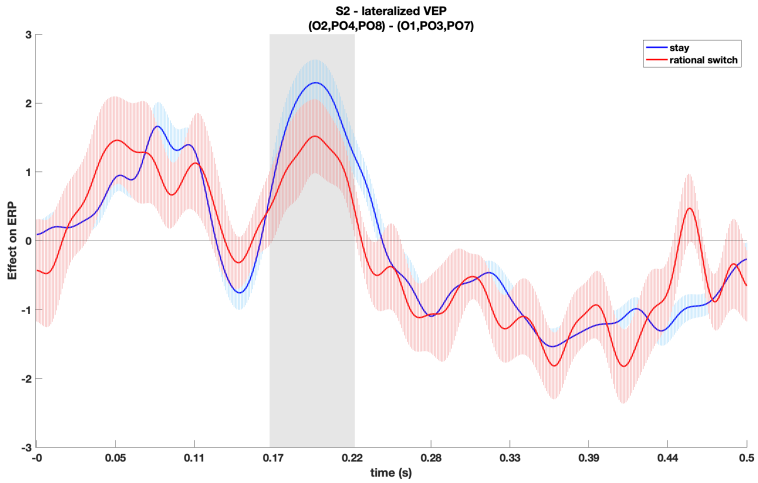
Enhanced attentional effects in rational switch trials at S2

In our behavioural analysis, we noticed a slight decrease in reaction times on rational switch trials. We hypothesized this might be associated with increased attention or arousal on these trials. We found slightly (although not significant) enhanced lateralized alpha desynchronization in the rational switch trials (**Figure 53**) than stay and irrational switch trials, but the effect was not re-

flected by lateralized VEP (**Figure 54**). The VEP lateralization did not differ between stay and rational switch trials at S1 ($t = 1.2$, $p = 0.241$) and decreased in rational switch trials at S2 ($t = 2.4947$, $p = 0.0193$). Therefore, visual selective attention was not enhanced in rational switch trials.



(a) S1



(b) S2

Figure 54: Lateralized VEPs in stay (blue) and rational switch (red) trials, time-locked to S1 (top) and S2 (bottom). Signals were plotted from 0 to 500 ms post stimulus onset. Y-axis: beta weights of 'stay' and 'rational switch' from GLM, in arbitrary units proportional to the actual potential). X-axis: time in seconds.

To summarise, the following conclusions can be derived from the above analyses regarding neural signatures of attentional effects:

- VEP lateralisation was found from 170 to 220 ms around both S1 and S2.
- Lateralised VEP found at both S1 and S2 reflected perceptual accuracy.
- Lateralised VEP at both S1 and S2 reflected WAM (estimated cognitive effort calculated based on difficulty and perceptual accuracy).
- VEP lateralisation was missing in irrational switch trials.
- Lateralised VEP was not enhanced in rational switch trials, compared to stay trials. Therefore, slightly (although not significant) quicker response found in rational switch trials in behaviour was not attributed to enhanced selective visual attention.
- Alpha desynchronisation was found from -1.3 s to S2 onset.
- Alpha power was not modulated by WAM in the run-up to S2 period.
- Alpha desynchronisation did not reflect perceptual accuracy.
- Alpha desynchronisation did not reflect decision types such as stay, rational switch or irrational switch.

3.6.2.2 Results time-locked to response

The time taken to produce a response depended primarily on task difficulty (**Figure 33**) but also varied between switch and stay conditions, being slightly faster for rational switch trials and slower for irrational switch trials. Furthermore there was additional residual variation in response times that could reflect endogenous variables determining response readiness (such as endogenous arousal). We analysed two ERP effects, the lateralized readiness potential and pre-stimulus beta desynchronization, to characterise their variation with factors affecting RT.

Lateralized readiness potential

The lateralized readiness potential (LRP) has been identified as a negative deflection on the contralateral side of the motor cortex relative to the responding hand, indexing the motor activation process and was found to be affected by a series of factors, such as stimulus intensity, stimulus-response congruency and speed stress (reviewed by Kappenman and Luck, 2011). In particular, LRPs have been found to exhibit an early signal of evidence accumulation and a late motor activation process (van Vugt et al., 2014). We hypothesise to see (1) reduced LRP amplitudes on trials where participants were less committed (i.e. switch trials); and (2) LRP amplitudes are negatively related to reaction time. We applied equation 24 to denoised data from -500 to 500 ms around the response. They are similar to the equations for VEPs above, except the attended side was replaced by the responding hand (e.g. hand_L: 1=left hand, 0=right hand):

$$\begin{aligned} LRP = \beta_1 \times hand_L + \beta_2 \times hand_R + \beta_3 \times WAM_L \\ +\beta_4 \times WAM_R + \beta_5 \times Zero_L + \beta_6 \times Zero_R \end{aligned} \tag{24}$$

Additionally, as the reaction time should modulate brain signals and its effect might interfere with difficulty, we built a regression model including the residuals of RT from equation 25 and took difficulty as an additional regressor as specified by equation 26:

$$RT = \beta_0 + \beta_1 \times difficulty \tag{25}$$

$$
\begin{aligned}
LRP = {} & \beta_1 \times hand_L stay + \beta_2 \times hand_R stay + \beta_3 \times residRT_L stay \\
& + \beta_4 \times residRT_R stay + \beta_5 \times difficulty_L stay + \beta_6 \times difficulty_R stay \\
& + \beta_7 \times hand_L rswt + \beta_8 \times hand_R rswt + \beta_9 \times residRT_L rswt \\
& + \beta_{10} \times residRT_R rswt + \beta_{11} \times difficulty_L rswt + \beta_{12} \times difficulty_R rswt \\
& + \beta_{13} \times hand_L irrswt + \beta_{14} \times hand_R irrswt + \beta_{15} \times residRT_L irrswt \\
& + \beta_{16} \times residRT_R irrswt + \beta_{17} \times difficulty_L irrswt + \beta_{18} \times difficulty_R irrswt
\end{aligned}
\tag{26}
$$

Here *difficulty* as a regressor used in equation 25 and 26 was normalised. Same as in VEP analyses, each regressor contained values only in the specified condition, and its values were set to zero for other conditions. The regressor '$residRT$' was obtained by first running a linear regressor between RT and difficulty (with equation 25) to control the confounding effect of difficulty on RT (so that this regressor conveyed only the effect of RT on signals after controlling for difficulty). This regression allowed us to examine if (1) LRPs existed (showed by '$hand_L$' and '$hand_R$') and (2) if response preparation conveyed by LRP reflected reaction time, difficulty and switch.

LRPs reflect response preparation in switch vs stay trials

In our behavioural analysis we noted that reaction times were slightly faster in rational switch trials than stay trials; in irrational switch trials, RTs were slower and did not depend on task difficulty. We therefore asked if the LRP, a response preparation signal, differed across trial types. Topoplots (**Figure 55**) show reduced LRPs in rational switch and irrational switch trials compared to stay. We next computed LRP by subtracting four electrodes in the ipsilateral

hemisphere from contralateral hemisphere around the centro-parietal area and found the LRPs were noisier in two types of switch trials, compared to stay (**Figure 56**), and there was no significant difference in LRPs between stay and either of the switch trials (t-test on LRPs -150 ms to 0 ms: stay vs. rational switch: t = -1.4255, p = 0.1659; stay vs. irrational switch: t = -0.9829, p = 0.3347). Therefore, response preparation was not notably reduced in switch trials but indeed was noisier and may be accountable for differences found in RTs, although the noisiness could be attributed to reduced number of trials in switch conditions. It thus cannot be concluded that response preparation in switch trials was reduced compared to stay trials, as seemingly plausible in topoplots (**Figure 55**). Note that the topoplots were illustrative and used to guide subsequent analyses, but we did not perform statistical analyses to draw any conclusions on them.

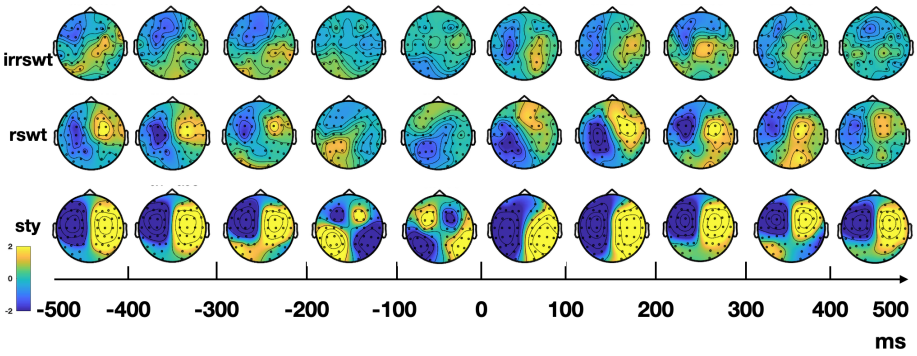
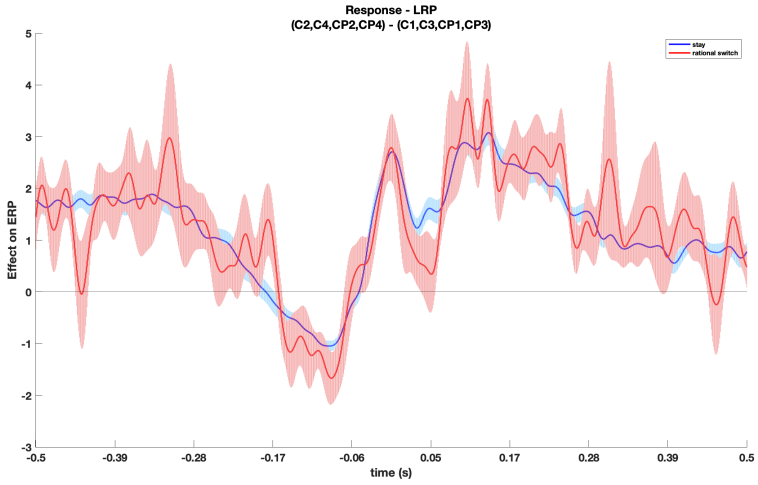
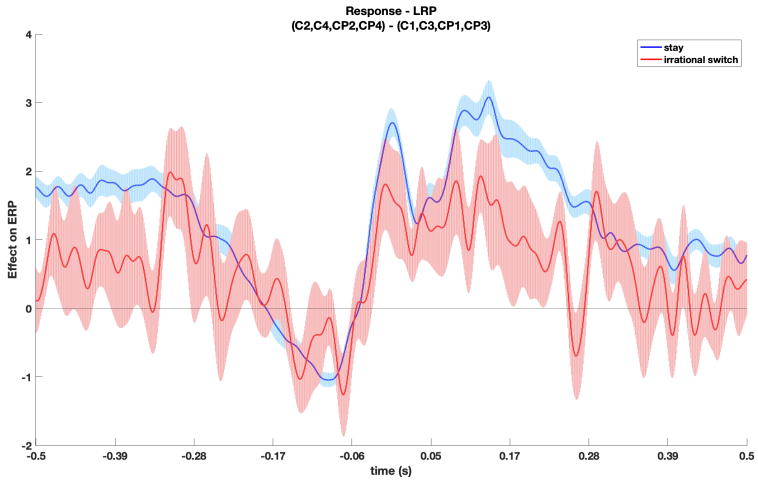


Figure 55: LRPs (left hand minus right hand) around response in stay, rational switch and irrational trials. Topographies were plotted in t-values of beta weights from regression analysis with 50 ms increments from -500 to 500 ms around the event.



(a) stay - rational switch



(b) stay - irrational switch

Figure 56: LRPs from -500 ms to 500 ms time-locked to response compared in stay and rational switch trials (top) as well as stay and irrational switch trials (bottom). Y-axis: beta weights of 'stay', 'rational switch' and 'irrational switch' from GLM, in arbitrary units proportional to the actual potential). X-axis: time in seconds.

LRPs reflect random variations in response speed

Whilst reaction time could be predicted from perceptual difficulty, there were also random variations in reaction time. We reasoned that if the LRP amplitude reflected response readiness, then it should increase in amplitude in trials where the RT was faster than expected based on task difficulty. To test this, we conducted a GLM analysis in which the LRP effect was predicted from the 'residual RT' – the residual after RT was predicted from difficulty. Trials with high residual RT were trials where the response was slower than expected. Topographies (**Figure 57**) have shown that LRP was negatively predicted by residual RT, so trials that were faster than expected had a stronger LRP amplitude, and the effect was present in stay and rational switch trials, but not in irrational switch trials. We then computed the effect of residual RT on LRP with selected electrodes in the centro-parietal area (**Figure 58**). Residual RT negatively predicted LRP in stay (t-test against zero with LRP averaged from -150 to 0 ms: t = -5.2734, p = 1.6356e-05), but not in rational switch (t = -1.9264, p = 0.0651) or irrational switch trials (t =-0.7256, p = 0.4746), and there was no difference between stay, rational switch and irrational switch trials (ANOVA: $F(2, 26) = 0.16$, p = 0.8509). Taken together, these analyses demonstrate that LRPs did not reflect stay/switch decisions, but enhanced LRP reflected faster response (as negatively predicted by RT - shorter RT, stronger LRPs) in stay trials.

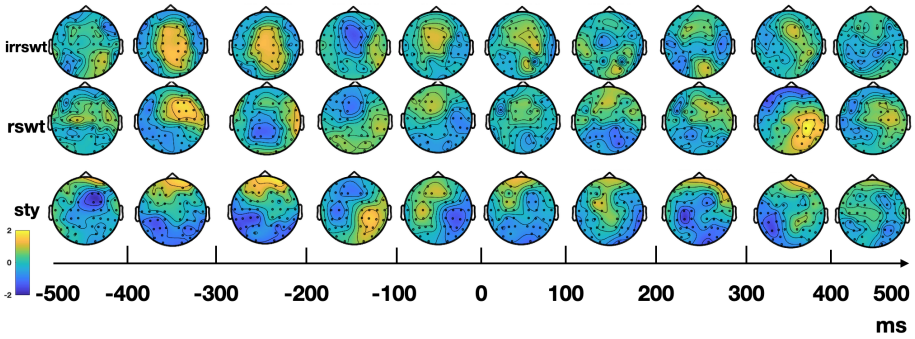


Figure 57: Effect of residual RT (the residual after RT was predicted from difficulty) on LRPs (left hand minus right hand) around response in stay, rational switch and irrational trials. Topographies were plotted in t-values of beta weights from regression analysis with 50 ms increments from -500 to 500 ms around the event.

