REINFORCEMENT LEARNING DESCRIBES THE COMPUTATIONAL AND NEURAL PROCESSES UNDERLYING FLEXIBLE LEARNING OF VALUES AND ATTENTIONAL SELECTION

MATTHEW BALCARRAS

A DISSERTATION SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN BIOLOGY
YORK UNIVERSITY
TORONTO, CANADA

JUNE 2015

Abstract

Attention and learning are cognitive control processes that are closely related. This thesis investigates this inter-relatedness by using computational models to describe the mechanisms that are shared between these processes. Computational models describe the transformation of stimuli to observable variables (behaviour) and contain the latent mechanisms that affect this transformation. Here, I captured these mechanisms with the reinforcement learning (RL) framework applied in two different task contexts and three different projects to show 1) how attentional selection of stimuli involves the learning of values for stimuli, 2) how the learning of stimulus values is influenced by previously learned rules, and 3) how explorations of value-related mechanisms in the brain benefit from using intracranial EEG to investigate the strength of oscillatory activity in ventromedial prefrontal cortex.

In the first project, the RL framework is applied to a feature-based attention task that required macaques to learn the value of stimulus features while ignoring non-relevant information. By comparing different RL schemes I found that trial-by-trial covert attentional selections were best predicted by a model that only represents expected values for the task relevant feature dimension.

In the second project, I explore mechanisms of stimulus-feature value learning in humans in order to understand the influence of learned rules for the flexible, on-going learning of expected values. I test the hypothesis that naive subjects will show enhanced learning of feature specific reward associations by switching to the use of an

abstract rule that associates stimuli by feature type. I found that two-thirds of subjects (n=22/32) exhibited behaviour that was best fit by a ‘flexible-rule-selection’ model.

Low-frequency oscillatory activity in frontal cortex has been associated with cognitive control and integrative brain functions, however, the relationship between expected values for stimuli and band-limited, rhythmic neural activity in the human brain is largely unknown. In the third project, I used intracranial electrocorticography (ECoG) in a proof-of-principle study to reveal spectral power signatures in vmPFC related to the expected values of stimuli predicted by a RL model for a single human subject.

DEDICATION

This thesis is dedicated to my family, and especially my partner Deanna. I could not have done this without all of you. We are a team. And I promise: no more degrees.

I also dedicate this to my parents, Alan and Dianne Balcarras. I have been blessed with the opportunity to study, and I learned by your example what hard work looks like.

ACKNOWLEDGMENTS

I would like to acknowledge the support and teaching of my supervisor, Dr. Thilo Womelsdorf. You are an excellent scientist, and I am glad to have had the opportunity to contribute to the work you are doing.

I also acknowledge the support and input from my supervisory committee, Drs. James Elder and J. Douglas Crawford, who have enabled me to bring this dissertation to a solid conclusion on a difficult and tight timeline. I also thank my examination committee for their input and support.

I am grateful to the incredible, and amazingly unique, access given to me by Dr. Taufik Valiante. Working with his patients in the epilepsy unit at Toronto Western Hospital was a powerful experience for which I will always be thankful. Dr. Valiante is a gifted surgeon and clinician, but still found time and space to make my project possible. I am also grateful to Dr. Valiante's clinical staff, especially his nurses Darcia Paul and Alina Shcharinsky who are so flexible in accommodating the iEEG research group at Toronto Western on top of being efficient and caring nurses.

To my colleagues in the attentional control lab, especially Salva, Steffi, Mariann, Chen, and Ben: It has been difficult, and I wish we had more opportunities to relax and de-stress together, but I have appreciated all of you, your input along the way, and your uncompromising commitment to doing the best science possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

"One of the most remarkable aspects of an animal's behaviour is the ability to modify that behaviour by learning, an ability that reaches its highest form in human beings."

Erik Kandel - Nobel Lecture

1.1 - Learning to control attention using values for stimuli

Learning is a hallmark of intelligent systems and is a central characteristic of the human brain - our brains are continually being changed through experience (Rosenzweig et al., 1972). Learning is also a complex phenomena that is still not very well understood and the last century has seen an abundance of work producing new insights into the computational and neuronal processes linking adaptive learning behaviour and brain function. Attentional selection, the process of selecting information for prioritized processing, is an adaptive behaviour that is improved as we learn information about our environments that relate to our goals (Peck et al., 2009; Chelazzi et al., 2013). Learning and attention are tightly linked, but it is unclear how the mechanisms of both processes function in the brain and how they interact (Rombouts et al., 2015). One way to gain new insight into this problem, and what we propose here, is to use the formal account of learning outlined in the Reinforcement Learning (RL) framework (Sutton and Barto,1998) to link changes in attentional behaviour to the circuits underlying associative learning in pre-frontal cortex (Frank & Badre, 2015; Dayan et al., 2000).

1.1a - Selective attention

Attention is the brain's solution to the problem of living in an informationally dense world with a limited capacity for processing information (Tsotsos, 2011). Only some environmental information, thoughts, and sensory inputs can be processed at any point in time and attention is the mechanism by which processing priority is accomplished. Representing and analyzing sensory information is a difficult problem for the brain to solve because not all sensory input is behaviourally relevant, and processing irrelevant information can be biologically very costly, i.e. you might not survive if you attend to the wrong thing. Therefore, because representation of stimulus information is competitive, i.e. some information is represented at the expense of representing other information, the brain requires a mechanism for selectively directing information gathering and processing systems to the most informative aspects in our environment (Desimone and Duncan, 1995; Dayan et al., 2000; Reynolds and Chelazzi, 2004; Tsotsos, 2011). This process has been labelled selective attention.

1.1b - Types of attentional control

There is evidence that selective attention can be directed by two types of mechanisms, voluntary (top-down) and involuntary (bottom-up)(Anderson et al., 2013). This top-down/bottom-up distinction distinguishes between goal-directed (top-down) mechanisms and salience-directed (bottom-up) mechanisms (**Fig. 1**). Salience-directed attentional selection prioritizes sensory information independently of current behavioural goals. Objects in our environment that move fast, or make sudden loud noises, are highly salient, which means that they are often important to our long term interests even if they are not related to our current activity.

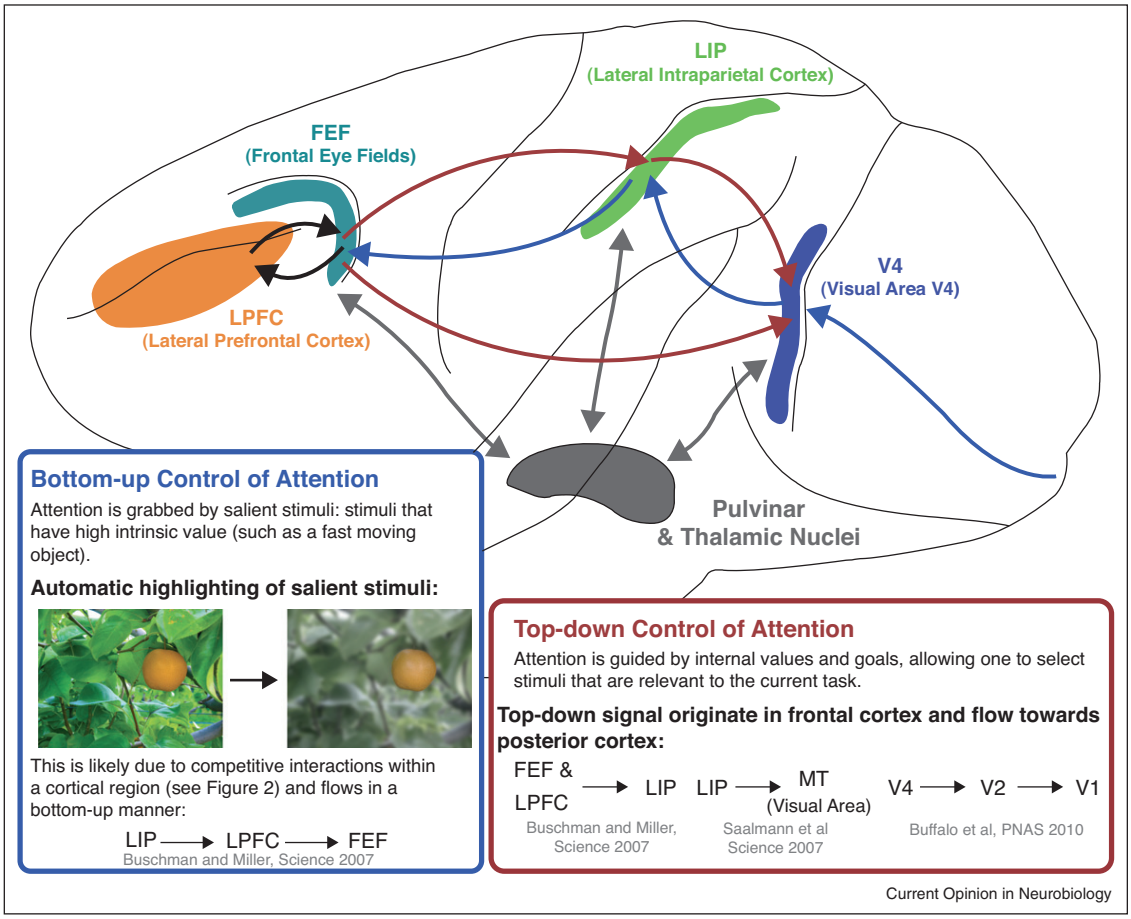


Figure 1. Bottom-up versus top down control of attention in the brain. Separate circuits in the brain have been identified as playing unique roles in different sources of attentional control. Bottom-up attention related to salience-based control moves information from primary visual areas towards pre-frontal cortex and brainstem areas. Top-down attention related to goal directed behaviour selectively routes information from pre-frontal areas 'backwards' toward integration areas (LIP) and primary visual areas to flexibly direct the uptake of new information. Adapted from Miller et. al., 2013.

Even if it is not part of our current goal of collecting fruit, rapidly processing information about a newly-arrived predator is always adaptive and highly salient. Goal-directed mechanisms of attentional selection, on the other hand, represent a powerful tool for directing the flow of information in the brain according to current interests. Accuracy and reaction time improve when responses are driven by attending to a relevant target (Chelazzi et al., 2013). Studies have shown that being cued to a future target location

informs goal-directed mechanisms to give priority to processing stimuli appearing at that location, and improves task performance (Anderson, 2013). However, the relevance of stimuli for goals can change over time or be uncertain - it is not always clear what is most relevant for accomplishing a current goal. Spoiled food may still look edible or a previously desired food item might become less appetizing as we reach satiety, which would make these things less important to focus on. Therefore, we need to be able to flexibly control our attention as we learn new things about our environment, such as value of stimuli for receiving reward. Stimulus values, or stimulus feature values, are typically defined as the predicted, or expected reward associated with that stimulus given the history of experiences with that stimulus (Niv, 2009; Wunderlich et al., 2010). Due to the dynamics of value, and the dynamic nature of human environments in general, goal-directed selective attention requires mechanisms that track value across time and provide current estimates of 'stimulus value' (Rombouts et al., 2015).

1.1c - Learned values and attentional selection

There is recent evidence that value-based learning mechanisms in the brain play a central role in selective attention (Peck et al., 2009; Anderson et al., 2011; Kaping et al., 2011; Gottlieb, 2012; Anderson and Yantis, 2013; Rombouts et al., 2015). Traditionally attention has been studied in cued attention tasks, where subjects learn during training associations between cue signals and sensory features by receiving rewards for correct responses to stimuli. Thus after training, attentional selection is triggered by the cue (Kaping et al., 2011). However, in the absence of explicit instructions, which is a normal experience in everyday environments, the control of attentional selection needs to rely on internal mechanisms that dynamically track the relevance of sensory information in

the environment (Droll et al., 2009; Anderson et al., 2011; Gottlieb, 2012; Anderson et al., 2013). Recent work by Anderson et. al (Anderson et al., 2011) has shown how learned values for stimulus features influences subsequent attentional selection. Their study suggests that one of the internal mechanisms for controlling attentional selection is likely to be found in the brain system underlying value-based learning.