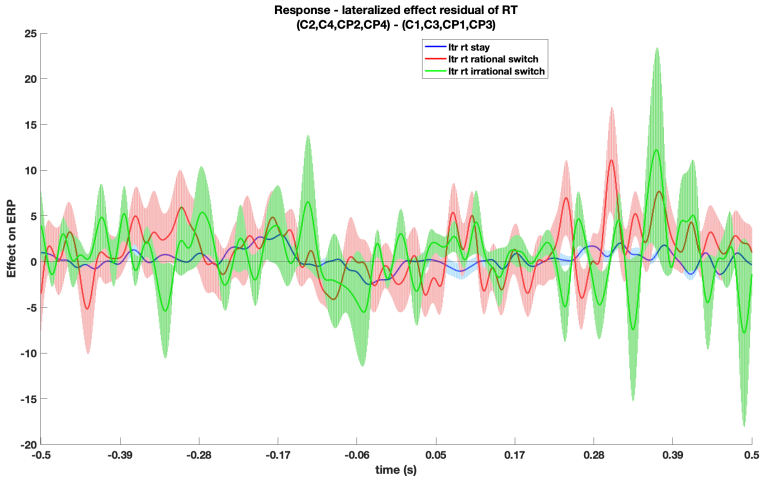


Figure 58: Effect of residual RT on LRPs (averaged (C2,C4,CP2,CP4) - (C1,C3,CP1,CP3)) around response in stay, rational switch and irrational trials. Y-axis: beta weights (e.g. 'RT_stay') from GLM, in arbitrary units proportional to the actual potential). X-axis: time in seconds.

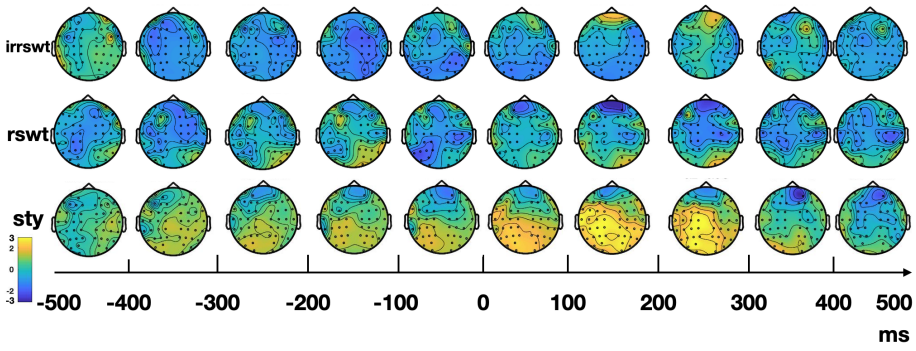
Response-related power changes

To examine the behaviour related power changes, we baseline-corrected signals from -800 to -500ms before the response, and collapsed the signals across beta

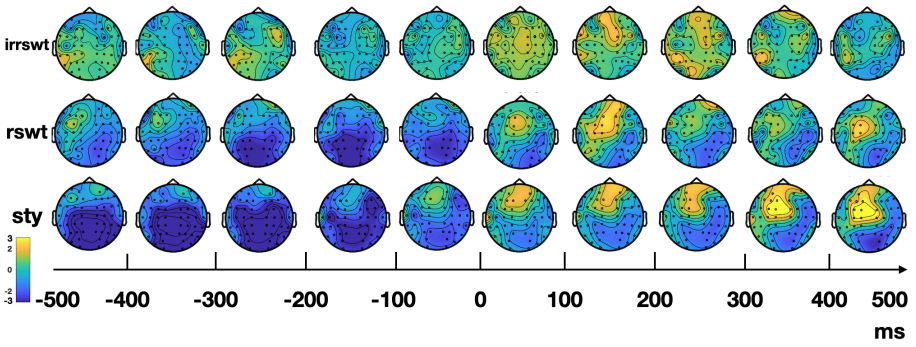
band (14-30 Hz) from -500 to 500ms around the response. To first test the lateralised effect, we performed equation 24 and 26 on the beta power. As the results showed non-lateralized effects, we performed subsequent analyses on the overall beta weights from both hands.

Non lateralized beta power effect in the run-up to response

In speeded reaction tasks like the stop-signal task, response preparation is associated with a bilateralized beta desynchronization in the run-up to the response; stronger desynchronizations are associated with faster reaction times and failure to stop when signalled (J. Wessel, 2020). We therefore asked if there was a lateralized beta desynchronization in the run-up to response in our task. We plotted beta power changes over time around response (**Figure 59(a)**) and did not find any lateralized beta desynchronization in the run-up to response in our task. However, there was a non-lateralized beta desynchronization over the sensorimotor sites in the run-up to response so we analysed that further (**Figure 59(b)**). There is an increase in beta power in the frontal cortex after a response, matching the movement cancellation traits found by Jan Wessel (J. Wessel, 2020). Note that our task is not speeded, and therefore may be disparate from previous studies with speeded tasks.



(a) Lateralized beta



(b) Non-lateralized beta

Figure 59: Lateralized (top) and non-lateralized (bottom) Beta power changes around response split by stay, rational switch and irrational switch. Topographies were plotted in t-values of beta weights from regression analysis with 50 ms increments from -500 to 500 ms around the event.

Non-lateralized beta desynchronization reflects response readiness, which associated with switch readiness

We asked if the non-lateralized beta desynchronization effect that we observed in the run-up to response depended on response-readiness (reaction time). Although rational switch trials were slightly faster than stay trials, this was not reflected by beta desynchronization (**Figure 59(b)**). The non-lateralized beta desynchronization indeed decreased in irrational switch trials, which were much slower than the other two.

Non-lateralized beta desynchronization decreased with residual RT (increased beta power in faster trials) when modelled over all trials, and this effect was greater in stay trials than rational switch trials (**Figure 60**). In irrational switch trials there was no such effect.

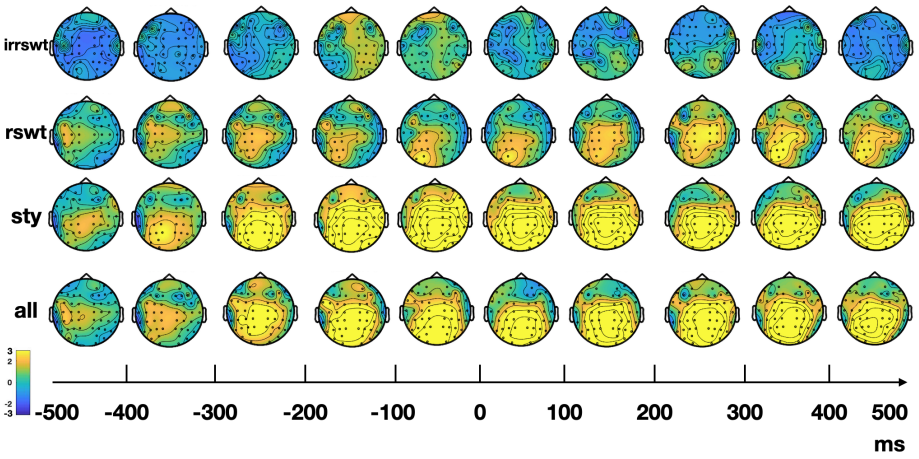


Figure 60: Effect of residual RT on non-lateralized beta power changes around response in all, stay, rational switch and irrational trials. Topographies were plotted in t-values of beta weights from regression analysis with 50 ms increments from -500 to 500 ms around the event.

Difficulty slightly reduced beta desynchronization with data collapsed from all trials, but the effect was mostly driven by stay trials (**Figure 61**). The fact that the effect of RT was much stronger than difficulty suggests that beta desynchronization reflects a general factor or response-readiness, rather than the time taken to generate a response based on an easy vs. difficulty stimulus.

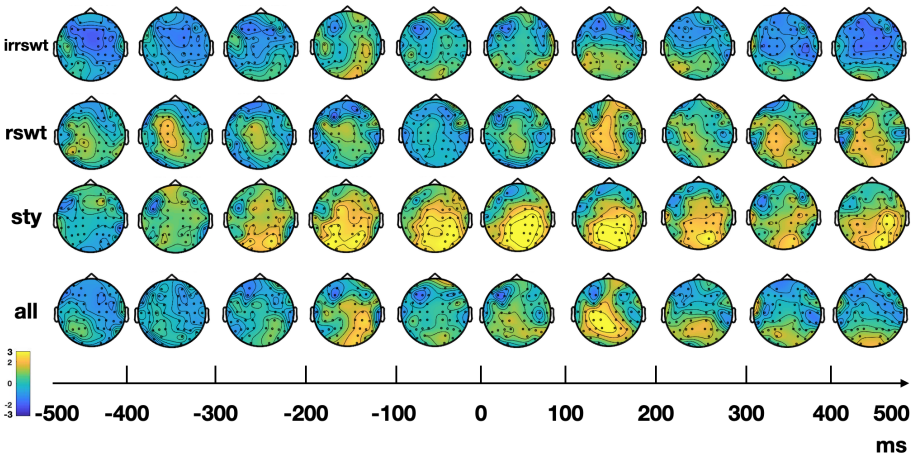


Figure 61: Effect of difficulty on non-lateralized beta power changes around response in all, stay, rational switch and irrational trials. Topographies were plotted in t-values of beta weights from regression analysis with 50 ms increments from -500 to 500 ms around the event.

Response-readiness appears to increase with switch-readiness, as RTs are faster on rational switch trials. In the behavioural analysis we characterised switch readiness in a GLM predicting switch behaviour from a number of factors (**Figure 62**). We hypothesised that beta desynchronization should depend on switch-readiness across all trials, not just be increased on switch trials. We found that beta desynchronization does indeed increase with switch-readiness, but is unaffected by difficulty, even though difficulty has the stronger effect on RT, suggesting that beta desynchronization is linked to the threshold setting process, not the actual speed of response per se.

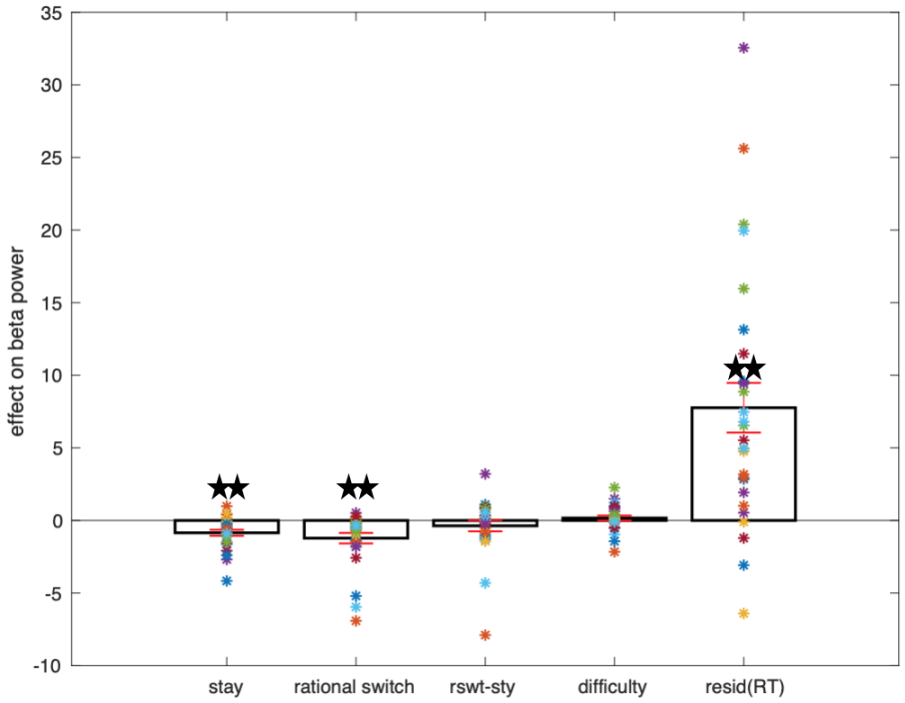


Figure 62: Effect of stay, rational switch, difference between rational switch and stay, difficulty and residuals of RT on beta power around parietal (CP1, CP3, CPZ, CP2 and CP4) area. Beta power (14-30 Hz) was averaged from -200 to 0 ms before response across participants. Double asterisks represent the significant regression weights ($p < 0.01$).

To summarise, the following conclusions are made based on above presented analyses on response related signals:

- LRPs did not differ among decision types (i.e. stay, rational switch or irrational switch).
- LRPs reflected response speed in stay trials (negatively predicted by RT), but not in any of the switch trials.
- Lateralised beta desynchronisation was not found in our task.
- Non-lateralised beta desynchronisation was found in stay and rational

switch trials, and did not differ between the two.

- Non-lateralised beta desynchronisation reflected response speed after controlling for difficulty, but not difficulty itself.

3.6.2.3 Time-locked to feedback

When participants receive feedback in this task, a number of processes are activated. Firstly, the feedback may signify a reward or a no-reward event. Secondly, the reward or reward omission may be more or less expected, based on the difficulty of the trial. Thirdly, the prediction error associated with the feedback must be interpreted in the context of other cues (such as recent reward history) to determine whether to switch sides. We investigated how the EEG response to feedback delivery related to each of these processes.

Feedback related potentials

We first split EEG signals from 0 to 500 ms around feedback by decision type, reward type, difficulty and perceptual correctness to compare the effect of these variables on feedback processing. Additionally we built a regression model to determine the effect:

$$\begin{aligned} FRP = \beta_0 + \beta_1 \times (R - NoR) + \beta_2 \times (R_{\text{swt-sty}}) + \beta_3 \times R_{\text{difficulty}} \\ + \beta_4 \times (NoR_{\text{swt-sty}}) + \beta_5 \times NoR_{\text{difficulty}} \end{aligned} \tag{27}$$

Here *R-NoR* was coded as (+1) for reward and (-1) for no-reward, and *swt-sty* was coded as (+1) for switch and (-1) for stay. The *difficulty* was normalised in both switch and stay conditions.

Reward, difficulty and switch all reflected by post-feedback signals

We first plotted the effects of reward (vs no-reward) and difficulty (parametric effect of difficulty) on the current trial, as well as switch (vs. no-switch) on the next trial on signals in the post-feedback period (**Figure 63**).