In the study by Anderson et. al. (Anderson et al., 2011) subjects were trained to associate stimulus colour with rewarding outcomes (**Fig. 2**). Following this training

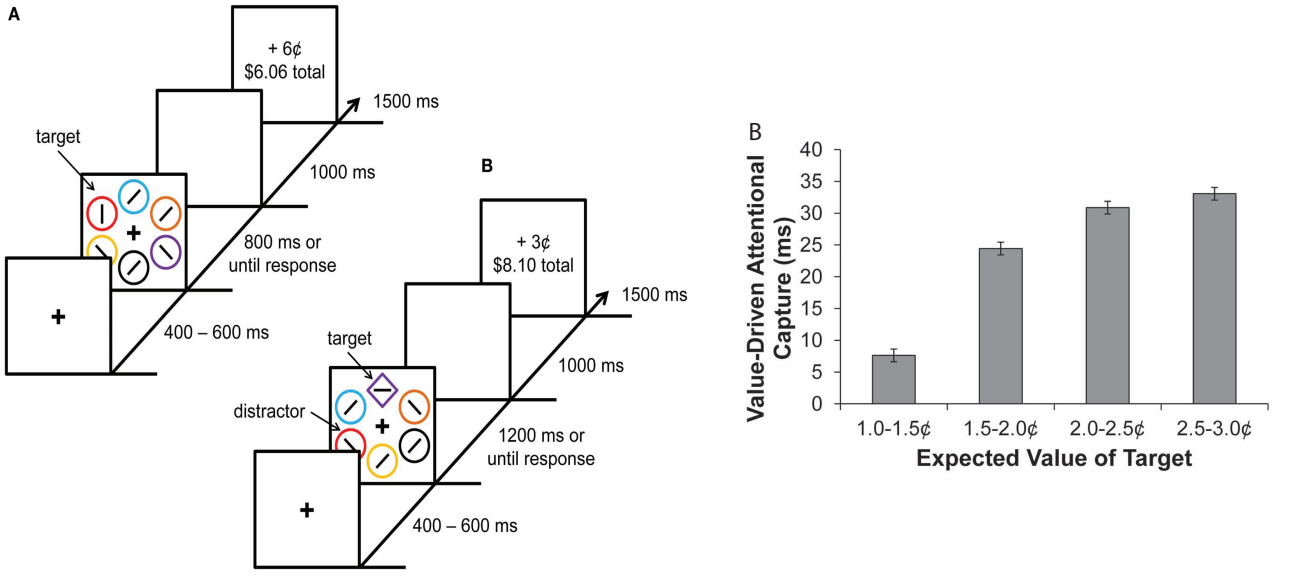


Figure 2. Attentional selection is influenced by reward learning. A) Subjects in this study first learn to associate stimulus colour with reward value. Following this learning, subjects learn to respond to a non-colour feature dimension. B) Reaction time on non-colour stimulus feature responses is proportionally influenced by the previously learned value for stimulus colour. Adapted from Anderson et.

session, subjects were then required to make choices in response to the shape of stimuli. Anderson et. al found that response time to the non-colour feature was proportional to the previously learned value of the stimulus colour. For example, if red had previously been associated with and expected value of 2.5-3¢ (the highest value for

a colour), then in the testing phase if the a non-target stimulus was coloured red, the response time to the target shape is slower than for all trials with other distractor colours. This result shows that value learning systems directly influence top-down sources of attentional control.

Similar to the results found by Anderson et. al., recent work by Kaping et. al. (Kaping et al., 2011) has shown how stimulus reward values influence covert shifts of attention, where information gathering systems are shifted independently of the overt

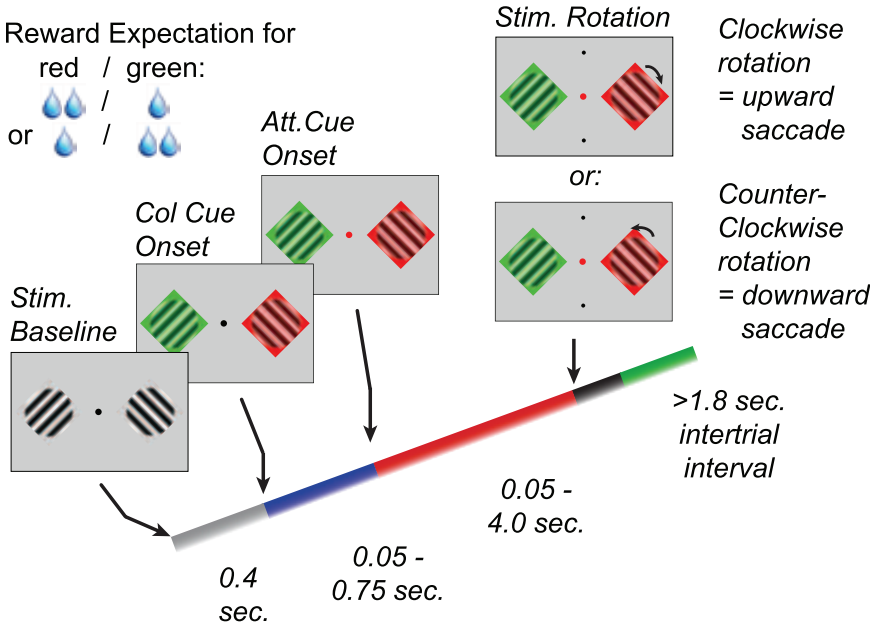


Figure 3. Cued attentional selection task. Monkeys were required to covertly shift attention towards the cued stimulus (target defined by cue colour), while maintaining fixation and ignoring the distractor effect of differing reward magnitudes for different colours in order to discriminate the transient rotation of the target and identify the proper saccade direction. Adapted from Kaping et al., 201).

gaze response. In the study by Kaping et. al. monkeys were trained to associate a cue

colour with a target stimulus of the same colour at random locations and irrespective of the reward that was varied for different stimulus colours in a block fashion (**Fig. 3**).

In discrete trials organized by blocks, monkeys were cued to covertly attend (they had to maintain fixation to the cue) to a stimulus by colour in order to discriminate the

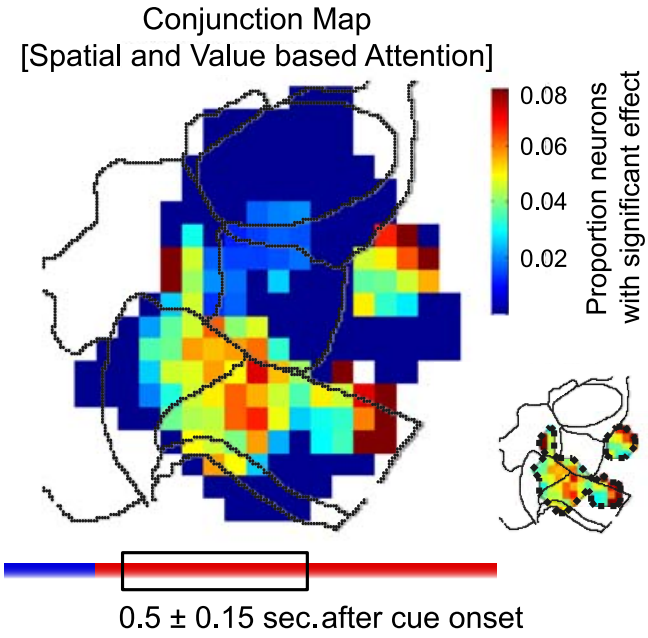


Figure 4. Flat map of macaque prefrontal cortex shows the spatial distribution of neurons that show significant effect for both spatial location and stimulus value. Conjunctive selectivity is restricted to two small clusters of neurons. Adapted from Kaping et al., 2011.

transient rotation indicating the correct saccadic response direction. Kaping et. al. found that the reward magnitude of the target versus the distractor directly influenced the success rate of the monkey in covertly shifting attention to the target. In addition to the behavioural effect of stimulus feature value on attentional selection, a key finding of the Kaping et. al study was that the neuronal response in pre-frontal cortex to the attentional cue onset was modulated by the relative reward value of the target colour as well as by the location of the target stimulus (see **Fig. 4**).

Pre-frontal areas in the brain such as ventromedial pre-frontal cortex (vmPFC), lateral PFC (lPFC), and Anterior Cingulate Cortex (ACC), are known to be associated with brain networks underlying value-based learning (Wallis, 2007; Rushworth et al., 2011; 2012). The finding of Kaping et. al. suggests that further work exploring attentional selection when values for stimulus features are dynamic or uncertain is needed to clarify the interaction between mechanisms in the brain involved in top-down attentional control and mechanisms involved in flexibly learning values. It remains unclear how value-based learning relates to the attentional selection of stimulus features that precedes overt choices, as opposed to the learning of action values that immediately trigger overt choices.

1.2 Reinforcement Learning of expected values for stimuli

Reinforcement-Learning (RL) algorithms are used to model natural and artificial decision-making systems that optimize choice behaviour on the basis of experienced outcomes (Barto,1998). The likelihood of moving to a more rewarding state by performing an action or the rules linking stimuli to reward are learned by incrementally improving predictions about future outcomes (Daw et al., 2006; Dayan and Niv, 2008). Essentially all animals, humans included, are continually faced with the problem of accurately predicting future outcomes, and the RL framework provides computational-level models of how this problem is solved. The current influence of RL studies of learning and its neural basis is due to the fact that it quantifies the unobservable elements of learning behaviour, which allows for predictions about the neural substrate

that would otherwise not be possible (Daw and Doya, 2006; Dayan and Daw, 2008; Niv, 2009).

1.2a Reinforcement Learning - principles

RL models have become an increasingly successful tool by which to describe both the behaviour related to learning value through experience but also the neural processes that represent and track the values of stimuli and actions through time (Dayan and Daw, 2008; Dayan and Niv, 2008; Lau and Glimcher, 2008; Daw et al., 2011; Nakahara and Hikosaka, 2012). The RL framework has its roots in the basic problem of learning via operant or instrumental conditioning (Niv, 2009; Shteingart and

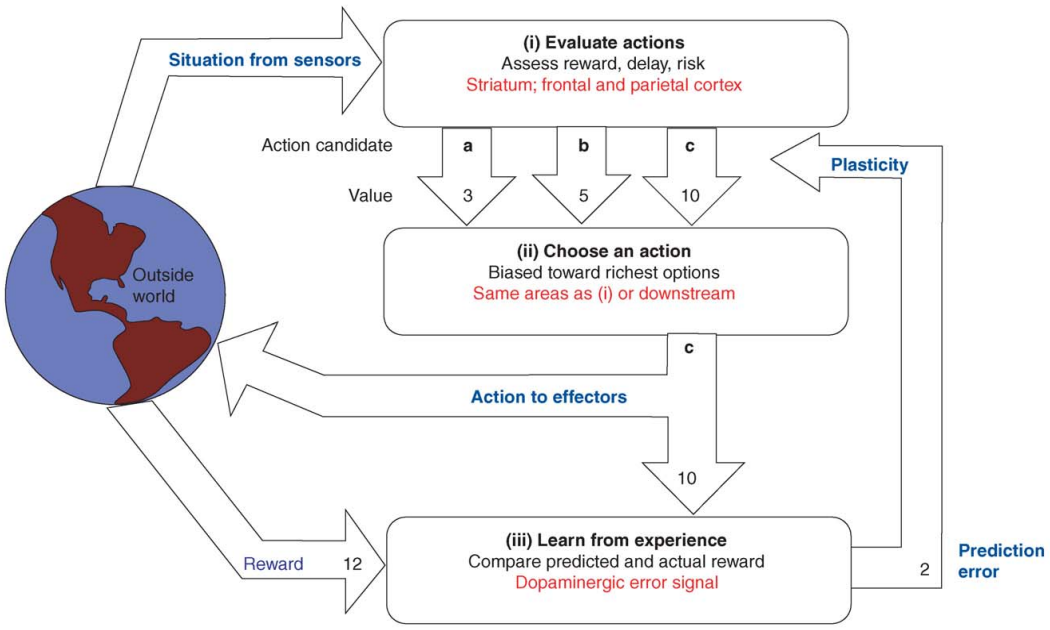


Figure 5. Reinforcement Learning describes the processes involved in instrumental conditioning. Adapted from Daw et al., 2006.

Loewenstein, 2014)(**Fig. 5**). In instrumental conditioning associations are made

between novel stimuli (CS-conditioned stimulus) and a stimulus linked to primary reward (US-unconditioned stimulus). Through these reinforced associations, animals learn to achieve goals (US) by selecting from available actions the action that is most optimal given its experience with present stimuli (CS). Learning via instrumental conditioning incorporates three separate components that are quantified in the RL framework (Daw and Doya, 2006). It involves (1) predicting the value, or the estimated long run utility (worth), of stimuli and actions. It also involves (2) the selection of an action from those available that increases the likelihood of achieving reward. And it involves (3) the updating of cached values by learning through experience. **Figure 5** provides a overview of the basic structure of learning via action and experience that is captured by reinforcement-learning (Daw and Doya, 2006).

The world - the environments in which we make decisions - provides us with sensory information about stimuli and the current situation. This sensory information is represented by the system and this representation includes estimates of the likely value or possible rewards for each stimulus. Actions for achieving outcomes, i.e. actions taken on stimuli with associated values, also have associated values based on the likelihood of achieving success, the time delay in receiving reward, and possible risks. Using these cached values, an action is chosen and an outcome received. The difference between the expected value and the experience value, what is called the reward prediction error, is used to update the expected value of stimuli and actions. The RL framework provides a computational description of each of these separable processes.

1.2a Reinforcement Learning - computational description

The basis for the formal account of learning in the RL framework is found in the Rescorla-Wagner model of pavlovian conditioning (Rescorla and Wagner, 1972). As a central principle, this model stipulates that learning occurs when expectations about outcomes are unmet (Niv, 2009). The association of a conditional stimulus with an affective stimulus, the unconditioned stimulus that provides primary reward signals, is changed in accordance with the formula in equation 1.

$$V_{new}(CS_i) = V_{old}(CS_i) + \eta \left[\lambda_{US} - \sum_i V_{old}(CS_i) \right] \qquad \text{Eq. 1}$$

Where the new associative strength (V_{new}) is equal to the value of the old strength (V_{old}) plus the scaled (η) difference between V_{old} and the actual outcome (λ). This formula provides the basis for much of the subsequent work on animal learning and the principle of the delta learning rule.

1.2c - TD Learning

The contemporary use of RL in neuroscience began with the work of Sutton and Barto and their adaptation of the Rescorla-Wagner framework (Barto, 1998; Barto and Sutton, 1998). One of the key developments in this work was the extension of the learning rule in the time domain. In the Rescorla-Wagner framework, time exists only in units of trials and this fails to take into account the timing of different events and the subsequent differences in value of different states, whereas the RL model developed by Sutton and Barto, the temporal difference (TD) learning model, represents the changing expectations of reward throughout a trial as a function of time (Niv, 2009; Sutton and Barto, 1998). In TD learning the goal of the learning system is to estimate the value of

the current state, with its available stimuli and actions, as a function of the future rewards or punishments predicted by the elements of the current state. TD learning assumes that estimates of current value are a function of their distance in time to receiving reward (**Eq. 2**).

$$V_{new}(S_{i,t}) = V_{old}(S_{i,t}) + \eta \left[r_t + \gamma \sum_{S_k @ t+1} V_{old}(S_{k,t+1}) - \sum_{S_j @ t} V_{old}(S_{j,t}) \right] \quad \text{eq. 2}$$

In this formulation of TD learning we see that the new value of the current state, $V_{new}(S_{i,t})$, is equal to the old value of the state, $V_{old}(S_{i,t})$, plus the scaled (η) information about rewards at this time step, which is equal to any reward received now (r_t) plus the discounted expectation of reward in the future. TD learning models have proved to be very effective at predicting the learning behaviour of many types of intelligent systems and has become one of the most influential developments in contemporary computational neuroscience of animal learning (O'Doherty et al., 2003; Niv et al., 2005; Maia, 2009).

1.2d Q-Learning

Another influential formal account of learning in the RL framework is the Q-Learning model. Q-Learning is an adaptation and a simplification of the actor-critic formulation of TD learning (Watkins and Dayan, 1992). An actor-critic learner employs separable systems of action and action critique, where the actor makes choices using a policy, and a policy is the set of actions that will lead to reward and minimize punishment in the current state. Policies are learned through the critiquing of actions via the critic that tracks the values of states, actions and stimuli (Niv, 2009). Q-Learning,

by contrast, eliminates the separable processes, and chooses actions directly based on the learned value associated with that action in this state, the Q-value (**Eq. 3**). The prediction error, δ_t, of Q-learning (**Eq. 4**) is similar to the learning rule of the Rescorla-

$$Q(S_t, a_t)_{new} = Q(S_t, a_t)_{old} + \eta\delta_t \qquad \text{Eq. 3}$$

$$\delta_t = r_t + \gamma Q(S_{t+1}, a_{t+1}) - Q(S_t, a_t) \qquad \text{Eq. 4}$$

Wagner model. The Q-value associated with any stimulus or action is updated according to the reward experienced at (t) plus the scaled difference between the expected value and its current value. Rather than representing policies that transition the model from one state to another, Q-learning represents, and then selects from,

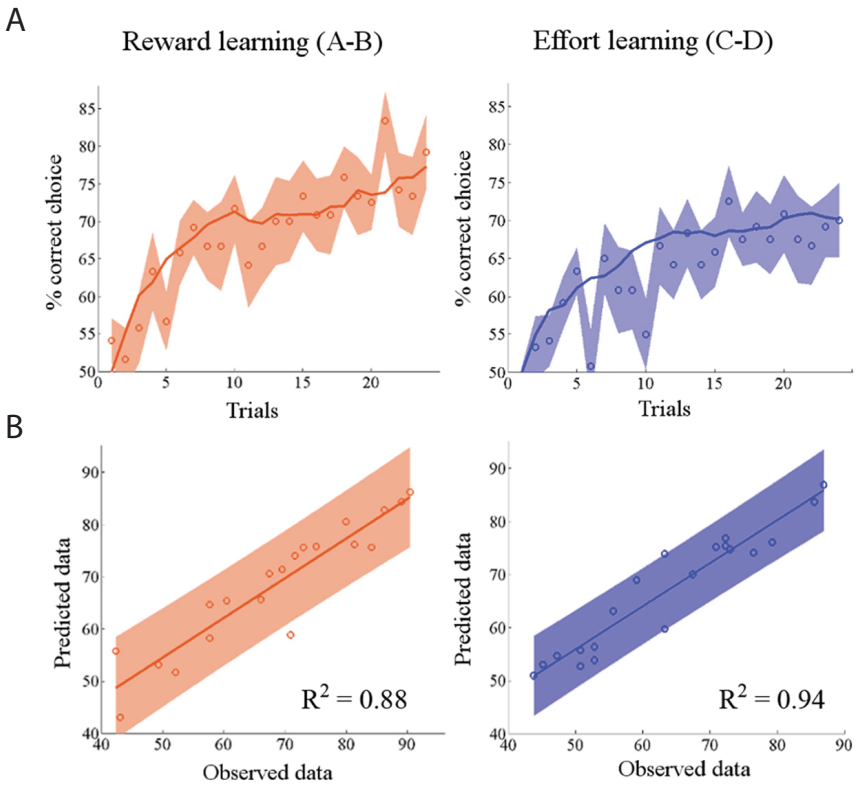


Figure 6. A Q-Learning model with four independent free parameters is highly predictive of both learning about rewards and efforts. A) Circles show average responses across sessions and subjects, with shaded area the SEM across subjects. Regression line is average choice behaviour from best fitting QL model. B) Plotting observed data against predicted shows the coefficient of determination for both task conditions. Adapted from Skvortsova et al., 2014.

values for actions and stimuli directly. Along with TD learning models, Q-learning models in various forms have also been successfully used to elucidate the computational processes underlying learning behaviour and its neural basis (Littman, 2001; Li and Daw, 2011).

Learning optimal responses to stimuli often involves learning about multiple things simultaneously, such as learning how to both maximize reward while learning how to minimize risks or efforts involved in receiving reward. Skvortsova et. al. (Skvortsova et al., 2014) recently showed how Q-Learning RL models are highly predictive of exactly this behaviour. Considering many different versions of Q-Learning models, each with parameter sets corresponding with different functional hypotheses, as well as other non-Q-Learning models, Skvortsova et. al. showed that Q-Learning is highly predictive of both learning values related to rewarding outcomes and to effort costs (**Fig. 6**) Q-Learning as a formal account of human learning processes provides powerful predictions about the mechanisms underlying adaptive choice behaviour.

1.2e Model-Free RL, Model-Based RL, and Hierarchical RL

RL models can be grouped into different classes based on their internal representation of the environment, i.e. whether or not they incorporate a world model. Model-Free RL, as the name suggests, does not include a model of the environment, which means that values are acted on directly through a selection process without the possibility of considering other non-value information about the current environment. Therefore a model-free learner is highly driven by recent reinforcement. Both the TD and Q-Learning models as outlined above are model-free.

In contrast, Model-Based RL incorporates a representation of the environment based on previous learning. Model-Based RL is a development of the RL framework that is meant to capture the flexibility of goal directed behaviour as opposed to the habitual responses of Model-Free(Dolan and Dayan, 2013; Worthy et al., 2014). Model-Free is habitual in its responses because it directly responds to reinforcement - rewarded actions tend to be repeated. Model-Based RL avoids this inflexibility because it is able to simulate possible outcomes independently of recent reward history(Glascher et al., 2010; Daw et al., 2011). For example, a rodent employing a model-free system in a maze that provides stochastic rewards is likely to return to a recently rewarded location, even if that reward is unlikely to be repeated. On the other hand, if the same rodent was employing a model-based system that contains a map of the maze that represents the likelihoods of achieving rewards, it will flexibly discount the recent reward in its consideration of possible actions and consider series of actions that are more adaptive (Doll et al., 2012).

To test the role of planning and model use in decision making, Daw et al (Daw et al., 2011) devised a task where subjects are trained on a two stage task where the likelihood of arriving in any state from some prior one is probabilistic, following a

schedule, and the likelihood of reward from all second stage states is also probabilistic.

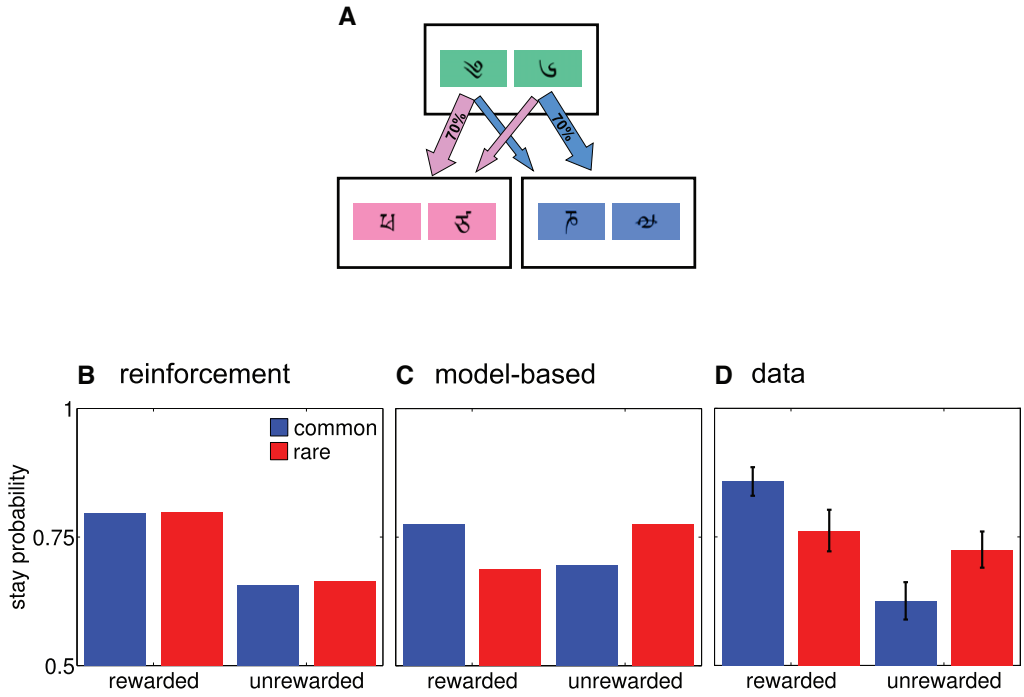


Figure 7. Model-Based RL proposes a strong hypothesis about subject responses to reinforcement under probabilistic circumstances compared to a model-free system. A) A subject that has no model about the task responds to reinforcement identically in both likely and unlikely circumstances. B-C) A model-based learner responds differentially to reinforcement by recognizing the likelihood of reinforcement being repeated. D) Observed data matches the different responses trends predicted by model-based RL. Adapted from (Daw et al., 2011).

According to the hypothesis of model-based RL, subjects that know the likelihoods of transition at each stage of the task are not likely to repeat the series of choices that led to a reward from a low probability transition. Using an internal model of the task, subjects will respond to rewards from low-probability transitions by selecting the alternate option on the next trial.

Daw et. al found that, as predicted, the likelihood of staying with a set of choices following a common transition, when rewarded, is significantly higher than following a rewarded uncommon transition. Conversely, Daw et. al. found that following a common

unrewarded transition, subjects were significantly less likely to stay, as compared to an unrewarded unlikely transition. This results suggests that subjects are employing previously learned information about the task to model state transitions.

Hierarchical RL (HRL) represents a more recently developed class of RL models that provides a solution to the problem of scale. Many normal decision making environments involve huge numbers of possible actions in responses to stimuli. Hierarchical strategies for learning optimal responses to stimuli extend computational power by means of temporal abstraction. Temporal abstraction in this context refers to the ability to apply a learned rule in a context that is temporally removed from the context where it was learned (Botvinick, 2012). As opposed to being limited in choice behaviour to primitive stimulus-action responses, a hierarchical agent make selections from 'high-level' options that group together sets of temporally abstract actions

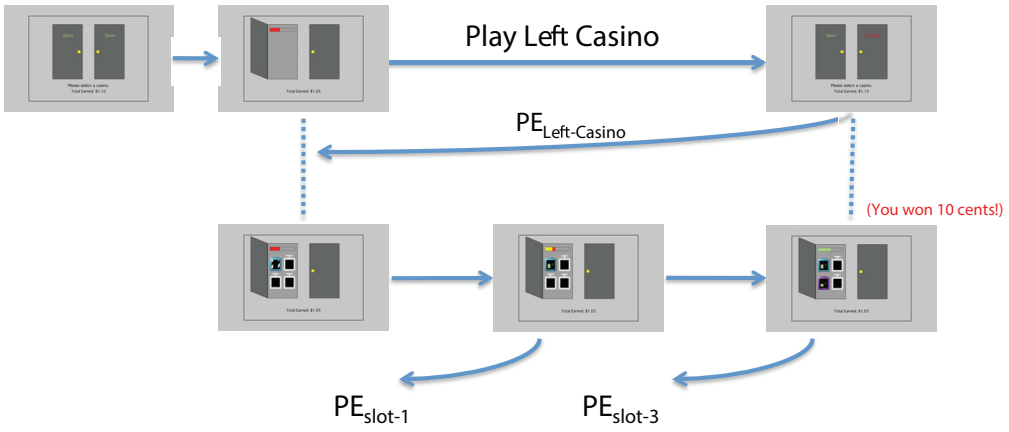


Figure 8. Prediction errors are separable in hierarchical learning according to the different levels of learning. Real-life learning involves learning about different task levels simultaneously, which results in unique prediction errors that overlap in time. Subjects performing this task learn about high level choices, the selection of a casino, and low-level choices, individual games, simultaneously as the outcomes of games informs about the value of a casino for goals. Adapted from Diuk et al., 2013.

(Botvinick et al., 2009; Botvinick, 2012). For example, I can choose to 'make breakfast',

as opposed to ‘go to work’, and my choice then initiates a sequence of actions - ‘turn on stove’, ‘get frying pan’, ‘get eggs’, etc - that expedites reaching a goal by abstracting an action sequence from previous learning.

Recent work from Diuk et al. (Diuk et al., 2013) has explored how subjects deploy hierarchical responses and the accompanying neural activations in tasks with hidden hierarchical structure. In their study, subjects gamble virtually in a set of casinos, where each casino contains a set of slot machines, with different payout probabilities.

The advantage of using a hierarchical strategy for learning is that choice behaviour can be adapted more rapidly when learning occurs at multiple levels, both at the level of local outcome - the reward of an individual slot machine, and also at the high level of sets of outcomes towards a goal - the overall outcome of the casino. Diuk et al. (Diuk et al., 2013) found that subjects learn at both levels, and also that the unique prediction errors produced by simultaneous local and higher learning correlated with activation in the basal ganglia.