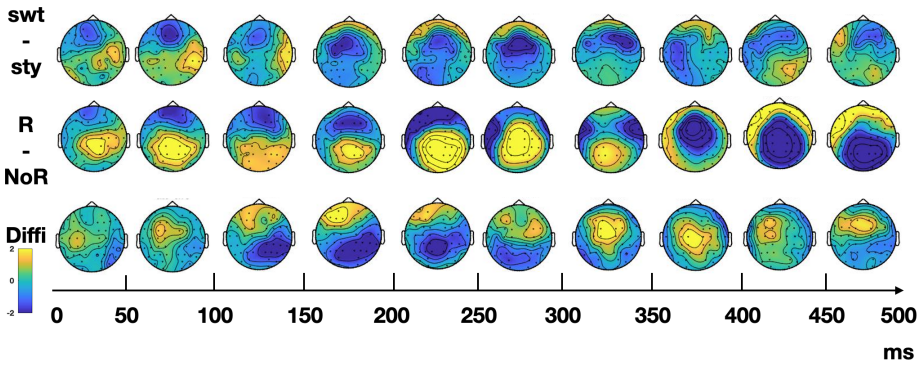


Figure 63: Effects of patch decision (switch minus stay in the next trial), feedback type (reward minus no-reward) and difficulty on ERPs from 0 to 500 ms after feedback onset. Topographies were plotted in t-values of beta weights from regression analysis with 50 ms increments from 0 to 500 ms around the event.

The FRN and P300 components reflect the interaction between feedback and expectation

Prior literature has identified two key feedback-related components: the feedback related negativity (FRN) occurring 200-300 ms post feedback (reviewed by Cohen, Wilmes et al., 2011; Ullsperger et al., 2014), and P300, a positive deflection found around the mid-line centro-parietal areas and peaking at 250 to 350 ms (reviewed by Picton, 1992; Nieuwenhuis et al., 2005). We aim to establish the two components and how the feedback processing they convey reflect task-related factors, such as difficulty, accuracy, switch and prior expectations. We picked 100 ms around the peak of each component found in our data that is close to previously identified time period: 270-320 ms for FRN and 300-400 ms for P300. We first plotted the ERP at frontal electrodes for rewarded and

non-rewarded trials, at each difficulty level – breaking down rewarded trials into those where the participant was perceptually correct (so should have expected a reward if he had perfect insight into his performance) or incorrect (**Figure 64**). Results have revealed a main effect of reward on FRN component (at FCZ round 270-320 ms, t = 7.5706, p = 5.9541e-13 (reward vs. no-reward)), as the difference between rewarded and unrewarded trials is large at this timepoint for all difficulty levels. There was no significant effect of accuracy on the component (t = 0.7042, p = 0.4819 (correct vs. incorrect)). Difficulty alone did not affect the component (ANOVA: F(4, 26) = 1.79, p = 0.1293). However, there was an interaction between reward and accuracy (F(1, 26) = 7.19, p = 0.0076). No interaction was found between reward and difficulty (F(4, 26) = 0.89, p = 0.4682) or accuracy and difficulty (F(4, 26) = 0.24, p = 0.9142). It seems in easier trials the no-reward but accurate condition evoked relatively larger deflections, both between 270-320 ms and later 400 to 500 ms, and the largest deflections at around 400 ms was found in the no-reward but correct condition in the easiest trial (**Figure 64**). In easy trials, participants could more easily be aware of incorrect responses, so should be more surprised if they were unrewarded for a correct, easy trial – these conditions drive switching behaviour (**Figure 32**) suggesting this component may be associated with the decision to switch. However, explicitly breaking trials down into switch- and stay (**Figure 65**) revealed no clear effect of switch in this time frame (stay vs. switch: t = 0.7042, p = 0.4819). There was an interaction between reward and switch (F(1, 26) = 54.15, p = 0), but no interaction between reward and difficulty (F(4, 26) = 0.09, p = 0.9853) or switch and difficulty (F(4, 26) = 0.9, p = 0.466). Taken together, results suggest that the FRN component was modulated by reward itself, an interaction between reward and accuracy, as well as an interaction between reward and switch.

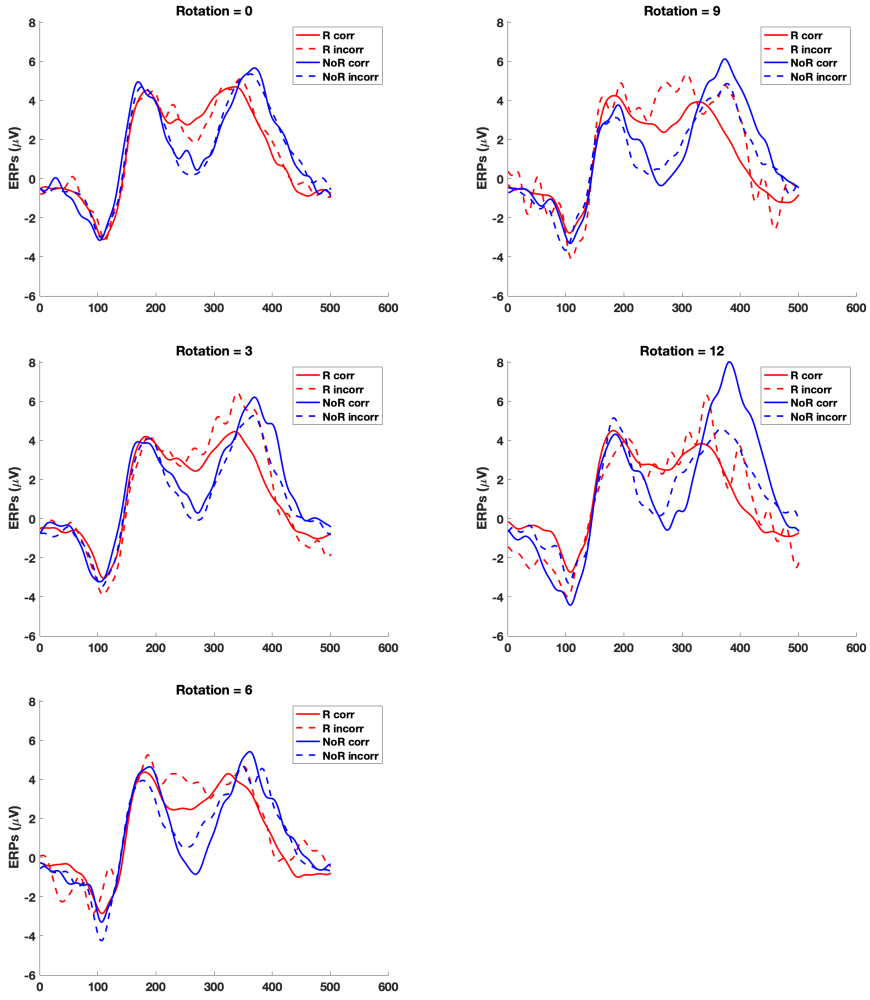


Figure 64: ERP of FCZ after feedback onset split by reward (R) / no-reward (NoR) and perceptual correctness plotted at five levels of difficulty (rotation angle from 0 to 12°).

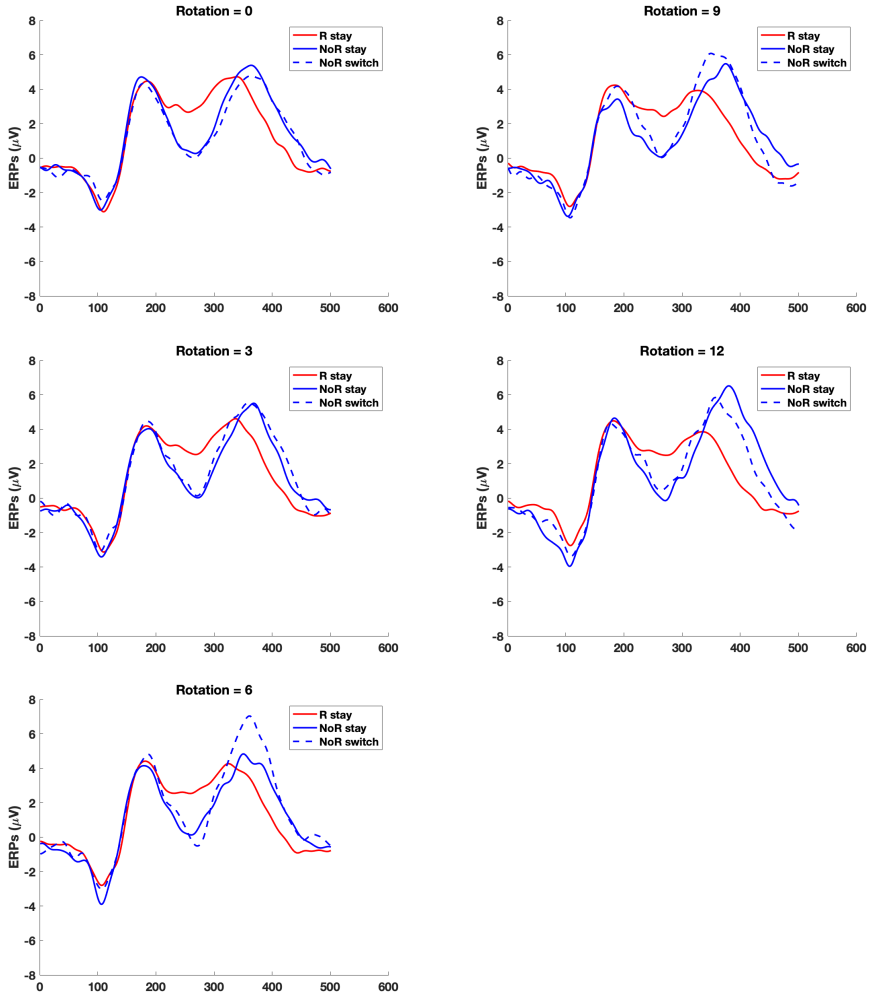


Figure 65: ERP of FCZ after feedback onset split by reward (R) / no-reward (NoR) and stay/switch plotted at five levels of difficulty (rotation angle from 0 to 12°). We removed "rewarded and switch" case because people rarely switched after a reward, and the curve looked rather noisy.

Next, we sought to establish the same effect on P300 component by plotting signals at CPZ within the same time frame split by correctness (**Figure 66**) and switch (**Figure 67**). The results show reduced peak amplitude between reward and no-reward trials, compared to what was found at FCZ, and the peak of no-

reward trials occur later than reward trials with approximately 100 ms delay. Notably there was a large N1 found with rewarded incorrect trials when rotation equalled 9 degrees (**Figure 66**). There was no significant effect of reward on signal averaged in 300-400 ms ($t = 1.8093$, $p = 0.0715$). Accuracy alone had an impact on the signals ($t = 2.9659$, $p = 0.0033$). So did difficulty ($F(4, 26) = 4.04$, $p = 0.0031$). There was no interaction between reward and accuracy found ($F(1, 26) = 0.36$, $p = 0.5504$), or between reward and difficulty ($F(4, 26) = 1.82$, $p = 0.1226$), or between accuracy and difficulty ($F(4, 26) = 1.47$, $p = 0.2095$). In terms of switch (**Figure 67**), signals significantly reflected switch (stay vs. switch: $t = 5.9328$, $p = 9.1322e\text{-}09$). There was an interaction between reward and switch ($F(1, 26) = 59.28$, $p = 0$), but no interaction between reward and difficulty ($F(4, 26) = 0.18$, $p = 0.9478$), or between switch and difficulty ($F(4, 26) = 0.91$, $p = 0.4559$). Taken together, P300 component was modulated by accuracy, difficulty, switch, as well as an interaction between reward and switch.

It seems both FRN and P300 were related to the processing of feedback, perceptual accuracy, and modulated commitment to the current patch, whereas P300 modulation involved additional reflection on difficulty, an effect that was not found with FRN.

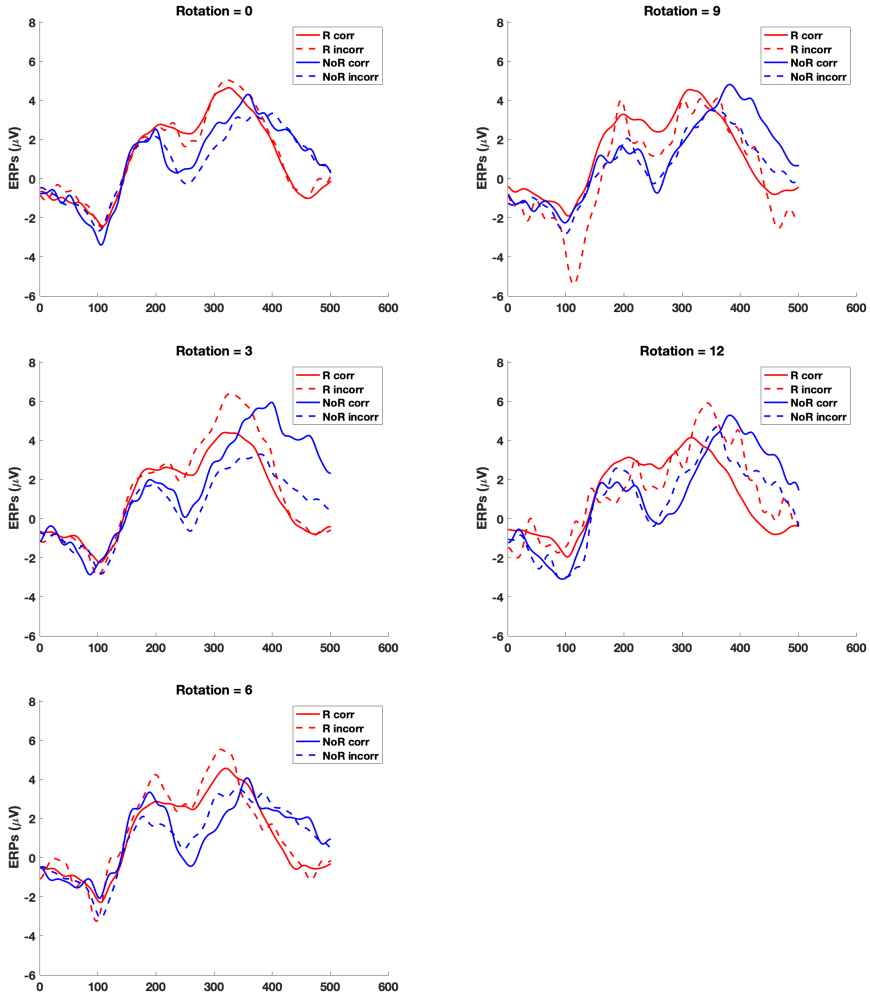


Figure 66: ERP of CPZ after feedback onset split by reward (R) / no-reward (NoR) and perceptual correctness plotted at five levels of difficulty (rotation angle from 0 to 12°).