1.2f Context, prior learning, and heterogenous world models

It is widely recognized that no single RL framework is capable of capturing the total dynamics of human learning, and that there are likely different systems that employ one set of mechanisms versus another given a set of circumstances, and/or that RL systems in the brain are interconnected and interact continuously (Dayan and Daw, 2008; Dayan and Niv, 2008; Seo and Lee, 2008; Glascher et al., 2010; Worthy et al., 2012; Worthy and Maddox, 2014). Perhaps the chief difficulty faced by formal accounts of learning is that human learners enter new environments with a wide range of previous experience

all of which directly impacts expectations and internal models. It has been noted that even in the simplest of learning task, human behaviour can be highly varied and that this is likely due to heterogenous world models (Shteingart and Loewenstein, 2014). Future work on the neural processes underlying the computational processes of human learning will likely need to take into account the influence of previous learning when evaluating choice behaviour in experimental contexts.

1.3 Neuronal basis of reinforcement learning of stimulus values

Beyond making highly accurate predictions of learning behaviour in human and non-human animals, the importance of RL is its ability to make clear predictions about the unobservable elements involved in learning. There have been many findings across multiple species that link specific variables of RL models to neuronal activity. Although an understanding of learning in the brain is far from complete, there are several significant findings that have been widely reproduced and implicate specific areas in the brain as playing a role in the computations of reinforcement learning, notably the striatum, and pre-frontal cortex.

1.3a Neuronal circuits underlying value-based decision-making

The earliest and most well-known result linking RL models to activity in the brain is known as the reward prediction error (RPE) hypothesis of dopaminergic activity. As was discussed above, the reward prediction error of an RL model represents the difference between expected and observed outcomes (Niv et al., 2005; Schultz, 2006).

The correlation between the phasic activity of striatal dopamine neurons and the reward prediction error was first observed in single unit recording of monkeys, but this result has since been replicated in humans. Seymour et. al. used fMRI to show that in

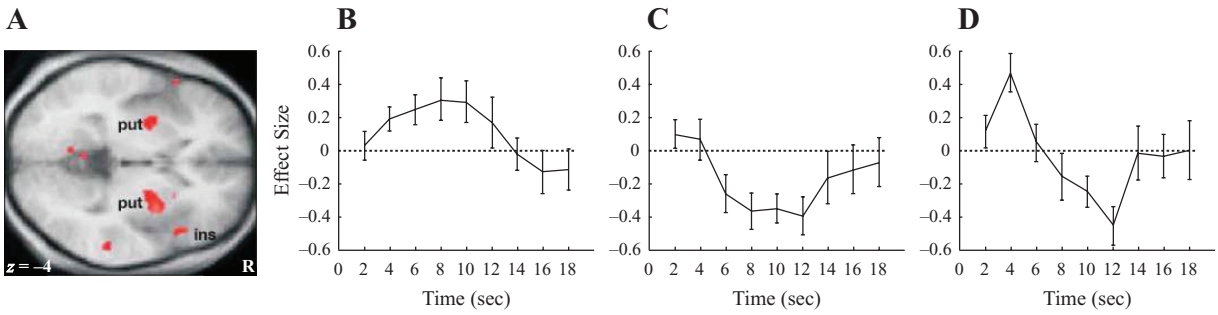


Figure 9. Activation of the putamen in human subjects is correlated with predictions of reward prediction error activity made by TD RL models. A) Figure shows the statistical parametric map of significant activations in the putamen related to RPE. B-D) Time-course of different prediction errors. B) positive prediction error. C) Negative prediction error. D) Biphasic prediction error, positive following the cue and negative following feedback. Adapted from (Seymour et al., 2004).

normal human subjects, the BOLD response in the putamen mirrored the predictions made by a TD RL model (**Fig. 9**)(Seymour et al., 2004).

According to the RL framework, the reward prediction error is used to update cached values for actions and stimuli, and in addition to finding neuronal evidence for the RPE there is also evidence that subjective values are represented in the activity of neurons in ventromedial and orbital frontal cortex. The study discussed above by Skvortsova et. al. (Skvortsova et al., 2014) used fMRI to analyze the BOLD response of subjects in relation to the variables produced by the best fit Q-Learning model. Skvortsova et. al. found that activation in ventromedial prefrontal cortex (vmPFC) was highly predictive of values for stimuli estimated by the model (**Fig. 10**)(Skvortsova et al., 2014).

Other studies have likewise implicated medial prefrontal areas in value-based

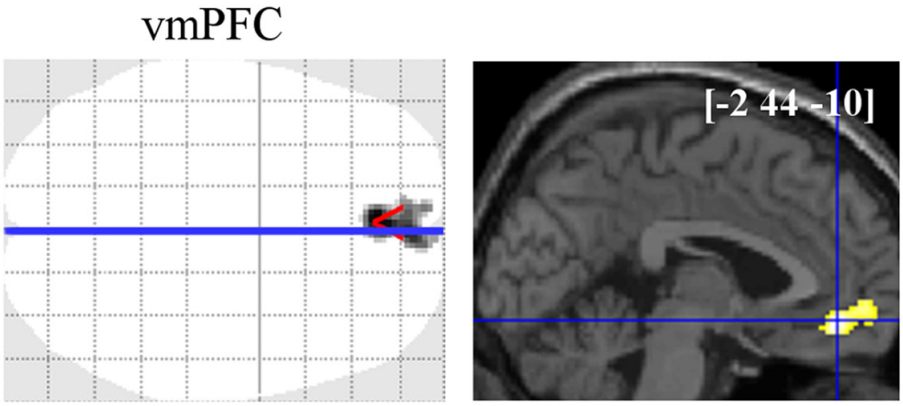


Figure 10. Figure shows statistical parametric maps of voxels where activations at cue onset are significantly correlated with expected reward. Adapted from Skvortsova et al., 2014.

learning and decision making. In another recent work, Chib et. al. (Chib et al., 2009) explored whether or not activity in vmPFC only represented some kinds of values, such as values related to goals, or if vmPFC represented a wide range of subjective values.

Figure 11. Figure shows statistical parametric maps for three studies overlaid, with peak activations co-occurring in vmPFC for values related to food, money, and other goods. Adapted from Chib et al., 2009.

They found that activations in vmPFC reflect a common currency of value - representations for value-based decision variables are localized in vmPFC despite reflecting different inter-subject intrinsic values and representing values for different classes of stimuli, such as food, money and other objects (see **Fig. 11)**.

There are many results linking decision variables from RL models to activation in pre-frontal cortex, however due to the fact that most studies of RL and the human brain use fMRI and are limited to the BOLD response, there are still many questions unresolved about the time course of activity and the variables of RL, as well as the dynamics of activity at the level of electrophysiological response.

1.3b ECoG and cellular activity in the cortex

Investigations into the relationship between learning, the computational mechanisms of RL and activity in human prefrontal cortex have frequently relied on fMRI measurements of the hemodynamic response (BOLD signal)(Hare et al., 2008; Chib et al., 2009; Wunderlich et al., 2010; Hare et al., 2011). While there are many advantages to using fMRI, such as its non-invasive nature and the ability to study many different subject populations, there are advantages to using other methods as well due to some of the limitations of collecting and analyzing BOLD activity, such as its dependence on inter-subject averages and low-resolution (comparatively) in the time domain (Logothetis, 2008). One alternative method is the use of intracranial electroencephalography (iEEG) which measures the electrocorticogram, or ECoG, a time series signal of voltage changes in the cortical surface.

One of the strengths of using ECoG to understand brain function as it relates to learning behaviour is that it is a direct measurement of cortical activity (Miller et al., 2007; 2010) as opposed to the surrogate signal of the BOLD response (Logothetis, 2008). Cellular processes in the brain give rise to electrical currents that superimpose in a volume of tissue to produce a electrical potential relative to a reference potential (Buzsáki et al., 2012). The difference between the potential and the reference measured via an electrode on the tissue surface is the ECoG signal, and studies suggest that the electrical potential of ECoG is primarily driven by the summed synaptic activity in a local population or ensemble of cells, although there are minor contributions from other sources as well (Buzsáki et al., 2012). The same activity measured at the scalp is the EEG (electroencephalography) signal and when measured from within a volume of tissue using micro-electrodes is the LFP (local field potential)(Buzsáki et al., 2012). One of the particular strengths of ECoG is that electrical currents and fields can be recorded with sub-millisecond precision, which allows for analysis about basic mechanisms of brain function that operate on the millisecond timescale, as well as the relationship between basic mechanisms and higher function, such as learning and attentional selection (Buzsáki et al., 2012).

Another strength of ECoG as a tool for understanding the mechanisms of cognitive functions, over fMRI, is that as a time series it can be decomposed into its elementary components, frequency specific oscillations. Oscillations in the ECoG signal are related to the rhythmic fluctuations of neuronal groups (Buzsáki et al., 2012) and oscillations in different frequencies have been related to a range of cognitive function (Fries, 2005; Womelsdorf et al., 2007), including selective attention (Fries, 2001; Womelsdorf et al.,

2005), long term memory (Buzsáki and Watson, 2012; Johnson and Knight, 2014), and working memory (Axmacher et al., 2008).

1.3c ECoG analysis in the time-frequency domain

One method of decomposing the ECoG signal is accomplished by using the Fourier theorem to convolve its continuous periodic activity into frequency specific activity. The Fourier theorem provides the analytic basis for time series signal analysis and according to the theorem any continuous oscillation can be represented by a set of frequency specific time series components. Convolution via the Fourier theorem, or by other methods (Bruns, 2004), decomposes the signal by combining the original signal with a number of wavelet kernels - sinusoidal functions with a specific periodic frequency. Convolving the input signal with wavelets according to equation 5 produces a time-series in the frequency domain with each time point having a corresponding

$$\tilde{x}(f) = \sum_{t=1}^{N} \exp(-2\pi i f t_n) \qquad \text{Eq. 5}$$

amplitude for each frequency kernel used. Spectral power, a common measure of functional change in neural activity, is produced by squaring the amplitude.

1.3e Spectral power changes and behaviour

Changes in spectral power have been linked to the behavioural adaptation of subjects in a variety of tasks. A recent study by van Driel et. al.(van Driel et al., 2012)

used EEG to explore power changes related to performance and attentional control in

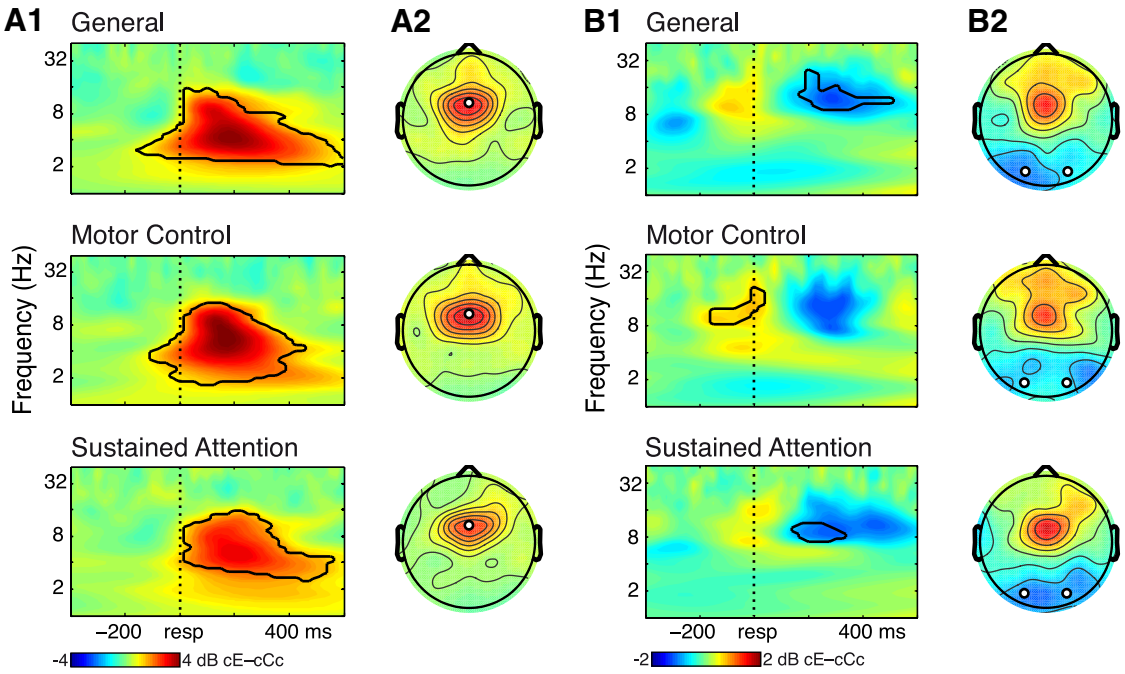


Figure 12. Spectral power plots for different electrode locations showing the difference in power between error and correct trials for three different conditions. A1) Frontal midline theta power increases significantly across all conditions on error trials that follow correct trials as compared to correct trials that follow correct trials (CE-CC). A2) Topographical plot showing power as function of electrode location. B1) Parietal-Occipital alpha power decreases significantly from baseline only in the sustained attention condition, as opposed to the motor control condition (black outline shows areas with significant change). B2) Plot shows power as a function of parietal-occipital electrode location. Adapted from van Driel et al., 2012.

variations of the Simon task (Simon and Rudell, 1967), a task designed to explore conflict monitoring. van Driel et al. found that separable error types, defined by task structure and cognitive demands, produced spectral power changes following an error that is separable by frequency and location. Errors in more attentionally demanding blocks produced power increases in frontal mid-line theta (**Fig. 12**). Errors in less attentionally demanding blocks were accompanied by power decrease in the alpha band in parietal-occipital areas (**Fig. 12**).