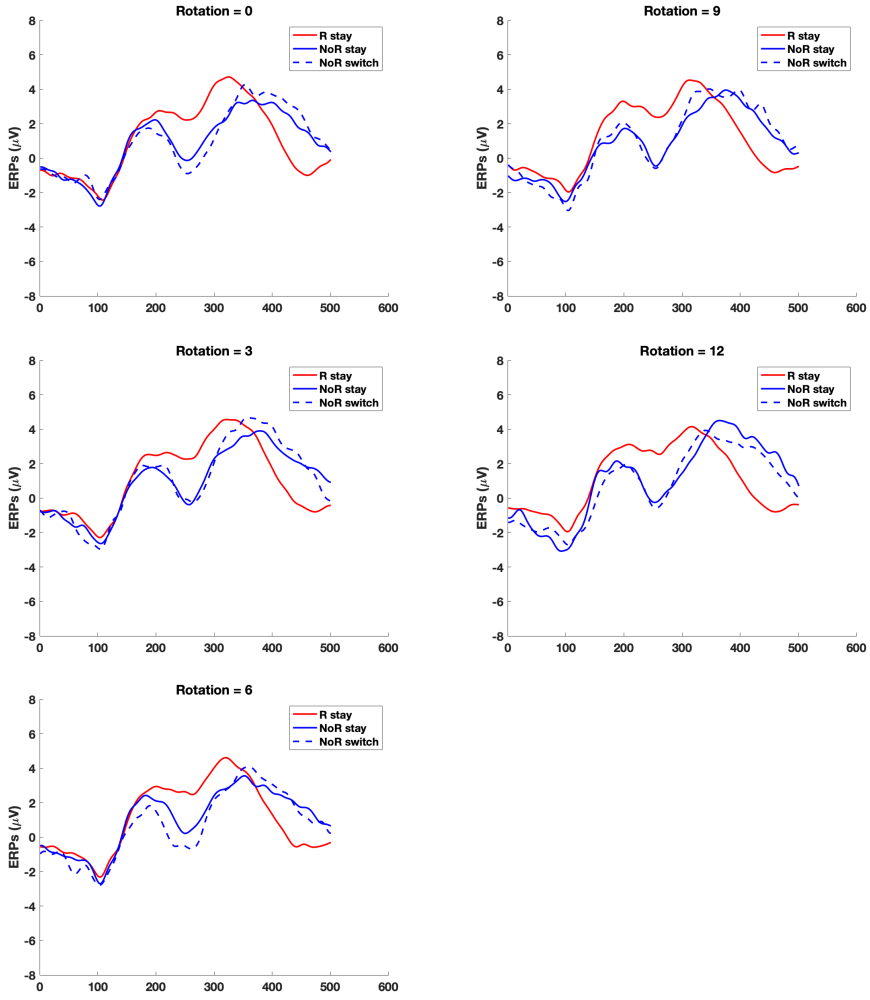


Figure 67: ERP of CPZ after feedback onset split by reward (R) / no-reward (NoR) and stay/switch plotted at five levels of difficulty (rotation angle from 0 to 12°). We removed "rewarded and switch" case because people rarely switched after a reward, and the curve looked rather noisy.

Switch evidence builds up across trials

As we hypothesise that the switch decision was not based on feedback from a single trial, but rather built upon accumulated evidence within a block, and the behavioural results revealed reaction time and accuracy adjustments around a

switch point, we next examined whether the behavioural changes were reflected by EEG signals. As previous analyses showed only a rational switch conveyed important information about participants' careful consideration about switch, only rational switch trials were included in the following analyses. In particular, we focused our analyses on patterns similar to FRN and P300 components. We first looked at signal changes from six electrodes from the fontal-central area at 270-320 ms (in accordance with FRN component as reviewed by Ullsperger et al., 2014) (shown by **Figure 68**) and found the signals slightly built up before a switch but the trend was not significant (t-test against the baseline (5th trial after a switch): -4: t = 0.2112, p = 0.8344; -3: t = 0.1925, p = 0.8488; -2: t = 0.7224, p = 0.4765; -1: t = 1.6539, p = 0.1102; 1: t = 0.2482, p = 0.8059; 2: t = 0.1219, p = 0.9039; 3: t = -0.3633, p = 0.7193; 4: t = -0.8603, p = 0.3975). We then tested the same effect with electrodes from centro-parietal area at 300-400 ms (in accordance with P300 component as suggested by S. Sutton et al., 1965). As **Figure 69** shows, signals around the centro-parietal area gradually built up from 4 trials prior to a switch, and returned to baseline after a switch (-4: t = -0.3687, p = 0.7153; -3: t = 1.1994, p = 0.2412; -2: t = 2.7633, p = 0.0104; -1: t = 3.4252, p = 0.0021; 1: t = -1.1916, p = 0.2442; 2: t = 0.3612, p = 0.7209; 3: t = 0.2181, p = 0.8290; 4: t = 1.0005, p = 0.3263). The results suggest P300 component modulation may serve as indices of switch evidence that built up across trials.

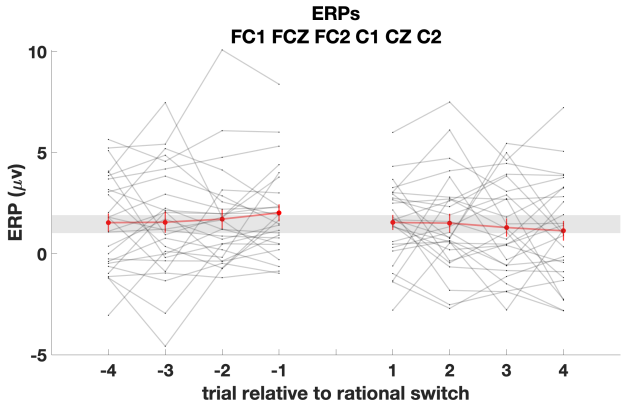


Figure 68: Event-related potentials averaged from six electrodes from frontal-central area (FC1, FCZ, FC2, C1, CZ, C2) 270-320 ms after the feedback presentation from four trials before and four trials after a rational switch. Only unrewarded trials were selected. Individual data are represented by the grey lines and data averaged across subjects are shown by the red line. The shaded grey bar represents the baseline voltage (5th trial after a switch).

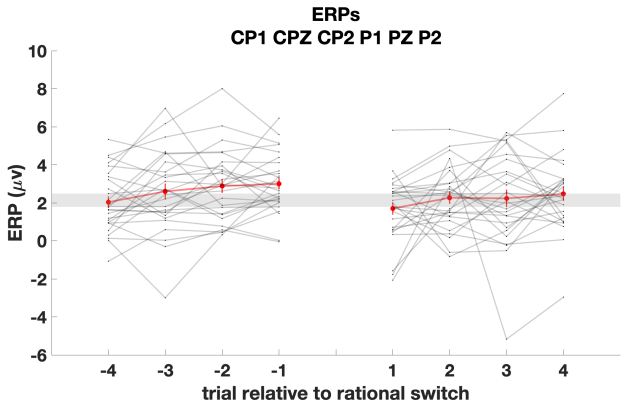


Figure 69: Event-related potentials averaged from six electrodes from centro-parietal area (CP1, CPZ, CP2, P1, PZ, P2) 300-400 ms after the feedback presentation from four trials before and four trials after a rational switch. Only unrewarded trials were selected. Individual data are represented by the grey lines and data averaged across subjects are shown by the red line. The shaded grey bar represents the baseline voltage (5th trial after a switch).

Responses to feedback reflect prior expectations and assessment of one's own performance

Although the perceptual confidence was not explicitly measured, it could be estimated by prior expectations and likelihood derived from the Bayesian model, given that the model successfully predicted both ground truth and participants' behaviour. We next looked into whether feedback related signals were reflected by Bayesian predictions by binning the effect of prior expectations and likelihood (probability of receiving reward given having picked the correct patch, which was equivalent to the chance of being perceptually correct) into two levels split by mean and looking at the no-reward - reward comparison. Topographies (**Figure 70**) revealed a positive centro-parietal component at around 300-400 ms for both high prior and high likelihood and a negative component for the low likelihood case. For the low prior case the negative component was slightly shifted later to 350-500 ms. Both high prior and high likelihood indicate high expectation of getting a reward, and revealed the component might reflect 'surprise'. There was also a flipped effect around the frontal pole between the low and high prior comparison, indicating the area might be related to an early effect prior expectation on the feedback processing.

To further dissociate the effect of high and low reward expectation on the feedback processing, we binned the Bayesian likelihood and prior into two levels and plotted their effects on the feedback related ERPs, as shown by **Figure 71**. The early effect of high Bayesian likelihood was significantly lower than the low likelihood on CPZ during the 100-200 ms period ($t = -3.0152$, $p = 0.0057$) but the difference was almost driven by the negative effect of the high likelihood. The later effect flipped during the 300-400 ms period ($t = 2.5460$, $p = 0.0172$) and again was driven by the negative effect of the low likelihood. The effect of high Bayesian prior was significantly higher than the low prior during 300-400

ms (t = 3.1243, p = 0.0043).

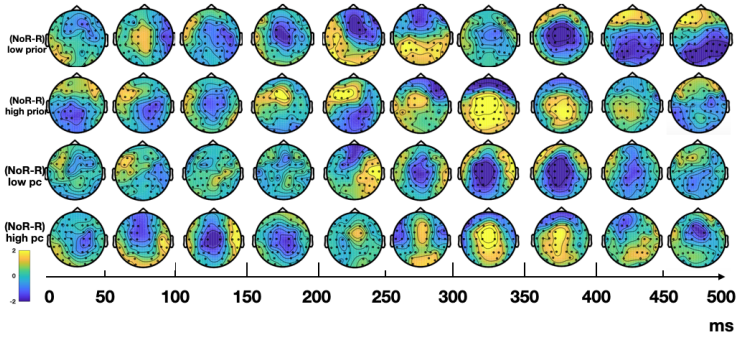
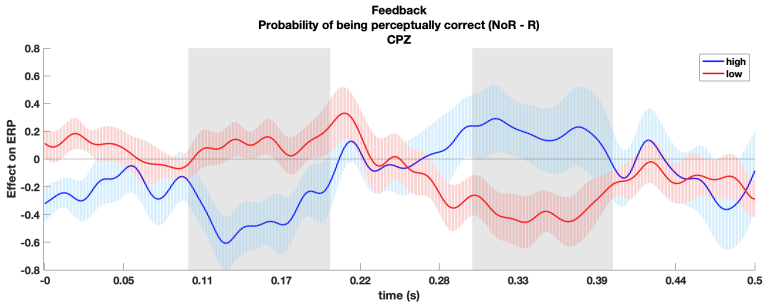
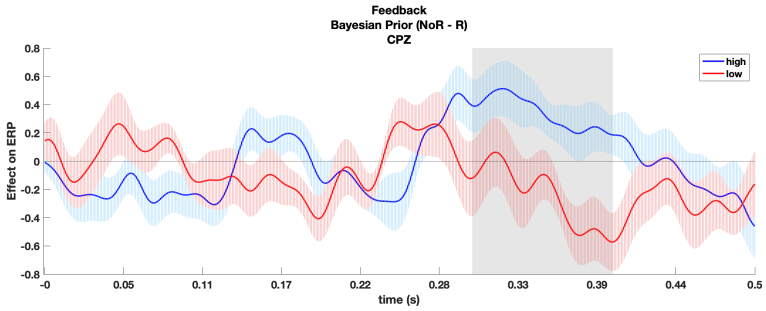


Figure 70: No-reward and reward comparison between low (below average) and high (above average) Bayesian prior and likelihood (probability of receiving reward given having picked the correct patch) on feedback signals. Topographies were plotted in t-values of beta weights from regression analysis with 50 ms increments from 0 to 500 ms around the event.)



(a) Bayesian likelihood



(b) Bayesian prior

Figure 71: Feedback related ERPs binned by high and low Bayesian likelihood and prior, plotted from 0 to 500 ms post feedback presentation. Y-axis: beta weights obtained from GLM using Bayesian likelihood ad Bayesian prior as predictors, respectively.

As a summary, we draw the following conclusions about FRN component (270-320 ms at FCZ) from our analyses:

- FRN was modulated by reward itself, with the component being larger in rewarded trials.
- It was modulated jointly by reward and accuracy (by an interaction of the two).
- No main effect or interaction effect related to difficulty was found on FRN.
- It did not reflect switch itself, but did an interaction between reward and switch.

- It did not reflect accumulated switch evidence.

The following conclusions about P300 component (300-400 ms at CPZ) can be derived:

- P300 was modulated by reward itself, being larger in rewarded trials.
- It was modulated jointly by accuracy and difficulty (by an interaction of the two).
- It reflected switch itself (reduced amplitudes in switch than in stay, regardless of reward), as well as an interaction between reward and switch.
- It reflected built-up switch evidence across trials prior to a switch decision.
- It reflected binned Bayesian likelihood (expected accuracy) and Bayesian prior (expected right side), with a significant difference found between high and low expectations.

3.6.3 EEG Discussion

We performed general linear model (GLM) approach on event related potentials (ERPs) and time frequency data around four time points: the first and second visual stimuli presentation, the response and feedback.

We found enhanced visual evoked potentials (VEP) on the contralateral hemisphere relative to the attended visual field 170-220 ms post stimulus onset. The lateralized VEP was also stronger in perceptually correct and higher WAM (our customised dummy regressor capturing both accuracy and difficulty) trials, suggesting the connection between lateralized VEP and selective attention. As the participants were instructed to fixate gaze on the centre and their eye movements were minimised by the eye tracker, the lateralized VEP reflects covert attention, aligned with previous studies reporting the connection between lateralized VEP and performance in covert attention tasks (Eimer et al., 2006; Hilo-Merkovich et al., 2018). The time window we chose was 170-220 ms where the largest peak was found, slightly different from but in general matching the first half of the 'late' period (180-300 ms) identified by Hilo-Merkovich et al., 2018 and the second time window (150-200 ms) identified by Eimer et al., 2006. We also found decreased alpha power on the contralateral hemisphere relative to the attended side during the period when participants anticipated the second stimulus onset. The lateralized alpha desynchronization did not significantly reflect perceptual accuracy, difficulty or switch, although in irrational switch trials the alpha desynchronisation was clearly absent. Previous studies addressing the cognitive processes underlying alpha desynchronization proposed two theories: (1) it reflects retrieving semantic content of information from the long term memory (Klimesch, 1999); (2) it is associated with anticipatory attention when the brain maintains the target information in the attentional buffer, a selectively activated part of the knowledge system (Klimesch, 2012). Our find-

ings thus provide additional evidence for the anticipatory attention theory and hopefully can shed light on the understanding of cognitive processes underlying alpha power.

We observed clear lateralized readiness potential (LRP) prior to response in stay, rational switch and irrational switch trials, but no significant difference in the height of LRP across three types of trials. Moreover, LRPs reflected random variations in response speed (over and above those predicted by difficulty and switch/stay trial type), but only in stay trials. The LRP was found to reflect accumulated evidence crossing the decision threshold and the height of LRP relates to changes in response bias (van Vugt et al., 2014). It is possible that in our study the evidence accumulation process was noisier in switch than in stay trials, whereas the decision threshold was the same, and that explains why LRP reflected response speed only in stay trials. We found non-lateralized beta desynchronization in stay and rational switch trials and the beta desynchronization decreased with residual RT but did not depend on difficulty. We also found post response increase in beta power over the frontal scalp sites in stay and rational switch trials. These findings parallel the properties of movement initiation (beta desynchronization over sensorimotor sites) and movement cancellation (beta synchronisation over frontal sites) addressed by previous studies (J. Wessel, 2020; McFarland et al., 2000; Pfurtscheller et al., 2003; Swann et al., 2009; Swann et al., 2011; Swann et al., 2012; Picazio et al., 2014; J. Wagner et al., 2018). These results also suggest that non-lateralized beta desynchronization reflects response readiness, which is associated with switching readiness, and is linked to the threshold setting process, rather than the actual speed of response per se. In other words, participants did not accumulate information faster as predicted by difficulty, but actually lowered their response threshold in some trials.

Previous literature identified a P300 component occurring around the midline centro-parietal regions (FZ, CZ and PZ) and peaking at 250 ms or 350 ms after stimuli presentation encodes 'valid information' and covaries with the choice certainty (reviewed by Picton, 1992). A special type of ERP appearing 200-300 ms after visual onset of performance feedback was also called feedback related negativity (FRN) and is believed to reflect feedback guided learning (reviewed by Cohen, Wilmes et al., 2011). We therefore sought to establish the feedback related potentials (FRP) in our task. We found the FRP was modulated by three factors: the feedback type, switch and difficulty. There was a large positive impact of reward versus no-reward on FCZ at 280 ms post feedback presentation, matching the classical FRN component but it does not seem to interact with correctness or difficulty, except in the easiest trials there was a large positive deflection of no-reward, possibly reflecting a 'surprise' signal. The fact that FRN was modulated by reward and accuracy, but not by switch or difficulty suggests FRN might reflect a prediction error. The cognitive meaning of FRN in the literature has been under debate: some argue that it reflects the magnitude or probability of feedback (Chase et al., 2011; Bellebaum et al., 2010; Talmi et al., 2012; Bellebaum and Daum, 2008; Gehring and Wiloughby, 2002; Goyer et al., 2008; Holroyd and Coles, 2002; Ichikawa et al., 2010; Nieuwenhuis et al., 2004; Pfabigan et al., 2011; Yasuda et al., 2004) and several others suggest it reflects a simple 'good-bad' evaluation (Hajcak et al., 2006; Holroyd et al., 2006; Philiastides et al., 2010; Sato et al., 2005; Toyomaki and Murohashi, 2005; Yeung and Sanfey, 2004). Our results tend to support the learning-unlearning framework proposed by Cohen, Wilmes et al., 2011: a positive feedback generally suggests good performance and indicates replicating the response; not receiving a reward given perceptual accuracy rings a bell of 'picking the wrong patch' and requires behavioural adjustment. Both cases

suggest learning from the trial evoked the highest FRN. P300 reflected extra processing of switch and difficulty, in addition to reward and accuracy. It also reflected accumulated switch evidence across trials, as well as expected accuracy and reward contingency estimated by the Bayesian model. This may indicate deeper cognitive processes the component conveys.

As the Bayesian model captured the behavioural characteristics well and gave estimates of choice confidence that served as essential parts of the meta-cognitive evaluation in the task, we next sought to establish the association between feedback related signals and Bayesian prediction. We found that high expectation on accuracy had an early (100-200 ms) positive effect on the signal at CPZ, whereas the the low expectation on accuracy displayed a late negative effect (300-400 ms). The expectation on the correct side showed an overall late (300-400 ms) effect. The early effect of high expected accuracy indicates when participants were very sure they were perceptually correct but did not get a reward, the 'surprise' came quite early and pointed to a straightforward conclusion: they picked the wrong side. When the expected accuracy was low, however, it entered the second level evaluation together with the expectation on side: did I pick the wrong side or reported the wrong rotation? Our results show the frontal and centro-parietal area are accountable for these two stages of feedback evaluation. The neural signatures underlying choice confidence found in our study add evidence to the findings from two other studies revealing that neurons in the parietal and orbitofrontal cortices encode the subjective confidence associated with choices (Kepecs et al., 2008; Kiani and Shadlen, 2009).

These results about P300 are consistent with a series of findings reviewed by Donchin and Coles, 1988, which discussed previous studies with regards to functional significance of P300 and presented a context updating model. In a paired associate learning task conducted by Horst et al., 1980, the amplitude of P300

was found to inversely correlate with participants' subjective confidence in the correctness of their responses and the actual outcome of the trial. In our study, P300 reflected expected accuracy and expected reward contingency estimated by the Bayesian model. Although model estimates may not accurately represent participants' subjective confidence, our results seem to support the plausible link between P300 and choice confidence and future research may examine this link combining confidence rating with computational models. In several other studies examining P300 and task relevance (Diamond et al., 1986; Gopher and Donchin, 1986), lower amplitudes of P300 were found when participants were occupied by another demanding task. In our task, lower amplitudes of P300 were found prior to switch in comparison to staying with the same patch regardless of reward. It is possible that when participants decided to switch, their attentional resources were allocated to the other patch, leading to reduced P300 amplitudes on the current trial. In Johnson and Donchin, 1982's oddball task in which the probabilities of tones reversed in midcourse, the amplitude of P300 increased when participants tried to identify the environment had changed. In our task, following unrewarded feedback, the amplitude of P300 climbed up prior to a switch. It seems the change of P300 amplitudes served as an index of accumulated switch evidence. The results we report are aligned with previous findings reviewed by Donchin and Coles, 1988 and therefore it is of our interest to refer to the context updating model proposed by them: that P300 is elicited by cognitive processes involved when the brain tries to represent a model of the environment and larger amplitudes of P300 are associated with larger changes in the model. However, as Donchin and Coles, 1988 clarified, P300 could be a byproduct of a cognitive process, rather than the cognitive process itself. Nevertheless, the results we report in our study bring up interesting possibilities of P300's functional role in accordance with previous findings and we agree with

Donchin and Coles, 1988's notion that understanding P300's functional significance was not for post hoc explanations, but for its prediction on participants' behaviour. Future research may use decoding approaches we applied in study 1 on P300 to examine its predictive power on behaviour to further understand the cognitive processes the component really conveys.

Although it is not possible to separate FRN and P300 as the timewindow we chose had a 20 ms overlap, we do believe the results shed light on some different cognitive aspects revealed by the two components because (1) the analyses on components were performed on different electrodes (with FCZ for FRN, as it was typically found in frontal-centro area (reviewed by Ullsperger et al., 2014); CPZ for P300 as typically found in centro-parietal area (S. Sutton et al., 1965); (2) components were found to reflect different factors, as FRN seemed to reflect a simple accuracy - reward evaluation, whereas P300 involved further reflection on difficulty and switch, in addition to reward and accuracy. These findings seem to corroborate previous theories about FRN's role in learning-unlearning framework and P300's role in context updating.

3.7 Conclusion

This study examined hierarchical perceptual decision making with a visual perception task, where participants had to infer reward contingency and report rotation of the target patch based on confidence in their own perceptual judgement as well as experience across trials. Participants were able to achieve an average accuracy of 80% throughout the task. Behaviour models suggested the high-level strategy followed a probabilistic inference as categorised by a simple Bayesian Learner, whereas the low-level sensory evidence integration fell out of the description of a classic drift diffusion model, probably due to low sample size and interacting processes between the two decision hierarchies. Visual evoked potentials and alpha power changes around the occipital area correlated with perceptual accuracy and switching, indicating task related attention modulations. We found a link between lateralised readiness potential and response speed, as well as the association between beta desynchronization and response threshold setting process. We also found feedback related negativity modulated by reward and accuracy, indicating its role in a learning-unlearning framework, and that P300 reflected reward, switch, accuracy and difficulty, in accordance with a previously proposed context updating model. Lastly, combining Bayesian model and EEG analyses, we found signals around the parietal and orbitofrontal areas related to hierarchical decision evaluation of choice confidence and integrated experience, although further research with source reconstruction is needed to pinpoint the neural representation. The study enable us to understand strategies and essential neural correlates involved in hierarchical perceptual decision making tasks.

4 General Discussion

Human decision making is complicated. The study of decision making reflects the interdisciplinary nature of science, with emerging evidence and theories from multiple subjects, including economics, psychology, neuroscience and computational biology. In preceding chapters I introduced two studies that tackled behavioural and neural mechanisms of human decision making with a virtual foraging task (chapter 2) and a metacognition task (chapter 3), with a combination of mathematical models and EEG. Both studies examine hierarchical decision making, which is not a one-stage decision based on a simple calculation of expected values or simple sensory perception, but to integrate historic evidence, dynamically adjust estimations on reward structures, compare choice options and monitor performance over time. The reward structures of both studies were dynamic, in that they varied according to pre-determined patterns and it was essential for participants to track the dynamic changes to excel in the tasks.

Both projects tackled the question of how people decide whether to change where they are or what they do. Specifically, *Gold Rush* examined when to quit a patch in a continuous time framework. The results demonstrated participants exploited the patch that maximised reward rate and left actively to explore when the reward rate declined to a certain level as predicted by MVT (Charnov, 1976). The finding is consistent with previous studies about exploration-exploitation trade-off (Daw et al., 2006; Bromberg-Martin and Hikosaka, 2011; Wilson et al., 2014) as well as patch leaving decisions (Kolling et al., 2012; Hayden et al., 2011; Meder et al., 2017; Wittmann et al., 2016). The *metacognition* project tackled how to evaluate the quality of perceptual evidence and one's own confidence, to inform how much it should be taken into account in switching decisions. Participants achieved overall 80% accuracy in this experiment, suggesting they were

able to use metacognition to inform a decision, aligned with previous findings in hierarchical perceptual decision tasks (Purcell and Kiani, 2016; Bahrami et al., 2010; Hanks et al., 2011). Together the experiments show people compared values of options for *within-state* decisions and tracked and evaluated across trial evidence for *state-change* decisions (as characterised by Kolling and O'Reilly, 2018). Next I will discuss how the findings from behavioural models and neural representations fit and extend previous theoretical frameworks of learning and decision making.

4.1 Three Learning Models

We built three behavioural models to capture three categories of learning processes in the tasks: the reinforcement leaning model for trial-and-error learning, the drift diffusion model for sensory evidence integration, and the Bayesian Learner for probabilistic inferences. I will now highlight the empirical evidence from our projects related to these learning models, as well as important considerations derived from the findings.

4.1.1 Learning through reinforcement

The *Gold Rush* involves what is known as the *value-based decision making*, a type of decisions guided by idiosyncratic preferences (Glimcher, 2014). Since the patch value was not explicitly given but to be learned through experiences, it was of our interests to understand the learning processes. The average reward rate of a patch should reflect the patch's quality and did not include past travel times, which were randomly drawn in each trial. It appeared to be tracked through the simple reinforcement learning (RL) model we built. It turned out the reward history derived from RL predicted patch learning and the reward history was also decoded from EEG signals around patch choice, patch leaving and feedback.

It helps answer an essential question in foraging: if the optimal leaving time is determined by comparing the current reward rate with the average reward rate in the overall environment (also excluding travel times) as modelled by MVT, how does an organism learn the average reward rate? Based on our results, we suggest that reward history serves as an overall estimate of the environment and is used to compare against the option in hand (i.e. the chosen patch) in determining when to leave. Note that unlike previously studied foraging scenarios where patch values were explicitly presented and fixed throughout a trial (Hayden et al., 2011; Kolling et al., 2012), in our experiment we added dynamic changes to patches. We propose the average reward rate is used to inform future decisions of when to leave. Overall this shows 1) patch leaving can be based on tracked reward experiences to infer future reward rates; 2) the average reward rate is flexible and neither simple time passing nor all task relevant time (as it excludes travel time) suggests it can be adapted to the circumstances.

Since we fitted a fixed learning rate to the data, the study is unable to capture how learning rates were adjusted according to the structure changes, as previous research suggested the learning rate varied as volatility of the environment (Behrens et al., 2007). However, although the exact values or direction of changes involved randomness, the rate of changes was relatively stable (i.e. reversal happened every 13-23 trials), and it was possible that the volatility of environment was seen as equal and the same learning rate was kept. Future research may examine how learning rates are adjusted with dynamic reward structures when the change rate is variable.

4.1.2 Learning through sensory evidence integration

In the second project we examined *metacognition* with a hierarchical visual perception task, in which participants needed to report rotation of a target

gabor patch. Although the decision included inference of the target patch, the basic level of cognition was still categorised into a simple visual discrimination decision. We built a drift diffusion model (DDM) that has been normally used to model decision processes involved in two-alternative discrimination tasks (Ratcliff and McKoon, 2008). The model posits that the evidence starts accumulating at a certain time point after the stimulus onset, and gradually builds up with a speed determined by drift rate. Once the evidence hits one of the decision boundaries, the choice is made and the decision process ends. The results captured accuracy and RT well in stay and rational switch trials, but contradicted the behavioural observations in irrational switch trials. It is possible that the parameters fitting requires large samples and after splitting data into three conditions, few data (15 out of 27 participants) remained in irrational switch trials and such drastic reduction in sample size led to unreliable fitting results, compared to the other two conditions. Another possibility is that the evidence integration process interacted with high-level cognition such as the inference of reward contingency, making the decision different from a simple perception that the model is normally applied to (for example, Fischer et al., 2018's flanker task). This means that while normally DDMs captures accuracy and RT distribution well, we have identified specific instances where the behaviour distracted from or interfered with the accumulation process. However, one limitation of our findings in the relatively low number of irrational switches, meaning future studies could design tasks to encourage such behaviour to test this further.