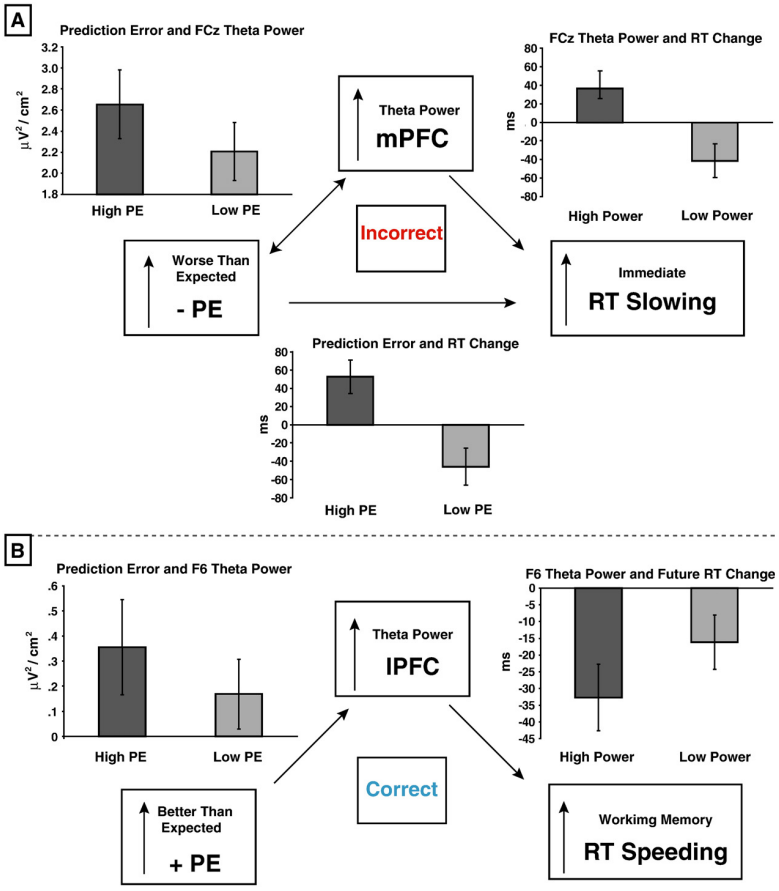


Figure 13. Prediction errors produced by a Q-Learning RL model, fit to the individual behaviour of subjects, are related to both spectral power changes and behavioural adaptation in the form of reaction time changes. A) The greater the magnitude of a negative prediction error on an error trial, the greater the increase in medial prefrontal theta power, and a slower reaction time on the following trial. B) The greater the magnitude of positive prediction error on a correct trial, the greater the increase in theta power in lateral prefrontal cortex and the faster the reaction time on the next trial of that type. Adapted from Cavanagh et al., 2010.

Frequency specific power has also been recently linked to the values estimated by RL models (Frank et al., 2015). One of the strengths of time-frequency analysis of ECoG time series is that it permits single trial regression analysis, enabling statistically powerful comparisons with the trial by trial estimates of value produced by RL models. In a study by Cavanaugh et. al (Cavanagh et al., 2010) subjects were required to learn the value (the expected likelihood of positive outcome) of visual stimuli, where each stimulus was associated with positive outcomes according to a probabilistic schedule.

Using EEG to monitor changes in activity related to learning the value of stimuli, Cavanagh et al fit Q-Learning models to the choice behaviour of subjects. Using a GLM regression analysis of the trial by trial frequency specific power changes in the EEG signal and the trial by trial prediction errors (RPE) produced by the best fit model, Cavanagh et. al. discovered a significant correlation between changes in theta band (4-8Hz) power and RPE following feedback (see **Fig. 13**).

1.4 Purpose of this dissertation

The purpose of this dissertation is to investigate the computational relationship between attentional selection and reinforcement learning processes in the brain by using formal models of RL to 1) model the choice behaviour of macaques performing a selective attention task, 2) model the choice behaviour of human subjects learning the value of stimulus features in an uncued and untrained context, and 3) produce single trial estimates of stimulus value for interrogating ECoG signals in ventromedial prefrontal cortex of human subjects as they learn values for stimulus features (Mars et. al 2012).

Learning is a fundamental characteristic of human behaviour and a basic element of brain function. Little is known about how learning mechanisms are related to other cognitive control mechanisms like attentional selection (Kruschke and Hullinger, 2010). RL provides a formal account of learning that allows for specific hypotheses about the mechanisms linking learning and shifts of attention, as well as learning related changes in the oscillatory activity in the human cortex. ECoG provides a unique insight into cortical activity due to its spatiotemporal sensitivity and allows for statistical powerful analyses of learning when paired with the computational predictions of RL.

The approach taken in this study is motivated by two significant gaps in the exploration of learning and attention. First, RL models have been used with great effect in cognitive neuroscience to elucidate the neuronal basis of learning in the brain, but despite the fact that attentional selection has been shown to be influenced by learned values for stimuli, RL models have not yet been used to quantify the trial by trial relationship between covert attentional selection of stimuli and their expected value (Anderson, 2013; Anderson et. al. 2013; Gottlieb, 2012). Second, it is not clear how value-related information is represented and processed in the brain. While RL models have been used to show that the activity of single neurons and the hemodynamic response of neuronal populations are related to the processing of stimulus value information, they have not yet been used to explore how this information is linked to the rhythmic fluctuations of neuronal activity recorded directly from the cortex (Buzsaki & Watson, 2012; Dayan & Niv, 2008). This study aims to fill these gaps.

1.5 References

Anderson BA (2013) A value-driven mechanism of attentional selection. Journal of Vision 13:7–7.

Anderson BA, Laurent PA, Yantis S (2011) Value-driven attentional capture. PNAS 108:10367–10371.

Anderson BA, Laurent PA, Yantis S (2013) Reward predictions bias attentional selection. frontiers in Human Neuroscience 7:262.

Anderson BA, Yantis S (2013) Persistence of value-driven attentional capture. Journal of Experimental Psychology: Human Perception & Performance 39:6–9.

Axmacher N, Schmitz DP, Wagner T, Elger CE, Fell J (2008) Interactions between Medial Temporal Lobe, Prefrontal Cortex, and Inferior Temporal Regions during Visual Working Memory: A Combined Intracranial EEG and Functional Magnetic Resonance Imaging Study. Journal of Neuroscience 28:7304–7312.

Barto, Andrew G. Reinforcement learning: An introduction. MIT press, 1998.

Botvinick MM (2012) Hierarchical reinforcement learning and decision making. Current Opinion in Neurobiology 22:956–962.

Botvinick MM, Niv Y, Barto AC (2009) Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. Cognition 113:262–280.

Bruns A (2004) Fourier-, Hilbert- and wavelet-based signal analysis: are they really different approaches? J Neurosci Methods 137:321–332.

Buzsáki G, Anastassiou CA, Koch C (2012) The origin of extracellular fields and currents--EEG, ECoG, LFP and spikes. Nature Publishing Group 13:407–420.

Buzsáki G, Watson BO (2012) Brain rhythms and neural syntax: implications for efficient coding of cognitive content and neuropsychiatric disease. Dialogues Clin Neurosci 14:345–367.

Cavanagh JF, Frank MJ, Klein TJ, Allen JJB (2010) Frontal theta links prediction errors to behavioral adaptation in reinforcement learning. NeuroImage 49:3198–3209.

Chelazzi L, Perlato A, Santandrea E, Libera Della C (2013) Rewards teach visual selective attention. Vision Res 85:58–72.

Chib VS, Rangel A, Shimojo S, O'Doherty JP (2009) Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. J Neurosci 29:12315–12320.

Daw ND, Doya K (2006) The computational neurobiology of learning and reward. Current Opinion in Neurobiology 16:199–204.

Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ (2011) Model-Based Influences on Humans' Choices and Striatal Prediction Errors. Neuron 69:1204–1215.

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. Nature 441:876–879.

Dayan P, Daw ND (2008) Decision theory, reinforcement learning, and the brain. Cognitive, Affective, & Behavioral Neuroscience 8:429–453.

Dayan P, Kakade S, Montague PR (2000) Learning and selective attention. Nat Neurosci 3:1218–1223.

Dayan P, Niv Y (2008) Reinforcement learning: The Good, The Bad and The Ugly. Current Opinion in Neurobiology 18:185–196.

Desimone R, Duncan J (1995) Neural mechanisms of selective visual attention. Annu Rev Neurosci 18:193–222.

Diuk C, Tsai K, Wallis J, Botvinick M, Niv Y (2013) Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia. Journal of Neuroscience 33:5797–5805.

Dolan RJ, Dayan P (2013) Goals and habits in the brain. Neuron 80:312–325.

Doll BB, Simon DA, Daw ND (2012) The ubiquity of model-based reinforcement learning. Current Opinion in Neurobiology:1–7.

Droll JA, Abbey CK, Eckstein MP (2009) Learning cue validity through performance feedback. Journal of Vision 9:18–18.

Frank MJ, Badre D (2015) How cognitive theory guides neuroscience. Cognition 135:14–20.

Frank MJ, Gagne C, Nyhus E, Masters S, Wiecki TV, Cavanagh JF, Badre D (2015) fMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. J Neurosci 35:485–494.

Fries P (2001) Modulation of Oscillatory Neuronal Synchronization by Selective Visual Attention. Science 291:1560–1563.

Fries P (2005) A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. Trends in Cognitive Sciences 9:474–480.

Glascher J, Daw N, Dayan P, Doherty JPO (2010) States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. Neuron 66:585–595.

Gottlieb J (2012) Attention, learning, and the value of information. Neuron 76:281–295.

Hare TA, O'Doherty J, Camerer CF, Schultz W, Rangel A (2008) Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. J Neurosci 28:5623–5630.

Hare TA, Schultz W, Camerer CF, O'Doherty JP, Rangel A (2011) Transformation of stimulus value signals into motor commands during simple choice. Proceedings of the National Academy of Sciences 108:18120–18125.

Johnson EL, Knight RT (2014) Intracranial recordings and human memory. Current Opinion in Neurobiology 31C:18–25.

Kaping D, Vinck M, Hutchison RM, Everling S, Womelsdorf T (2011) Specific Contributions of Ventromedial, Anterior Cingulate, and Lateral Prefrontal Cortex for Attentional Selection and Stimulus Valuation Behrens T, ed. PLoS Biol 9:e1001224.

Kruschke JK, Hullinger RA (2010) Evolution of attention in learning Schmajuk N, ed. Computational models of conditioning.

Lau B, Glimcher PW (2008) Value representations in the primate striatum during matching behavior. Neuron 58:451–463.

Li J, Daw ND (2011) Signals in human striatum are appropriate for policy update rather than value prediction. J Neurosci 31:5504–5511.

Littman ML (2001) EZProxy. ICML.

Logothetis NK (2008) What we can do and what we cannot do with fMRI. Nature 453:869–878.

Maia TV (2009) Reinforcement learning, conditioning, and the brain: Successes and challenges. Cognitive, Affective, & Behavioral Neuroscience 9:343–364.

Mars RB, Shea NJ, Kolling N, Rushworth MFS (2012) Model-based analyses: Promises, pitfalls, and example applications to the study of cognitive control. The Quarterly Journal of Experimental Psychology 65:252–267.

Miller EK, Buschman TJ (2013) Cortical circuits for the control of attention. Current Opinion in Neurobiology 23:216–222.

Miller KJ, Hermes D, Honey CJ, Sharma M, Rao RPN, Nijs den M, Fetz EE, Sejnowski TJ, Hebb AO, Ojemann JG, Makeig S, Leuthardt EC (2010) Dynamic modulation of local population activity by rhythm phase in human occipital cortex during a visual search task. frontiers in Human Neuroscience 4:197.

Miller KJ, Leuthardt EC, Schalk G, Rao RPN, Anderson NR, Moran DW, Miller JW, Ojemann JG (2007) Spectral Changes in Cortical Surface Potentials during Motor Movement. J Neurosci 27:2424–2432.

Nakahara H, Hikosaka O (2012) Learning to represent reward structure: A key to adapting to complex environments. Neuroscience Research 74:177–183.

Niv Y (2009) Reinforcement learning in the brain. Journal of Mathematical Psychology.

Niv Y, Duff MO, Dayan P (2005) Dopamine, uncertainty and TD learning. Behav Brain Funct 1:6.

O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ (2003) Temporal difference models and reward-related learning in the human brain. Neuron 38:329–337.

Peck CJ, Jangraw DC, Suzuki M, Efem R, Gottlieb J (2009) Reward modulates attention independently of action value in posterior parietal cortex. J Neurosci 29:11182–11191.

Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Classical conditioning

II: Current research and theory (Black AH, Prokasy WF, eds), pp 64–99. New York: Appleton-Century-Crofts.

Reynolds JH, Chelazzi L (2004) Attentional modulation of visual processing. Annu Rev Neurosci 27:611–647.

Rombouts JO, Bohte SM, Martinez-Trujillo J, Roelfsema PR (2015) A learning rule that explains how rewards teach attention. Visual Cognition:1–27.

Rosenzweig MR, Bennett EL, Diamond MC (1972) Brain changes in response to experience. Scientific American.

Rushworth MF, Kolling N, Sallet J, Mars RB (2012) Valuation and decision-making in frontal cortex: one or many serial or parallel systems? Current Opinion in Neurobiology 22:946–955.

Rushworth MFS, Noonan MP, Boorman ED, Walton ME, Behrens TE (2011) Frontal Cortex and Reward-Guided Learning and Decision-Making. Neuron 70:1054–1069.