4.1.3 Learning through probabilistic inference

Both *Gold Rush* and *metacognition* tasks involve making estimations and inferences under uncertainty. The estimation process has been thought of as a form of probabilistic inference (Gold and Shadlen, 2007) that works out the (unknown) state of the world given the noisy data. We built a Bayesian Learner to

characterise how participants estimated the components of underlying reward structures (scale factor and decay rate) in the foraging task and how people inferred the reward contingency (which patch was relevant to reward) in the metacognition task. Particularly, the model started with a uniform representation of possible values and updated the posterior probabilities based on newly observed evidence following the Bayes' rule (O'Reilly, 2013). The results demonstrated that, although with some delay (for new evidence to build up and modulate beliefs), the Bayesian model well captured how people tracked the reversals of reward components as well as reward contingency.

Across two different task paradigms and decision types, our results provide further evidence for the statistical decision theory, a conceptual framework that describes mathematical prescriptions of forming a decision based on computational elements (reviewed by Gold and Shadlen, 2007). Does that mean our brain is doing probabilistic calculation when making decisions? Some argue that the brain is more like a Bayesian sampler that instead of calculating probabilities, the brain samples information to form representations of the world following a sampling algorithm asymptotic to a Bayesian model, such as Markov Chain Monte Carlo (MCMC) (Sanborn and Chater, 2016). The Bayesian models we built for our tasks followed the laws of probability. Both models had predefined grids of possible states that were evaluated using the evidence, and the only difference was in *Gold Rush* the variables were not bounded and we made our own cut-off in the grid search. The fact that we successfully implemented Bayesian models in both tasks suggests that people could track abstract task features and structural elements such as decay rates, scale factors and reward contingency to making decisions while for leaving times they resorted to much simpler average value based heuristics. However, both tasks used relatively straightforward rules (e.g. simple exponential decay and reward contingency), making the cal-

culation of probabilities feasible. Future research may further examine if the decision processes still follow the ideal Bayesian reasoning or turn to a Bayesian sampler with more complex reward structures.

4.2 Neural Representations of Learning

The EEG results of our studies revealed important neural basis of learning. Next I will present how modulated anticipatory attention, P300 and FRN reflect context updating and choice commitment, as well as how the results from both studies fit into a theoretical framework about feedback learning and action selection.

4.2.1 Context updating and choice commitment

We sought to establish previously found EEG signals related to motivation and learning. We found lateralised VEP and alpha desynchronization related to attended side and accuracy, reflecting enhanced anticipatory attention in *meta-cognition* task. The findings are aligned with previous studies showing the relation between VEP and covert attention (Eimer et al., 2006; Hilo-Merkovich et al., 2018), as well as the connection between alpha-band oscillations and attentional control of the knowledge system (Klimesch, 2012; Klimesch et al., 1998; Klimesch et al., 1992). The fact that participants devoted more attention to the attended side during the working memory delay suggests high-level, motivation related cognitive control. We also found effect of low accuracy expectation on signals of electrodes CPZ and FCZ during 300-400 ms post feedback in *meta-cognition*, in accordance with previous literature suggesting the role of P300 in updating context (Donchin and Coles, 1988; see the EEG discussion section in study 2). We also decoded next trial's leaving time from 300-400 ms post feedback period in *Gold Rush*, indicating feedback modulated overall commitment

to the current environment. Taken together, it seems in *metacognition* P300 was linked to feedback processing but in *Gold Rush* there was a potential vigour signal related to future leaving time. It is possible that P300 encodes a model that the brain holds of the environment, and the amplitude of P300 scales with the need to adjust the model. Although without source reconstruction we cannot conclude the decoded neural signatures in *Gold Rush* were commonly identified P300, this finding is still intriguing, under the novel foraging scenario that has not been addressed by previous studies on P300, and our results suggest there may be common neural circuitry of context updating in both value-based and perceptual decision making.

Does reduced commitment imply a switch? Probably not. Although next trial's leaving was decoded from feedback signals in *Gold Rush*, next trial's switch was not. In *metacognition*, feedback-related negativity did not interact with switch either. Non-lateralised beta desynchronization was found to correlate with switch readiness in our task. It suggests that P300 has potentially multiple roles, but is not simply a switch signal as it was not predictive of trial-wise switching but other motivational and learning factors in both tasks.

4.2.2 Feedback learning and action selection

Learning from the past is essential in goal-directed decision making. Previous research on feedback guided learning argues that feedback modulates action selection through a system of performance monitoring and behavioural adjustments governed by ACC and basal ganglia (BG) (Cockburn and Frank, 2011). Evidence emerged showing error-related negativity (ERN or rERN) and feedback-related negativity (FRN or fERN) as indices of error signals generated by ACC (Martin et al., 2011; Hughes and Yeung, 2010; Steinhauser and Yeung, 2012; for review see Cockburn and Frank, 2011). Both our tasks provided explicit feedback, but information conveyed by feedback was ambiguous (in *Gold Rush* the

accumulated reward was given and needed to be considered together with time spent and overall environment; in *metacognition* a single reward/no-reward was given and needed to be evaluated with both accuracy and reward contingency). Furthermore, unlike in speeded reaction time tasks, it was not possible for our participants to realise an error immediately after a response. Since there was no unequivocal error awareness drawn from feedback or response, we did not aim to identify the cognitive role of ERN or FRN components in our studies. However, our results still fit into the important theoretical framework addressing feedback learning and action selection.

Two influential theories have been proposed to account for the role of ERN/FRN in learning. The *reinforcement learning theory* proposed by Holroyd and Coles, 2002 viewed ERN/FRN as indices of phasic dopamine signals impacting ACC. Specifically, prediction errors are reflected by the arrival of dopamine signals in ACC and unexpected bad outcomes elicit larger FRN than expected bad outcomes. We decoded time discounted accumulated reward as well as reward history from EEG signals around feedback in *Gold Rush* and found expectation modulated reward/no-reward signals post feedback presentation (0.3-0.4 s, centro-parietal) in *metacognition*. Although we cannot conclude the decoded signals were indeed FRN in our first project, the results are still in accordance with the theory suggesting FRN is related to reward contingency learning. Another theory, the *conflict monitoring theory* interprets ERN as a product of conflict monitoring processes in the ACC (Botvinick et al., 2001; Yeung et al., 2004). The authors argued that ERN facilitates a response conflict monitor system coordinated by ACC to select from mutually exclusive responses. The theory predicts fast errors, which were not found in our experiment: we found slow errors in irrational switch trials but no accuracy or RT difference in other trials in *metacognition*. This finding was more likely to result from low

concentration rather than response conflict. Our results seem to support the *reinforcement learning theory* in that ERN/FRN are related to prediction errors and serve to adjust people's beliefs of choice options. Again, since our tasks are different from the speeded reaction time tasks that *conflict monitoring theory* was built on, our results might reflect a different role of ERN/FRN in a distinct task paradigm.

In addition to the two theories, Cockburn and Frank, 2011 proposed a neural-network model that potentially explains biological mechanisms of the association between ACC, basal ganglia and ERN/FRN. The model was built upon a core model describing BG's function as gating information flow between thalamus and cortex and augmented the core model with an ACC layer (**Figure 72**).

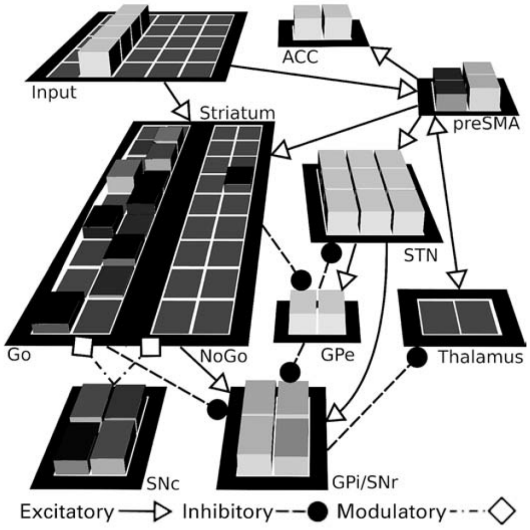


Figure 72: Neural network model of BG circuit from (Cockburn and Frank, 2011). BG consists of five core components: the striatum, globus pallidus (incl. GPi and GPe), substantia nigra pars compacta (SNc), thalamus, and subthalamic nucleus (STN). These core structures either facilitates or suppresses cortical activations through recurrent connectivity between thalamus and cortex (incl. preSMA and sensory cortex). The reciprocal connectivity between BG and ACC is completed via the limbic pathways and the pyramidal cells in ACC which project to multiple motor areas including BG and preSMA. The ACC represents available responses and receives inputs from preSMA. The feedback provided by SNc then facilitates response selection and changes response conflicts in ACC by modulating GO and NO-GO synaptic strengths.

The model is based on previous theories that have been proposed to characterise the physiological pathways between BG and ACC. As a brief summary, the hypothetical reciprocal connectivity between BG and ACC consist of two pathways: (1) the limbic loop which originate from the orbitofrontal cortex and ACC and pass through the BG and thalamus; (2) the pyramidal cells in ACC that project to multiple motor areas including BG, supplementary and primary motor cortices. Different responses are registered in units that comprise the ACC layer. The ERN and FRN are believed to be either the "indices of ACC's response to the arrival of phasic dopamine signals" (the reinforcement learn-

ing theory) or "the product of conflict monitoring processes in the ACC" (the conflict monitoring theory) (see Cockburn and Frank, 2011). According to the BG-ACC model (illustrated by **Figure 72**), action selection is performed by bottom-up support from the BG and top-down dopamine mediated response monitoring from ACC. The bottom-up pathway posits that the striatum excites using GO signal (i.e. signals from the striatal D1 receptor inhibiting GPi to remove tonic inhibition of the thalamus) or suppresses using NO-GO signals (i.e. signals from the striatal D2 receptor inhibiting GPe to activate GPi and prevent flow of thalamo-cortical activity) the selection of particular cortical actions. Dopamine, as a reinforcement learning signal, modulates model plasticity that allows the BG model to learn which cortical representation should be facilitated and which to be suppressed by further increasing GO units and decreasing NO-GO units, as simple Hebbian-learning principles (i.e. the connection strengths between active units increase and vice versa). In doing so, the feedback allows preSMA to learn to refine action selection based on previous actions and outcomes. The model proposes a relationship between response conflict and feedback: response conflicts in the ACC are driven by projections from the BG network (i.e. the relative strengths of the GO and NO-GO signals), which are modulated by feedback. Furthermore, as explained above, FRN may represent negative reinforcement learning signals that induce NO-GO signals in the BG.

Our studies involve two hierarchies of response conflicts: low level action selection based on choice signals guiding button press, and high level cognition of value comparisons, reward history, patch commitment etc. We found effect of difficulty on responses in *metacognition* and decoded lateralised travel time difference from patch choice EEG signals in *Gold Rush*, indicating bottom up response selection driven by basic features of stimuli, as proposed by the model. The relationship between response readiness and switch in *metacognition* as well

as the decoded reward history from feedback and patch choice signals suggest the built-up switch evidence might influence choice commitment and eventually the response by modulating the synaptic strengths as proposed by the top down feedback guided pathway in the model. Note that we did not find a direct link between FRN and switch in *metacognition*, probably because FRN did not directly lead to switch, but gradually contributed to a switch decision by changing the synaptic strengths. Future research may focus on finding neurobiological evidence of how feedback is learned and governs action selection in hierarchical decision making tasks.

4.3 Future Research

Although our findings have revealed important cognitive and neural mechanisms of decision making involving integration of multiple information sources and hierarchical inferences, some questions still remain and require further work. Here I present a few questions that point to future research:

- Where in the brain are dynamic reward changes tracked and decision strategies adjusted in foraging? Does the dissociation of vmPFC and ACC (Kolling et al., 2012) account for patch selection and patch leaving in foraging scenario with dynamic reward structures?

- How do neural circuits integrate perceptual evidence as a function of time in hierarchical perceptual decision making?

- How is probabilistic inference implemented in the brain? Is the mechanism universal or task-specific (e.g. ideal Bayesian reasoning vs. Bayesian sampler as proposed by Sanborn and Chater, 2016)?

- Can we find neurobiological evidence for the BG-ACC model (Cockburn and Frank, 2011) of how ACC tracks switch evidence and governs re-

sponse selection in hierarchical decision making? Does the model explain behaviour in similar task paradigms with more than two alternatives?

4.4 Concluding Remarks

In this thesis I described two projects that tackled hierarchical decision making with dynamic reward structures with a virtual foraging scenario (chapter 2) and a perceptual decision making task (chapter 3). The thesis makes important contributions to the field of decision making in three aspects. Firstly, it extends our understanding of human decision making with complex structures with two different decision making scenarios. It reveals some common mechanisms underlying seemingly very different task paradigms (e.g. value-based and perceptual decisions). Secondly, combining behavioural models and EEG analyses, including ERP, brain oscillations and decoding analyses, it unveils unobservable neural mechanisms of how human brain dynamically learns, integrates and utilises complex task structures and adjusts the decision strategies accordingly. Thirdly, it reveals neural dynamics for patch learning and selection and arbitration between performance lapse and environmental change. In other words, it looks at what environment to pick, when to leave (*Gold Rush*) and when to change what to be engaged with (*metacognition*), which in a way is also a question of when to 'leave' or abandon a side cognitively.

Understanding human decision making in complex, real-life scenarios is not easy. The work presented by this thesis only scratched the surface of complex decision making that we face in our daily life, and more work needs to be done to further examine the theoretical frameworks suggested by our findings. Still, I hope our work serves as a good starting point and sheds light on the understanding of some universal or context-dependent representations behind human behaviour that comprise the fabric of cognition.