Schultz W (2006) Behavioral Theories and the Neurophysiology of Reward. Annu Rev Psychol 57:87–115.

Seo H, Lee D (2008) Cortical mechanisms for reinforcement learning in competitive games. Philosophical Transactions of the Royal Society B: Biological Sciences 363:3845–3857.

Seymour B, O'Doherty JP, Dayan P, Koltzenburg M, Jones AK, Dolan RJ, Friston KJ, Frackowiak RS (2004) Temporal difference models describe higher-order learning in humans. Nature 429:664–667.

Shteingart H, Loewenstein Y (2014) Reinforcement learning and human behavior. Current Opinion in Neurobiology 25:93–98.

Simon JR, Rudell AP (1967) Auditory S-R compatibility: the effect of an irrelevant cue on information processing. J Appl Psychol 51:300–304.

Skvortsova V, Palminteri S, Pessiglione M (2014) Learning To Minimize Efforts versus Maximizing Rewards: Computational Principles and Neural Correlates. J Neurosci 34:15621–15630.

Sutton RS, Barto AG (1998) Reinforcement learning: An introduction. 1998.

Tsotsos JK (2011) A Computational Perspective on Visual Attention. MIT Press.

van Driel J, Ridderinkhof KR, Cohen MX (2012) Not all errors are alike: theta and alpha EEG dynamics relate to differences in error-processing dynamics. J Neurosci 32:16795–16806.

Wallis JD (2007) Orbitofrontal cortex and its contribution to decision-making. Annu Rev Neurosci 30:31–56.

Watkins CH, Dayan P (1992) Q-learning. Mach Learn 8:279–292.

Womelsdorf T, Fries P, Mitra PP, Desimone R (2005) Gamma-band synchronization in visual cortex predicts speed of change detection. Nature 439:733–736.

Womelsdorf T, Schoffelen JM, Oostenveld R, Singer W, Desimone R, Engel AK, Fries P (2007) Modulation of Neuronal Interactions Through Neuronal Synchronization. Science 316:1609–1612.

Worthy DA, Cooper JA, Byrne KA, Gorlick MA, Maddox WT (2014) State-based versus reward-based motivation in younger and older adults. Cognitive, Affective, & Behavioral Neuroscience:1–13.

Worthy DA, Hawthorne MJ, Otto AR (2012) Heterogeneity of strategy use in the Iowa gambling task: A comparison of win-stay/lose-shift and reinforcement learning models. Psychon Bull Rev 20:364–371.

Worthy DA, Maddox WT (2014) A comparison model of reinforcement-learning and win-stay-lose-shift decision-making processes: A tribute to W.K. Estes. Journal of Mathematical Psychology 59:41–49.

Wunderlich K, Rangel A, O'Doherty JP (2010) Economic choices can be made using only stimulus values. PNAS 107:15005–15010.

Chapter 2.

Attentional selection can be predicted by reinforcement learning of task-relevant stimulus features weighted by value-independent stickiness

under review at **Journal of Cognitive Neuroscience**

Matthew Balcarras[1]*, Salva Ardid[1,2]*, Daniel Kaping[1], Stefan Everling[3], Thilo Womelsdorf[1,3]

[1]Department of Biology, Centre for Vision Research, York University, 4700 Keele Street, Toronto, Ontario M6J 1P3, Canada. [2]Center for Computational Neuroscience and Neural Technology (CompNet), Department of Mathematics and Statistics, Boston University, Boston, Massachusetts 02215. [3]Department of Physiology and Pharmacology, University of Western Ontario, 100 Perth Drive, London, Ontario N6A 5K8, Canada.

* These authors contributed equally

Correspondence: Dr. Thilo Womelsdorf, Dr. Salva Ardid, Mr. Matthew Balcarras.
York University, Department of Biology, 4700 Keele Street,
Toronto ON M3J 1P3, Canada

2.1 Abstract

Attention includes processes that evaluate stimuli relevance, select the most relevant stimulus against less relevant stimuli, and bias choice behavior towards the selected information. It is not clear how these processes interact. Here, we captured these processes in a reinforcement learning framework applied to a feature-based attention task that required macaques to learn and update the value of stimulus features while ignoring non-relevant sensory features, locations, and action plans. We found that value based reinforcement learning mechanisms could account for feature-based attentional selection and choice behaviour, but required a value-independent stickiness selection process to explain selection errors while at asymptotic behavior. By comparing different reinforcement learning schemes we found that trial-by-trial selections were best predicted by a model that only represents expected values for the task relevant feature dimension, with non-relevant stimulus features and action plans having only a marginal influence on covert selections. These findings show that attentional control subprocesses can be described by (1) the reinforcement learning of feature values within a restricted feature space that excludes irrelevant features, (2) a stochastic selection process on feature specific value representations, and (3) value-independent stickiness towards previous selections. We speculate that these three mechanisms are implemented by distinct but interacting brain circuits and that the proposed formal account of attentional selection will be important to understand how attentional sub-processes are implemented in primate brain networks.

2.2 Introduction

Selective attention can be defined as a set of processes that work around resource limitations by prioritizing processing to goal relevant information (Tsotsos, 2011), while ensuring flexibility to adapt to new situations (Dayan et al., 2000; Kruschke and Hullinger, 2010; Ardid and Wang, 2013). Such a definition of attention implicitly assumes a continuous evaluation of the relevance of sensory information (Kaping et al., 2011; Gottlieb, 2012), which entails computing value predictions of stimulus features (Rushworth et al., 2011; Rangel and Clithero, 2012; Anderson, 2013; Chelazzi et al., 2013). Consistent with this suggestion, recent neurophysiological studies have shown that representations of stimulus value affect attentional search performance and gaze allocation in human subjects (Libera and Chelazzi, 2009; Anderson et al., 2011; Tatler et al., 2011) and underlie economic choices (Wunderlich et al., 2010; Hare et al., 2011; Padoa-Schioppa, 2011). Furthermore, neural correlates of those signals have been found in prefrontal and parietal neurons as well as in subcortical neural circuits (Peck et al., 2009; Kaping et al., 2011; Kennerley et al., 2011; Luk and Wallis, 2013; Peck et al., 2013; Cai and Padoa-Schioppa, 2014). However, it is unclear how value-based learning relates to the attentional selection of stimulus features that precedes overt choices, as opposed to the learning of action values that immediately triggers overt choices (Lau and Glimcher, 2005; Glimcher, 2011). To elucidate the mechanisms that underlie attention, task paradigms and analyses need to isolate the learning of covert (attentional) stimulus selection from processes linked to overt choice such as perceptual discrimination and action planning (Rangel and Clithero, 2012).

In the decision-making domain, reinforcement learning (RL) provides a framework that links stimulus or action valuation to choice behavior (Rushworth and Behrens, 2008; Rangel and Hare, 2010). Commonly applied RL realizes goal-directed choices by (1) the continuous updating of value predictions of sensory features, (2) a softmax stochastic choice process among features that ensures performance accuracy while allowing for occasional exploratory choices, and (3) rapid learning from the consequences (outcomes) of selections using prediction error signals (Rushworth and Behrens, 2008). These processing components could likewise account for the efficient top-down control of attention and may thus provide a framework to understand the interplay of attentional subprocesses (Dayan et al., 2000; Roelfsema and van Ooyen, 2005; Wilson and Niv, 2011). We therefore devised a task for macaque monkeys that allowed testing whether commonly used RL frameworks help to understand how the learning of efficient attentional control is implemented and integrated during goal directed behavior.

We found that the learning of attentional stimulus selections in non-human primates closely followed a RL model that acts on representations of a restricted set of task relevant features, rather than on a representation of all stimulus and action items that could be linked to the decision outcome (Rangel & Clithero, 2014). However, we also show that a 'feature-based' RL model of attention needed to be supplemented with a value-independent stickiness process to account for non-randomly distributed errors during asymptotic behaviour.

2.3 Materials & Methods

Experiments were performed in two male macaque monkeys following guidelines of the Canadian Council of Animal Care policy on the use of laboratory animals and of the University of Western Ontario Council on Animal Care. Monkeys sat in a custom made primate chair viewing visual stimuli on a computer monitor (85 Hz refresh rate,

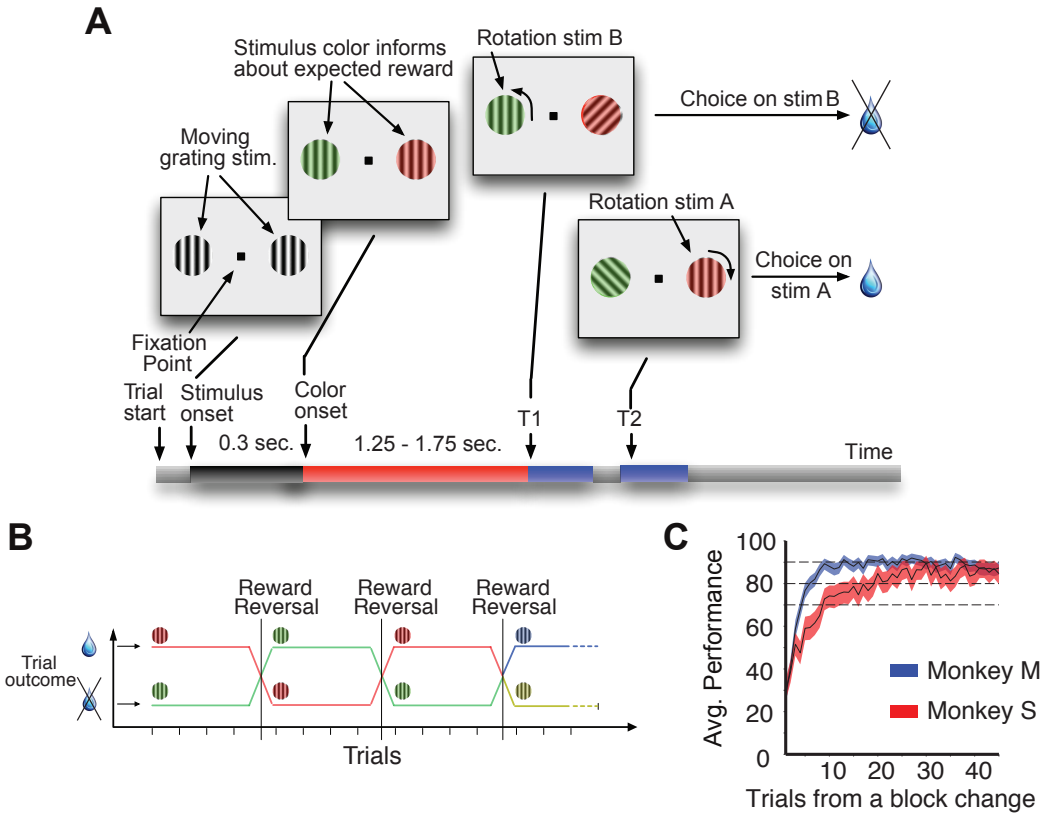


Figure 1. Feature based attentional learning task. (A) Uncued task design. Monkeys learned by practice that only the colour dimension of the stimuli was associated with reward, while others features (location, rotation direction or time onset of the rotation) were completely irrelevant. A proper allocation of covert attention allowed monkeys to successfully discriminate a transient rotation in the relevant stimulus while ignoring that of the distractor. Monkeys reported their response with an upwards vs downwards saccade according to the rotation direction, which was reversed in the two monkeys. (B) Colour-reward associations were changed in blocks of trials. (c) Average performance for monkeys M and S as a function of trial number in the block. The shaded area denotes the 95% confidence interval.

distance of 58 cm) in a sound attenuating isolation chamber (Crist Instrument Co., Inc.). The monitor covered 36° x 27° of visual angle at a resolution of 28.5 pixel/deg. Eye

positions were monitored using a video-based eye-tracking system (ISCAN, Woburn, US, sampling rate: 120 Hz) and were calibrated prior to each experiment to a 5 point fixation pattern. During the experiments eye fixation on a 0.2º gray square was controlled within a 1.4-2.5 degree radius window. Monitoring of eye positions, stimulus presentation, and reward delivery were controlled through MonkeyLogic (open-source software http://www.monkeylogic.net) running on a PC Pentium III (Asaad and Eskandar, 2008). Liquid reward was delivered by a custom made, air-compression controlled, mechanical valve system with a noise level during valve openings of ≤17 dB within the isolation chamber.

Task design.

We trained the monkeys on a feature-based reversal learning task (**Fig. 1A**). The task required monkeys to fixate and covertly attend to one of two peripherally presented stimuli. Stimuli had different colours and only one colour was associated with reward across trials within a block. To obtain reward the animals had to discriminate a transient rotation of the attended stimulus. Rotations also occurred in the stimulus with the non-reward associated colour. Monkeys indicated their choice by making a saccadic eye movement to one of two response targets presented 6.7º above or below the fixation point (clockwise/counter-clockwise rotations were mapped onto up-/downwards saccades for one monkey and onto down-/upwards saccades for the second monkey). In each block of trials, reward was associated only with one colour. No reward was given to rotation discriminations of the stimulus with the non-rewarded colour. Rotation direction (clockwise vs counter-clockwise), location (right vs left), and the time onset of rotation of the stimulus with rewarded and non-rewarded stimulus (first vs second vs

simultaneous) changed randomly across trials. In each trial, the stimulus with the rewarded colour and the stimulus with the non-rewarded colour rotated in opposite directions.

The event sequence in a trial was as follows (**Fig. 1A**). Monkeys initiated trials by directing and maintaining their gaze on a centrally presented, gray fixation point (on a black, 0.6 candela, background), followed 0.3 s later by the onset of two stimuli. Within the stimulus aperture, motion direction of a grating to the left from fixation was always to the upper left (-45º from vertically up) and motion direction of the stimulus on the right side from fixation was always to the right left (+45º from vertically up). After 0.4 s the stimuli were coloured. The rotation of the rewarded and non-rewarded stimulus occur either at 0.75 s or at 1.35 s. Trials in which the stimulus with the rewarded colour rotated before or after the stimulus with the non-rewarded colour were counterbalanced. In 10-50% (on average 30%) of all trials, the rotation of stimulus with the rewarded and non-rewarded colour occurred at the same time (1 s following the colour onset). Trials with rotations at the same time were introduced to validate that animals succeeded at selecting the relevant stimulus prior to discriminating the relevant rotation direction. Following stimulus rotation, animals made a saccadic response towards either of two target dots located vertically, above vs below, with respect to the fixation point, to report the rotation direction of the chosen stimulus. To obtain reward, a saccade had to be made 0.05-0.5 s following rotation onset of the stimulus associated with the rewarded colour. Animals received a fluid reward with a delay of 0.4 s following the saccadic response.

Within an experimental session, the colour-reward association was alternated in blocks of 60-100 trials, either maintaining the same pair of colours or by introduction of a new pair (**Fig. 1a**). After a minimum of 60 trials, a new block was introduced as soon as either of three performance criteria was achieved: i) running average performance (over 15 trials) of rewarded correct sensory-response associations exceeded 80%, ii) a total number of 60 rewarded trials; or iii) a total number of 100 trials independent on whether the choice was rewarded or not. Each experimental session also included shorter blocks of (n = 30) cued trials, which besides the cue instruction, had identical timing and stimulus events as the uncued trials described above. In cued trials, the fixation point was coloured to match the colour of one of the peripheral stimulus, which was indicative of that stimulus being relevant. Stimulus colours used in the cued trials were never used in the uncued trials. Cued trials were not analyzed in this report.

Stimuli.

We used square wave gratings with rounded-off edges for the peripheral stimuli (**Fig. 1A**), moving within a circular aperture at 1 deg/s, a spatial frequency of 1.4 Hz/deg and a radius of 2.2º. Gratings were presented at 6º eccentricity to the left and right of fixation. The grating on the left (right) side always moved within the aperture upwards at -45º (+45º) relative to vertical. The angle of rotation ranged between ± 13º and ± 19º. The rotation proceeded smoothly from the standard direction of motion towards maximum tilt within 60 ms, staying at maximum tilt for 235 ms, rotated back to the standard direction within 60 ms, and continued moving at their pre-changed direction of motion at -45º or +45º relative to vertical thereafter.

Performance analysis within a block.

Data analysis was done with custom written Matlab scripts (Mathworks Inc.). Analysis was performed on n=200 experimental sessions (n=100 sessions for monkey M and n=100 sessions from monkey S). To estimate how fast the animals learned a new colour-reward association when a new block started we calculated the number of trials needed to reach an average of 80% correct choices, with average behavior calculated by a moving Gaussian kernel with a sigma of five trials.

Logistic regression analysis.

We developed a logistic regression analysis over the complete set of trials under consideration to check whether reinforcement learning mechanisms are overall consistent with monkeys performance in the task, and if so, to infer specific reinforcement learning characteristics that we then used in the implementation of the reinforcement learning models.

In particular, we analyzed and ranked the predictive power for attentional selection of stimulus features. We tested four different versions of the regression analysis depending on how features in trial T predicted attentional selection of one of the two stimuli, and hence the choice, in the following trial T+1: version 1) features predict attentional selection of the stimulus they belong (trial T+1) if they formed part of the previously selected stimulus (inferred from choice in trial T) regardless of outcome (note that this is a control case in which the regression analysis is actually not consistent with reinforcement learning mechanisms); version 2) features predict attentional selection of the stimulus they belong if those features formed part of the previously selected stimulus and the trial (T) was rewarded (this case is compatible with reinforcement learning, such that positively correlated feature-reward associations are reinforced for

subsequent attentional selection); version 3) features predict attentional selection of the stimulus they belong if those features did not form part of the previously selected stimulus and the trial (T) was not rewarded; and version 4) that combines the previous two conditions, so features predict attentional selection of the stimulus they belong if they formed part of the previously selected stimulus and the trial (T) was rewarded, as well as if they did not form part of the previously selected stimulus and the trial was not rewarded.

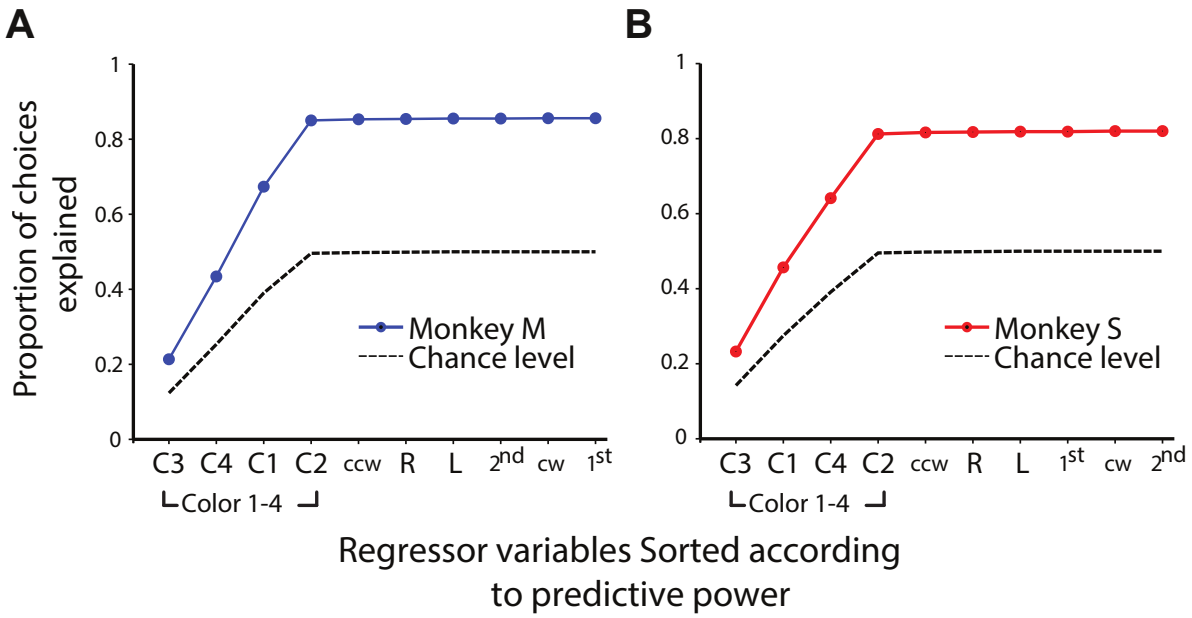


Figure 2. Logistic regression analysis of monkey performance in the task. We ranked the predictive power of each stimulus feature for subsequent performance according to a logistic regression analysis (see Materials & Methods for details). In both monkeys, the regression that included only colour information was the best predictor of next choices, which confirmed that non-human primates primarily utilize reward-associated stimulus features to guide their covert attentional selection. (A) Monkey M. (B) Monkey S. In both panels, labels 'C1, C2, C3, C4' in the figure denote the individual stimulus colours used; 'ccw', 'cw', denote stimulus rotation direction; 'R', 'L', denote stimulus locations, and '1st', '2nd', denote the relative time of movement onset of the rewarded stimulus in relation to that of the distractor.

Interestingly, results from this latter condition was the best predictive of monkey choices (**Fig. 2**, other conditions not shown), which suggests a value-update

generalization of the features in the two stimuli, even when monkeys in each trial only acted on one of the two stimuli.

We computed the ranking of features on half of the sessions (odd session numbers) and validated this ranking on the other half of the sessions (even session numbers). Figure 2 shows the proportion of choices explained with respect to chance level from a collection of regression analyses, in which each analysis included one more regressor than the previous, beginning with the regressor with the largest predictive power, following according to the predictive power ranking, until all regressors were taken into account. These results confirmed that colours were the best predictors of next choices, supporting the hypothesis of value-based covert attentional selection guiding monkeys' behavior.

Due to the two-alternative choice, the chance level was computed as 50% of the trials in which at least one of the features in the feature set formed part of the stimulus associated with reward (by task design) in trial T and present in trial T+1 (hence predictive of choice; see above). For instance for monkey M, the first regressor (colour C3) correctly predicted 21.35% of the next choices. This represented 86.44% of the trials in which the colour C3 determined the stimulus associated with reward by task design (24.7% of the whole set of trials, far beyond the chance level at 12.35%). Importantly, the proportion of trials explained initially grew at a similar pace while including colours in the regression analysis, but then drastically stopped, showing that including other features did not improve further the predictive power of monkey choices (**Fig. 2**). Note also the increasing separation with respect to the chance level when incorporating colour features to the regression analysis, until it reached the maximum

when all colours were added. This separation remained the same even though other features were included in the regression analysis (**Fig. 2**).

R**einforcement learning modelling.**

To model monkeys' behavior and the processes related to covert attentional selection we used Rescorla-Wagner type reinforcement learning (RL) employing standard Q-Learning and Boltzman softmax selection algorithms (Glimcher, 2011). We initially compared two distinct value based reinforcement learning models that differed in whether a restricted, optimal internal representation of the task was prioritized or not.

In order to explain a specific pattern of error trials shown by the monkeys, which was not reproduced by value-based models, we explored additional non-value based mechanisms. First, we accounted for an influence of selection perseveration that is unaffected by values, which has been previously shown to improve action selection (Lau and Glimcher, 2005). This Value-History Model (**Fig. 6A**) transforms feature values into probabilities of attentional selection just as the feature-based RL does, but it then incorporates a weighted bias towards whatever feature was selected on the previous trial.

The second extension of feature-based RL, the Hierarchical Value-History model, is similar to the previous Value-History Model, but in this formulation the value-based selection process is concatenated with a subsequent final attentional selection between the selected feature in the previous trial vs the current trial value-based selected feature (**Fig. 6B**). This sequential selection can therefore be conceived of as a hierarchical two-step decision process.

Third, we quantified the influence of a mechanism that dynamically adjusted the exploration vs exploitation trade-off based on performance. This Adaptive Selection Model incorporated a meta-learning parameter that scaled up or down the non-linearity in the transformation from value to probability of attentional selection according to reward outcome (**Fig. 6C**). Thus, when model performance is low, typically at the beginning of a block, more exploratory behavior is produced due to a low β value, since it increases the stochasticity of selection among features. As rewarded outcomes become more frequent, β increases, which makes attentional selection more deterministic.

In a fourth model extension we incorporated non-value based noise into the attentional selection process (**Fig. 6D**). In this Intrinsic Noise Model, such noise is evenly distributed among all stimulus features. Thus, there is no dependence on value, reward, or selection-history in this module of the model, but rather an explicit influence of noise, intrinsic to the transformation of value-based selections to motor commands due to influences, such as decreased motivation, imperfect sensory-motor mappings, or selection biases, among others, under the assumption that these influences do not show a preference for specific features in the internal model representation of the task.

RL model algorithms.

In its basic form, the value of any predictor of reward (Q_i) is updated on the next time step (trial) from its previous value through the scaled reward-prediction error: the difference between the binary reward outcome (R, either 0 or 1) and the predictor itself. The scaling factor (α) represents the learning rate:

$$Q_i(n+1) = Q_i(n) + \alpha[1 - R(n) - Q_i(n)] \quad \text{(eq. 1)}$$

We implemented reinforcement learning models that assumed value generalization. Thus, all stimulus features associated with the selected stimulus updated their value according to equation 1. Stimulus features associated with the other stimulus were updated according to:

$$Q_i(n+1) = Q_i(n) + \alpha[R(n) - Q_i(n)] \quad \text{(eq. 2)}$$

Our RL approach assumed that performance in a trial only depended on a correct covert attentional selection of the relevant stimulus, which implied an infallible rotation discrimination and its associated saccadic response.

The *feature-based RL* took only the systematically relevant colour dimension into account as predictor of attentional selection, and therefore of reward (**Fig. 3A**). In contrast, in the *non-selective RL*, all stimulus features (colours, locations, rotation directions and time onsets of the rotation) were considered potential predictors of reward (**Fig. 3B**).

The final attentional selection of one stimulus against the other obeyed a covert, value-based softmax decision-making process acting on the feature space, in particular upon non-linearly transformed values that represented the probabilities of selecting different stimulus features, according to the Boltzmann equation:

$$P_i(n) = \frac{e^{\beta Q_i(n)}}{\sum_j e^{\beta Q_j(n)}} \quad \text{(eq. 3)}$$

where β represents the inverse temperature and establishes the strength of the non-linearity. The two RL models thus included two free parameters (α and β) that we optimized to best predict monkey behavior on a trial by trial basis (**Fig. 3**).

Value-History Model. The first extension of the feature-based formulation introduces an explicit factor that influences the value based selection mechanism by biasing the selection towards the feature that was selected previously, irrespective of whether it was rewarded or not (Fig. 6A). The selection of this Value-History Model is formally implemented as:

$$P_i(n) = \frac{e^{\beta Q_i(n)} + e^{\gamma \delta_{ik}} - 1}{\sum_j e^{\beta Q_j(n)} + e^{\gamma \delta_{jk}} - 1} \qquad \text{(eq. 4)}$$

where in the γ term k represents the previously selected feature and appears inside a Kronecker delta function, which takes a value of 1 if i is equal to k, or 0 otherwise. The term -1 is included to remove any impact of the γ term when γ is 0. The effect of γ can be described as an increase in the probability to reselect the immediate previous selection, which in principle might be beneficial to diminish the impact of noise in the value system implementation, at the cost of a reduced celerity in the adaptation to changed feature-reward contingencies.