References

Adams, G. K., Watson, K. K., Pearson, J. & Platt, M. L. (2012). Neuroethology of Decision-making. *Current Opinion in Neurobiology*, *22*(6), 982–989. https://doi.org/10.1016/j.conb.2012.07.009

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G. & Frith, C. D. (2010). Optimally Interacting Minds. *Science*, *329*(5995), 1081–1085. https://doi.org/10.1126/science.1185718

Bang, D. & Fleming, S. M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(23), 6082–6087. https://doi.org/10.1073/pnas.1800795115

Behrens, T. E., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Bellebaum, C. & Daum, I. (2008). Learning-related changes in reward expectancy are reflected in the feedback-related negativity. *The European journal of neuroscience*, *27*(7), 1823–1835. https://doi.org/10.1111/j.1460-9568.2008.06138.x

Bellebaum, C., Polezzi, D. & Daum, I. (2010). It is less than you expected: the feedback-related negativity reflects violations of reward magnitude expectations. *Neuropsychologia*, *48*(11), 3343–3350. https://doi.org/10.1016/j.neuropsychologia.2010.07.023

Bonnefond, M. & Jensen, O. (2012). Alpha Oscillations Serve to Protect Working Memory Maintenance against Anticipated Distracters. *Current Biology*, *22*(20), 1969–1974. https://doi.org/10.1016/j.cub.2012.08.029

Borst, J. P. & Anderson, J. R. (2015). The discovery of processing stages: Analyzing EEG data with hidden semi-Markov models. *NeuroImage*, *108*, 60–73. https://doi.org/10.1016/j.neuroimage.2014.12.029

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S. & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652. https://doi.org/10.1037/0033-295X.108.3.624

Botvinick, M. M., Cohen, J. D. & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. *Trends in Cognitive Sciences*, *8*(12), 539–546. https://doi.org/10.1016/j.tics.2004.10.003

Brodmann, M. (n.d.). Anterior cingulate cortex. https://commons.wikimedia.org/w/index.php?curid=8402164

Bromberg-Martin, E. S. & Hikosaka, O. (2011). Lateral habenula neurons signal errors in the prediction of reward information. *Nature Neuroscience*, *14*(9), 1209–1216. https://doi.org/10.1038/nn.2902

Charnov, E. L. (1976). Optimal Foraging, the Marginal Value Theorem. *Theoretical Population Biology*, *9*, 129–136. https://doi.org/10.1016/0040-5809(76)90040-X

Chase, H. W., Swainson, R., Durham, L., Benham, L. & Cools, R. (2011). Feedback-related negativity codes prediction error but not behavioral adjustment during probabilistic reversal learning. *Journal of cognitive neuroscience*, *23*(4), 936–946. https://doi.org/10.1162/jocn.2010.21456

Cockburn, J. & Frank, M. (2011). Reinforcement learning, conflict monitoring, and cognitive control: An integrative model of cingulate-striatal interactions and the ern. In R. B. Mars, J. Sallet, M. F. Rushworth & N. Yeung (Eds.), *Neural basis of motivational and cognitive control* (pp. 311–331). MIT Press.

Cohen, M. X., Wilmes, K. A. & van de Vijver, I. (2011). Cortical electrophysiological network dynamics of feedback learning. *Trends in Cognitive Sciences*, *15*(12), 558–566. https://doi.org/10.1016/j.tics.2011.10.004

Cohen, M. X., Wilmes, K. & van de Vijver, I. (2011). Cortical electrophysiological network dynamics of feedback learning. *Trends in Cognitive Sciences*, *15*(12), 558–566. https://doi.org/10.1016/j.tics.2011.10.004

Creel, D. J. (2019). Visually evoked potentials. *Neurology*, *160*, 501–522. https://doi.org/10.1016/B978-0-444-64032-1.00034-5

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. (2006). Cortical Substrates for Exploratory Decisions in Humans. *Nature*, *441*, 876–879. https://doi.org/10.1038/nature04766

Daw, N. D. & Tobler, P. N. (2014). Value learning through reinforcement: The basic of dopamine and reinforcement learning. In P. W. Gilmcher & E. Fehr (Eds.), *Neuroeconomics* (pp. 283–298). Academic Press.

Diamond, D. M., Kramer, A. & Wickens, C. (1986). Applications of event-related brain potentials to problems in engineering psychology. In M. G. Coles, E. Donchin & S. W. Porges (Eds.), *Psychophysiology: Systems, processes and applications* (p. 280). Guilford Press.

Donchin, E. & Coles, M. G. (1988). Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences*, *11*(3), 357–427. https://doi.org/10.1017/S0140525X00058027

Duncan-Johnson, C. (1981). Young Psychophysiologist Award address, 1980. P300 latency: a new metric of information processing. *Psychophysiology*, *18*(3), 207–215. https://doi.org/10.1111/j.1469-8986.1981.tb03020.x

Duncan-Johnson, C. & Donchin, E. (1977). On quantifying surprise: the variation of event-related potentials with subjective probability. *Psycho-*

physiology, *14*(5), 456–467. https://doi.org/10.1111/j.1469-8986.1977.tb01312.x

Eimer, M., Velzen, J. V., Gherri, E. & Press, C. (2006). Manual response preparation and saccade programming are linked to attention shifts: ERP evidence for covert attentional orienting and spatially specific modulations of visual processing. *Brain Research*, *1105*(1), 7–19. https://doi.org/10.1016/j.brainres.2005.10.060

Fiorillo, C., Tobler, P. & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, *299*, 1898–1902. https://doi.org/10.1126/science.1077349

Fischer, A., Nigbur, R., Klein, T. A., Danielmeier, C. & Ullsperger, M. (2018). Cortical beta power reflects decision dynamics and uncovers multiple facets of post-error adaptation. *Nature Communications*, *9*(5038). https://doi.org/10.1038/s41467-018-07456-8

Fleming, S. M., Hujigen, J. & Dolan, R. J. (2012). Prefrontal Contributions to Metacognition in Perceptual Decision Making. *Journal of Neuroscience*, *32*(18), 6117–6125. https://doi.org/10.1523/JNEUROSCI.6489-11.2012

Gehring, W. & Wiloughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science*, *295*(5563), 2279–2282. https://doi.org/10.1126/science.1066893

Glimcher, P. W. (2014). Value-based decision making. In P. W. Gilmcher & E. Fehr (Eds.), *Neuroeconomics* (pp. 373–391). Academic Press.

Glimcher, P. W. & Fehr, E. (2014). *Neuroeconomics*. Academic Press.

Gold, J. I. & Heekeren, H. R. (2014). Neural mechanisms for perceptual decision making. In P. W. Gilmcher & E. Fehr (Eds.), *Neuroeconomics* (pp. 355–372). Academic Press.

Gold, J. I. & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*, 535–574. https://doi.org/10.1146/annurev.neuro.29.051605.113038

Gopher, D. & Donchin, E. (1986). Workload: An examination of the concept. In K. Boff, L. Kaufman & J. Thomas (Eds.), *Handbook of perception and human performance* (pp. 1–49). John Wiley & Sons.

Goyer, J. P., Woldorff, M. G. & Huettel, S. A. (2008). Rapid electrophysiological brain responses are influenced by both valence and magnitude of monetary rewards. *Journal of cognitive neuroscience*, *20*(11), 2058–2069. https://doi.org/10.1162/jocn.2008.20134

Hajcak, G., Moser, J. S., Holroyd, C. B. & Simons, R. F. (2006). The feedback-related negativity reflects the binary evaluation of good versus bad outcomes. *Biological Psychology*, *71*(2), 148–154. https://doi.org/10.1016/j.biopsycho.2005.04.001

Händel, B. F., Haarmeier, T. & Jensen, O. (2011). Alpha oscillations correlate with the successful inhibition of unattended stimuli. *Journal of Cognitive Neuroscience*, *23*(9), 2494–2502. https://doi.org/10.1162/jocn.2010.21557

Hanks, T. D., Mazurek, M. E., Kiani, R., Hopp, E. & Shadlen, M. N. (2011). Elapsed Decision Time Affects the Weighting of Prior Probability in a Perceptual Decision Task. *Journal of Neuroscience*, *31*(17), 6339–6352. https://doi.org/10.1523/JNEUROSCI.5613-10.2011

Hayden, B. Y., Pearson, J. M. & Platt, M. L. (2011). Neuronal basis of sequential foraging decisions in a patchy environment. *Nature Neuroscience*, *14*(7), 933–939. https://doi.org/10.1038/nn.2856

Hebart, M. N., Schriever, Y., Donner, T. H., Haynes, J.-D. & Notes, A. (2016). The Relationship between Perceptual Decision Variables and Confid-

ence in the Human Brain. *Cerebral Cortex*, *26*(1), 118–130. https://doi.org/10.1093/cercor/bhu181

Heekeren, H. R., Marrett, S., Bandettini, P. A. & Ungerleider, L. G. (2004). A general mechanism for perceptual decision-making in the human brain. *Nature*, *431*, 859–962. https://doi.org/10.1038/nature02966

Hillyard, S. A., Vogel, E. K. & Luck, S. J. (1998). Sensory gain control (amplification) as a mechanism of selective attention: electrophysiological and neuroimaging evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *353*, 1257–1270. https://doi.org/10.1098/rstb.1998.0281

Hilo-Merkovich, R., Carrasco, M. & Yuval-Greenberg, S. (2018). Task performance in covert, but not overt, attention correlates with early laterality of visual evoked potentials. *Neuropsychologia*, *119*, 330–339. https://doi.org/10.1016/j.neuropsychologia.2018.08.012

Holroyd, C. B. & Coles, M. G. H. (2002). The Neural Basis of Human Error Processing: Reinforcement Learning, Dopamine, and the Error-Related Negativity. *Psychological Review*, *109*(4), 679–709. https://doi.org/10.1037//0033-295X.109.4.679

Holroyd, C. B., Hajcak, G. & Larsen, J. T. (2006). The good, the bad and the neutral: electrophysiological responses to feedback stimuli. *Brain research*, *1105*(1), 93–101. https://doi.org/10.1016/j.brainres.2005.12.015

Holroyd, C. B. & Krigolson, O. E. (2007). Reward prediction error signals associated with a modified time estimation task. *Psychophysiology*, *44*(6), 913–917. https://doi.org/10.1111/j.1469-8986.2007.00561.x

Horst, R. L., Johnson, R. & Donchin, E. (1980). Event-related brain potentials and subjective probability in a learning task. *Memory & Cognition*, *8*, 476–488. https://doi.org/10.3758/BF03211144

Hubbard, J., Kikumoto, A. & Mayr, U. (2019). EEG decoding reveals the strength and temporal dynamics of goal-relevant representations. *Scientific Reports*, *9*(9051). https://doi.org/10.1038/s41598-019-45333-6

Hughes, G. & Yeung, N. (2010). Dissociable correlates of response conflict and error awareness in error-related brain activity. *Neuropsychologia*, *49*(3), 405–415. https://doi.org/10.1016/j.neuropsychologia.2010.11.036

Ichikawa, N., Siegle, G. J., Dombrovski, A. & Ohira, H. (2010). Subjective and model-estimated reward prediction: Association with the feedback-related negativity (FRN) and reward prediction error in a reinforcement learning task. *International Journal of Psychophysiology*, *78*(3), 273–283. https://doi.org/10.1016/j.ijpsycho.2010.09.001

Johnson, R. & Donchin, E. (1982). Sequential expectancies and decision making in a changing environment: an electrophysiological approach. *Psychophysiology*, *19*(2), 183–200. https://doi.org/10.1111/j.1469-8986.1982.tb02545.x

Kappenman, E. S. & Luck, S. J. (2011). *The oxford handbook of event-related potential components*. Oxford University Press.

Kelly, S. P., Lalor, E. C., Reilly, R. B. & Foxe, J. J. (2006). Increases in alpha oscillatory power reflect an active retinotopic mechanism for distracter suppression during sustained visuospatial attention. *Journal of Neurophysiology*, *95*(6), 3844–3851. https://doi.org/10.1152/jn.01234.2005

Kepecs, A., Uchida, N., Zariwala, H. A. & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227–231. https://doi.org/10.1038/nature07200

Kiani, R. & Shadlen, M. N. (2009). Representation of Confidence Associated with a Decision by Neurons in the Parietal Cortex. *Science*, *324*(5928), 759–764. https://doi.org/10.1126/science.1169405

Klimesch, W. (1999). EEG alpha and theta oscillatins reflect cognitive and memory performance: a review and analysis. *Brain research. Brain research reviews*, *29*(2-3), 169–195. https://doi.org/10.1016/s0165-0173(98)00056-3

Klimesch, W. (2012). Alpha-band oscillations, attention, and controlled access to stored information. *Trends in Cognitive Sciences*, *16*(12), 606–617. https://doi.org/10.1016/j.tics.2012.10.007

Klimesch, W., Doppelmayr, M., Russegger, H., Pachinger, T. & Schwaiger, J. (1998). Induced alpha band power changes in the human EEG and attention. *Neuroscience Letters*, *244*(2), 73–76. https://doi.org/10.1016/s0304-3940(98)00122-0

Klimesch, W., Pfurtscheller, G. & Schimke, H. (1992). Pre- and post-stimulus processes in category judgement tasks as measured by event-related desynchronization (ERD). *Journal of Psychophysiology*, *6*(3), 185–203.

Kolling, N., Behrens, T. E., Mars, R. B. & Rushworth, M. F. (2012). Neural Mechanisms of Foraging. *Science*, *336*(6077), 95–98. https://doi.org/10.1126/science.1216930

Kolling, N. & O'Reilly, J. X. (2018). State-change decisions and dorsomedial prefrontal cortex: the importance of time. *Current Opinion in Behavioral Sciences*, *22*, 152–160. https://doi.org/10.1016/j.cobeha.2018.06.017

Kolling, N., Wittmann, M. & Rushworth, M. F. (2014). Multiple neural mechanisms of decision making and their competition under changing rish pressure. *Neuron*, *81*(5), 1190–1202. https://doi.org/10.1016/j.neuron.2014.01.033

Korn, C. W. & Bach, D. R. (2003). Heuristic and optimal policy computations in the human brain during sequential decision-making. *Nature Communications*, *9*(325). https://doi.org/10.1038/s41467-017-02750-3

Lebreton, M., Abibol, R., Daunizeau, J. & Pessiglione, M. (2015). Automatic integration of confidence in the brain valuation signal. *Nature Neuroscience*, *18*, 1159–1167. https://doi.org/10.1038/nn.4064

Lynch, P. J. (n.d.). Ventromedial prefrontal cortex. https://creativecommons.org/licenses/by/3.0

Martin, M. E., Yeung, N. & Steinhauser, M. (2011). Error related brain activity and adjustments of selective attention following errors. *Neuroimage*, *56*(4), 2339–2347. https://doi.org/10.1016/j.neuroimage.2011.03.083

Martino, B. D., Fleming, S. M., Garrett, N. & Dolan, R. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105–110. https://doi.org/10.1038/nn.3279

McFarland, D. J., Miner, L. A., Vaughan, T. M. & Wolpaw, J. R. (2000). Mu and beta rhythm topographies during motor imagery and actual movements. *Brain topography*, *12*(3), 177–186. https://doi.org/10.1023/a:1023437823106

Meder, D., Kolling, N., Verhagen, L., Wittmann, M. K., Scholl, J., Madsen, K. H., Hulme, O. J., Behrens, T. E. & Rushworth, M. F. (2017). Simultaneous representation of a spectrum of dynamically changing value estimates during decision making. *Nature Communications*, *8*(1), 1942. https://doi.org/10.1038/s41467-017-02169-w

Myers, N., Rohenkohl, G., Wyart, V., Woolrich, M. W., Nobre, A. C. & Stokes, M. G. (2015). Testing sensory evidence against mnemonic templates. *eLife*, *4*, e09000. https://doi.org/10.7554/eLife.09000

Nieuwenhuis, S., Aston-Jones, G. & Cohen, J. D. (2005). Decision Making, the P3, and the Locus Coeruleus-Norepinephrine System. *Psychological Bulletin*, *131*(4), 510–532. https://doi.org/10.1037/0033-2909.131.4.510

Nieuwenhuis, S., Holroyd, C. B., Mol, N. & Coles, M. G. (2004). Reinforcement-related brain potentials from medial frontal cortex: origins and functional significance. *Neuroscience & Biobehavioral Reviews*, *28*(4), 441–448. https://doi.org/10.1016/j.neubiorev.2004.05.003

Nitta, E., Onoda, K., Ishitobi, F., Okazaki, R., Nagai, S. M. A. & Yamaguchi, S. (2017). Enhanced Feedback-related negativity in Alheimer's Disease. *Frontiers in Human Neuroscience*, *11*, 179. https://doi.org/10.3389/fnhum.2017.00179

O'Connell, R. G., Shadlen, M. N., Wong-Lin, K. & Kelly, S. P. (2018). Bridging Neural and Computational Viewpoints on Perceptual Decision-Making. *Trends in Neurosciences*, *41*(11), 838–852. https://doi.org/10.1016/j.tins.2018.06.005

O'Reilly, J. X. (2013). Making predictions in a changing world—inference, uncertainty, and learning. *Frontiers in Neuroscience*, *7*(105). https://doi.org/10.3389/fnins.2013.00105

O'Reilly, J. X., Jbabdi, S., Rushworth, M. F. & Behrens, T. E. (2013). Brain Systems for Probabilistic and Dynamic Prediction: Computational Specificity and Integration. *PLOS Biology*, *11*(9), e1001662. https://doi.org/10.1371/journal.pbio.1001662

Pfabigan, D. M., Alexopoulos, J., Bauer, H. & Sailer, U. (2011). Manipulation of feedback expectancy and valence induces negative and positive reward prediction error signals manifest in event-related brain potentials.

Psychophysiology, *48*(5), 656–664. https://doi.org/10.1111/j.1469-8986.2010.01136.x

Pfurtscheller, G., Graimann, B., Huggins, J. E., Levine, S. P. & Schuh, L. A. (2003). Spatiotemporal patterns of beta desynchronization and gamma synchro- nization in corticographic data during self-paced movement. *Clinical neurophysiology*, *114*(7), 1226–1236. https://doi.org/10.1016/s1388-2457(03)00067-1

Philiastides, M. G., Biele, G., Vavatzanidis, N., Kazzer, P. & Heekeren, H. R. (2010). Temporal dynamics of prediction error processing during reward-based decision making. *Neuroimage*, *53*(1), 221–232. https://doi.org/10.1016/j.neuroimage.2010.05.052

Picazio, S., Veniero, D., Ponzo, V., Caltagirone, C., Gross, J., Thut, G. & Koch, G. (2014). Prefrontal control over motor cortex cycles at beta frequency during movement inhibition. *Current biology*, *24*(24), 2940–2945. https://doi.org/10.1016/j.cub.2014.10.043

Picton, T. (1992). The P300 wave of the human event-related potential. *Journal of clinical neurophysiology : official publication of the American Electroencephalographic Society*, *9*(4), 456–479. https://doi.org/10.1097/00004691-199210000-00002

Platt, M. L. & Plassmann, H. (2014). Multistage valuation signals and common neural currencies. In P. W. Gilmcher & E. Fehr (Eds.), *Neuroeconomics* (pp. 237–258). Academic Press.

Purcell, B. A. & Kiani, R. (2016). Hierarchical decision processes that operate over distinct timescales underlie choice and changes in strategy. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(31), E4531–E4540. https://doi.org/10.1073/pnas.1524685113

Quinn, A. J., Vidaurre, D., Abeysuriya, R., Becker, R., Nobre, A. C. & Woolrich, M. W. (2018). Task-Evoked Dynamic Network Analysis Through Hidden Markov Modeling. *Frontiers in Neuroscience*, *12*, 603. https://doi.org/10.3389/fnins.2018.00603

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163. https://doi.org/10.2307/271063

Ratcliff, R. & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Rushworth, M. F., Kolling, N., Sallet, J. & Mars, R. B. (2012). Valuation and decision-making in frontal cortex: one or many serial or parallel systems? *Current Opinion in Neurobiology*, *22*(6), 946–955. https://doi.org/10.1016/j.conb.2012.04.011

Sanborn, A. N. & Chater, N. (2016). Bayesian Brains without Probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893. https://doi.org/10.1016/j.tics.2016.10.003

Sato, A., Yasuda, A., Ohira, H., Miyawaki, K., Nishikawa, M., Kumano, H. & Kuboki, T. (2005). Effects of value and reward magnitude on feedback negativity and P300. *Neuroreport*, *16*(4), 407–411. https://doi.org/10.1097/00001756-200503150-00020

Shenhav, A., Straccia, M. A., Botvinick, M. M. & Cohen, J. D. (2016). Dorsal anterior cingulate and ventromedial prefrontal cortex have inverse roles in both foraging and economic choice. *Cognitive, Affective, & Behavioural Neuroscience*, *16*, 1127–1139. https://doi.org/10.3758/s13415-016-0458-8

Steinhauser, M. & Yeung, N. (2012). Error awareness as evidence accumulation: effects of speed-accuracy trade-off on error signalling. *Frontiers in Human Neuroscience*, *6*, 240. https://doi.org/10.3389/fnhum.2012.00240

Stephens, D. W. & Krebs, J. R. (1987). *Foraging theory*. Princeton University Press.

Stevens, F. L., Hurley, R. A. & Taber, K. H. (2011). Anterior cingulate cortex: unique role in cognition and emotion. *The Journal of Neuropsychiatry and Clinical Neurosciences*, *23*(2), 121–125. https://doi.org/10.1176/jnp.23.2.jnp121

Störmer, V. S., McDonald, J. J. & Hillyard, S. A. (2009). Cross-modal cueing of attention alters appearance and early cortical processing of visual stimuli. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(52), 22456–22461. https://doi.org/10.1073/pnas.0907573106

Sutton, R. S. & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT Press.

Sutton, S., Braren, M., Zubin, J. & John, E. R. (1965). Evoked-potential correlates of stimulus uncertainty. *Science*, *150*(3700), 1187–1188. https://doi.org/10.1126/science.150.3700.1187

Swann, N., Cai, W., Conner, C. R., Pieters, T. A., Claffey, M. P., George, J. S., Aron, A. R. & Tandon, N. (2012). Roles for the pre-supplementary motor area and the right inferior frontal gyrus in stopping action: electrophysiological responses and functional and structural connectivity. *Neuroimage*, *59*(3), 2860–2870. https://doi.org/10.1016/j.neuroimage.2011.09.049

Swann, N., Poizner, H., Houser, M., Gould, S., Greenhouse, I., Cai, W., Strunk, J., George, J. & Aron, A. R. (2011). Deep brain stimulation of the

subthalamic nucleus alters the cortical profile of response inhibition in the beta fre- quency band: a scalp EEG study in Parkinson's disease. *Journal of Neuroscience*, *31*(15), 5721–5729. https://doi.org/10.1523/JNEUROSCI.6135-10.2011

Swann, N., Tandon, N., Canolty, R., Ellmore, T. M., McEvoy, L. K., Dreyer, S., Disano, M. & Aron, A. R. (2009). Intracranial EEG reveals a time- and frequency- specific role for the right inferior frontal gyrus and primary motor cortex in stopping initiated responses. *Journal of Neuroscience*, *29*(40), 12675–12685. https://doi.org/10.1523/JNEUROSCI.3359-09.2009

Talmi, D., Fuentemilla, L., Litvak, V., Duzel, E. & Dolan, R. J. (2012). An MEG signature corresponding to an axiomatic model of reward prediction error. *NeuroImage*, *59*(1), 635–645. https://doi.org/10.1016/j.neuroimage.2011.06.051

Tervo, D. G. R., Kuleshova, E., Manakov, M., Prokurin, M., Karlsson, M., Lustig, A., Behnam, R. & Karpova, A. Y. (2020). Anterior cingulate cortex directs exploration of alternative strategies. *bioRxiv*. https://doi.org/10.1101/2020.05.23.098822

Thut, G., Nietzel, A., Brandt, S. A. & Pascual-Leone, A. (2012). α-Band Electroencephalographic Activity over Occipital Cortex Indexes Visuospatial Attention Bias and Predicts Visual Target Detection. *Journal of Neuroscience*, *26*(37), 9494–9502. https://doi.org/10.1523/JNEUROSCI.0875-06.2006

Toyomaki, A. & Murohashi, H. (2005). Discrepancy between feedback negativity and subjective evaluation in gambling. *Neuroreport*, *16*(16), 1865–1868. https://doi.org/10.1097/01.wnr.0000185962.96217.36

Twomey, D. M., Murphy, P. R., Kelly, S. P. & O'Connell, R. G. (2015). The classic P300 encodes a build-to-threshold decision variable. *European Journal of Neuroscience*, *42*, 1636–1643. https://doi.org/10.1111/ejn.12936

Ullsperger, M., Danielmeier, C. & Jocham, G. (2014). Neurophysiology of performance monitoring and adaptive behavior. *Physiological Review*, *94*(1), 35–79. https://doi.org/10.1152/physrev.00041.2012

Vandekerckhove, J. & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review*, *14*(6), 1011–1026. https://doi.org/10.3758/BF03193087

Vandekerckhove, J. & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods*, *40*(1), 61–72. https://doi.org/10.3758/BRM.40.1.61

van Vugt, M. K., Simen, P., Nystrom, L., Holmes, P. & Cohen, J. D. (2014). Lateralized Readiness Potentials Reveal Properties of a Neural Mechanism for Implementing a Decision Threshold. *PLoS ONE*, *9*(3), e90943. https://doi.org/10.1371/journal.pone.0090943

Vidaurre, D., Myers, N. E., Stokes, M., Nobre, A. C. & Woolrich, M. W. (2019). Temporally Unconstrained Decoding Reveals Consistent but Time-Varying Stages of Stimulus Processing. *Cerebral Cortex*, *29*(2), 863–874. https://doi.org/10.1093/cercor/bhy290

Vogt, B. A. (2005). Pain and emotion interactions in subregions of the cingulate gyrus. *Nature Reviews Neuroscience*, *6*, 533–544. https://doi.org/10.1038/nrn1704

Wagner, A. R. & Rescorla, R. A. (1972). Inhibition in pavlovian conditioning: Application of a theory. In R. Boakes & M. Halliday (Eds.), *Inhibition and learning* (pp. 301–336). Academic Press.

Wagner, J., Wessel, J. R., Ghahremani, A. & Aron, A. R. (2018). Establishing a right frontal beta signature for stopping action in scalp EEG: implications for testing inhibitory control in other task contexts. *Journal of cognitive neuroscience*, *30*(1), 107–118. https://doi.org/10.1162/jocn_a_01183

Weichwald, S., Meyer, T., Özdenizci, O., Schölkopf, B., Ball, T. & Grosse-Wentrup, M. (2015). Causal interpretation rules for encoding and decoding models in neuroimaging. *NeuroImage*, *110*, 48–59. https://doi.org/10.1016/j.neuroimage.2015.01.036

Wessel, J. (2020). β-Bursts Reveal the Trial-to-Trial Dynamics of Movement Initiation and Cancellation. *Journal of Neuroscience*, *40*(2), 411–423. https://doi.org/10.1523/JNEUROSCI.1887-19.2019

Wessel, J. R. (2012). Error awareness and the error-related negativity: evaluating the first decade of evidence. *Frontiers in Human Neuroscience*, *6*, 88. https://doi.org/10.3389/fnhum.2012.00088

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A. & Cohen, J. D. (2014). Humans Use Directed and Random Exploration to Solve the Explore–Exploit Dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074–2081. https://doi.org/10.1037/a0038199

Wittmann, M. K., Kolling, N., Akaishi, R., Chau, B. K., Brown, J. W., Nelissen, N. & Rushworth, M. F. (2016). Predictive decision making driven by multiple time-linked reward representations in the anterior cingulate cortex. *Nature Communications*, *7*, 12327. https://doi.org/10.1038/ncomms12327

Worden, M. S., Foxe, J. J., Wang, N. & Simpson, G. V. (2000). Anticipatory Biasing of Visuospatial Attention Indexed by Retinotopically Specific α-Bank Electroencephalography Increases over Occipital Cortex. *Journal*

of Neuroscience, *20*(6), RC63. https://doi.org/10.1523/JNEUROSCI.20-06-j0002.2000

Yang, T. & Shadlen, M. N. (2007). Probabilistic reasoning by neurons. *Nature*, *447*, 1075–1080. https://doi.org/10.1038/nature05852

Yasuda, A., Sato, A., Miyawaki, K., Kumano, H. & Kuboki, T. (2004). Error-related negativity reflects detection of negative reward prediction error. *Neuroreport*, *15*(16), 2561–2565. https://doi.org/10.1097/00001756-200411150-00027

Yeung, N., Botvinick, M. M. & Cohen, J. D. (2004). The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychological Review*, *111*(4), 931–959. https://doi.org/10.1037/0033-295x.111.4.939

Yeung, N. & Sanfey, A. G. (2004). Independent coding of reward magnitude and valence in the human brain. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *24*(28), 6258–6264. https://doi.org/10.1523/JNEUROSCI.4537-03.2004

Yoon, B.-J. (2009). Hidden Markov Models and their Applications in Biological Sequence Analysis. *Current Genomics*, *10*(6), 402–415. https://doi.org/10.2174/138920209789177575