Hierarchical Value-History Model. As indicated above, the second extension of the model-based RL is similar to the Value-History Model, but in this formulation the selection process based on values is concatenated with a subsequent selection between the feature choice of the previous trial and the current trial value-based selected feature (**Fig. 6B**). This sequential selection can be conceived of as a

hierarchical two-step decision process. The first process fully corresponds to the model-based selection process defined in equation 3. From this selection, feature k is selected with 'confidence' P_k dictated by the softmax function, and used in a second step to compete with the previously selected feature P_l (if feature l is different than feature k):

$$P_k^{'} = \frac{P_k}{P_k + P_l} \qquad \text{(eq. 5)}$$

vs.

$$P_l^{'} = \frac{P_l}{P_k + P_l} \qquad \text{(eq. 6)}$$

where, $P_l = e^{\gamma} - 1$. When the value-based selected feature k and the previously selected feature l are the same, both terms add together and the probability to select the feature trivially collapses to 1.

Adaptive Selection Model. The third extension of the feature-based formulation introduces a mechanism that adjusts the probabilistic nature of the value-based selection process to either trigger more exploratory selections, or to more deterministically follow the valuation mechanism (corresponding to an exploitation regime with high confidence, see **Fig. 6C**). The selection of values in the *Adaptive Selection Model* uses equation 3, but with the difference that β is not a constant, but instead obeys an equation similar to the Q-values (equations 1 and 2; note that R is a binary teaching signal and then only one of the two terms in equation 7 is different than zero in each trial):

$$\beta(n+1) = \beta(n) + R(n)\mu[\beta_H - \beta(n)] - [1 - R(n)]\mu\beta(n) \qquad \text{(eq. 7)}$$

where μ is the rate of change of β. β values are bounded between 0 and β_H. β tends to either one or the other depending on the outcome (R). If the outcome is 1, then β grows towards β_H, otherwise it decreases to 0. Thus, following positive outcomes, the impact of β is to make the softmax function (P_i above) more similar to a winner-take-all, but to otherwise encourage more exploratory behavior. Therefore, this model becomes more or less confident on the value system depending on outcome evaluation.

Intrinsic Noise Model. The fourth extension assumes that part of behavioral variability is in principle not explainable by value-based updating and selection mechanisms, but rather is due to random behavioral variability, and hence evenly distributed among features (Fig. 6D):

$$P_i(n) = \frac{P_R}{N_F} + (1 - P_R)\frac{e^{\beta Q_i(n)} - 1}{\sum_j e^{\beta Q_j(n)} - 1} \quad \text{(eq. 8)}$$

The term P_R denotes the random behavioral probability that is evenly distributed among task features (N_F refers to the number of those). Note that -1 is introduced to remove the contribution of Q-values that are equal to 0. The value system is scaled down by the factor 1-P_R. This random weighting factor could theoretically fit the data better compared to the pure value based model if the noise significantly splits into two parts: one noise component in the softmax (among Q-values that are not strictly 0), and another noise component that is non-value based. This is because a single β parameter in principle does not necessarily capture the two sources of noise at once, but instead is designed to capture value-based stochasticity. This *Intrinsic Noise Model* is similar to

the *Value-History Model* by adding a non-value based process that competes with value, but for the *Intrinsic Noise Model* the non-value based process operates at random among features instead of favouring the previous attentional selection.

Model evaluation and optimization.

Three independent criteria (outlined in detail below) were combined to evaluate RL models. Such a multi-score evaluation was critical to (1) account for the dynamics of learning of monkeys in the task (Performance-SSD), (2) analyze the plausibility of a RL mechanism for explaining monkey performance (Mechanism-SSD), and (3) maximize the total number of trials in which monkeys and model performance matched, corrected to penalize model biases against the least frequent outcome (i.e. the overall proportion of trials explained was corrected by subtracting the highest between the proportion of false positives and false negatives).

The first score represented the sum of square differences between the block-averaged performance of the model with respect to the monkey over the same blocks of trials (**Fig. 4A** and **Fig. 7A,C**). The second score quantified the extent to which the RL mechanism employed by a model was compatible with monkeys' behavior (**Fig. 4B** and **Fig. 7B,D**). The average model performance only depends on the probability to select the relevant stimulus, and a direct test of this mechanism can be applied to monkeys' behavior: we binned the probability to select the relevant stimulus and computed for them the averaged monkeys' performance as well as its 95% confidence intervals. If the averaged performance of the monkeys were largely different from the probability to select the relevant stimulus according to the model, we can then conclude that such mechanism would not be fully compatible with monkeys' behavior, and this is visualized

by deviations from the diagonal in Figure 4B and Figure 7B,D. After calculating our measure of Performance SSD and Mechanism SSD, we normalized these scores across all models and parameter sets, but independently in each monkey, to ensure that all scores were bounded in the same range [0,1].

The third measure evaluating the model performance compared the outcome experienced by the monkey on every trial to that of the model and calculate the total proportion of correctly matching trials. The common denotation of this measure is proportion of total explained trials. We modified this score to correct for the fact that it is important for a model to not only predict a high proportion of trials correctly but it must ideally predict the correct proportion of rewarded and unrewarded trials, avoiding any potential bias. For example, a toy model that merely predicts a rewarded choice on every trial would provide no insight into the mechanisms driving monkey behavior, but would report a total proportion explained >80% due to the overall high proportion of rewarded behavior shown by the monkeys. We then corrected the total explained score by subtracting the proportion of false positives or false negatives (whichever was higher) to provide a single score that combines both raw explanatory power and a measure of predictive accuracy.

The score appeared inverted (i.e. 1-score, corrected proportion of unexplained trials, so a lower score reflected a better model performance) to be in agreement with the two previously described scores. Each model was optimized by a grid search across the total parameter space and through cross-validation between odd and even numbered sessions. Model performance was first assessed using odd numbered sessions of monkey data by calculating the mean score for each parameter set across the three

different measures, with each score representing the mean of ten model replications to diminish the impact of fluctuations due to the stochasticity in the model. Then, best aggregate scores for each model computed on odd numbered sessions were used to assess model performance on even numbered sessions. Cross-validation of scores confirmed that parameters were not fit to non-systematic behavior (e.g. which would have followed from over-fitting), but instead represented a generalizable version of the model.

Analysis of error patterns.

Consecutive unrewarded trials during asymptotic performance (towards the end of a block, after the learning period; **Fig. 4A**) were unlikely events in feature-based and non-selective RL systems (**Fig. 5**), because feature values were very dissimilar and changed only minimally at asymptote, so errors were only due to the stochasticity of the selection process under such conditions. This suggests a random, and independent distribution of errors during this period, which would be expected to happen also in monkey behavior if following directives of the feature-based or non-selective RL system.

To test this null hypothesis we counted all errors made during asymptotic behavior in a blockwise fashion (typically from trial 15-20 after the block change to the end of the block) and calculated the proportion of errors occurring in sequences of increasing length. To compare to a random distribution, we subtracted the proportion of errors for each error sequence from the theoretical proportion given by a random distribution. This transformation eased the identification of clusters of errors (i.e. unrewarded trials made consecutively), which occurred in monkeys more frequently than predicted by the

stochasticity of learned values according to RL models (**Fig. 5**). This finding suggested an additional selection mechanism influenced by non-value based sources (**Fig. 6**).

2.4 Results

We devised a reversal learning task for macaques that isolates the covert attentional selection of relevant sensory information from the perceptual discrimination and action planning processes directly involved in overt decision making (**Fig. 1A**). Covert attention was required to select one of two peripherally (left and right) presented stimuli for prioritized processing. Overt decision making was required to obtain reward through discriminating a transient (clockwise/counterclockwise) rotation of the stimuli by making an (up/downward) saccadic eye movement. Monkeys were rewarded only if the decision about the rotation was performed on one of the two stimuli with no reward given if the animal acted on the alternative stimulus. The rewarded stimulus was defined by its colour with the reward-associated colour changing between blocks of trials. The task design ensured that the stimulus colour varied independently from (1) the stimulus location (right or left), (2) the decision variable of the overt choice (clockwise or counterclockwise rotation) that eventually provided the outcome, (3) the action plan (up or downward saccades) used to indicate overt choice, and (4) the three possible time points at which the stimulus rotation could occur. Thus, in contrast to previous learning paradigms in nonhuman primates (Sugrue et al., 2004; Lau and Glimcher, 2008), the actual reward associated feature (colour) was independent of action, location and timing. A related study using a similar task, but with an cue indicating the relevant stimulus, shows that the monkeys shift attention in response to stimulus color, with the only difference in our study being the the monkey's need to learn the relevant color for

reward (Kaping, et. al, 2011). The task enforced learning reward predictions about specific colours by changing the reward associated colour after a performance criteria or a maximum number of trials was reached in a block of trials with constant colour-reward association (see **Fig. 1B** and **Materials & Methods**).

Monkeys successfully use feature values to guide attention.

Both monkeys were successful in 82.5% of trials (monkey M = 84.7% out of 84,417 trials; monkey S = 80.3% out of 86,689 trials). Within blocks, monkeys required on average 12.5 trials to locally reach a performance level of 80% rewarded trials (Monkey M: mean 8.5 trials, SEM ± 0.37 trials, Monkey S: mean 16.5 trials, SEM ± 0.84 trials; see also Materials & Methods for details of how local performance was computed). This performance level was stable across experimental sessions as the monkeys had learned the task structure during behavioral training sessions, which are not included in the analysis. Asymptotic performance measured across trials following initial learning was on average 87.3% correct (Monkey M: 89.1 ± 0.02%, S: 85.5 ± 0.02%; **Fig. 1C**).

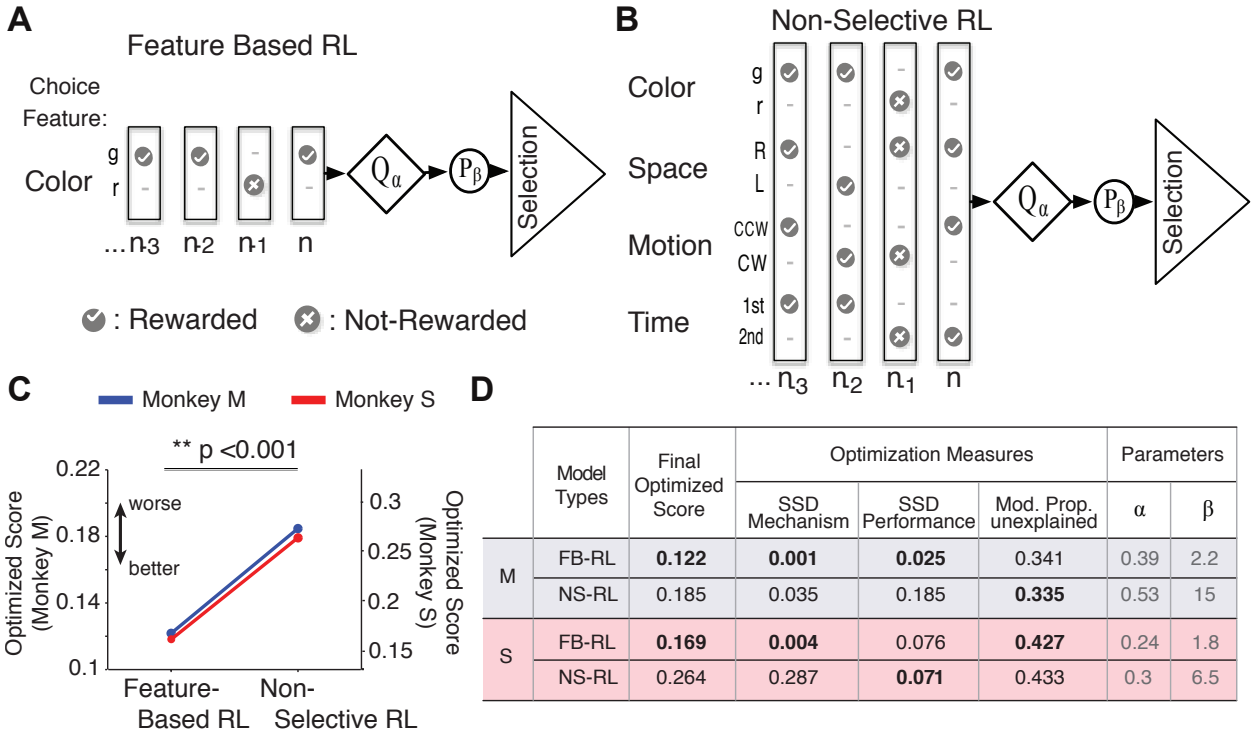
By task design, reversal blocks, where stimulus colour was maintained but the colour-reward association was reversed from the previous block, present the animal with a more difficult learning problem than other blocks where a new colour pair is presented. Reversal blocks represent 34 & 39% of total blocks for monkeys ‘M’ and ’S’ respectively, and average performance overall in these blocks is not significantly different from other blocks for monkey ‘M’ (83.2 & 83.6 % correct choices in Rev. vs Other blocks respectively, $p>.05$ Mann–Whitney–Wilcoxon), and is slightly worse in reversal blocks for monkey ’S’ (79.0 & 81.4% correct choices in Rev. vs. Other blocks,

p<.05 Mann–Whitney–Wilcoxon). The biggest difference in performance in reversal blocks versus other blocks is seen in the dynamics of choice behaviour following the block switch (**Appendix D - Chapter 2 Supplementary figures**). Both monkeys are slower to learn the correct colour-reward association, but the learning rate is different between them, with monkey 'S' taking on average nine trials in reversal blocks to match performance in other blocks, while monkey 'M' takes on average only four. For monkey 'S' the mean proportion of correct choices on trials early in a reversal block is significantly lower than in other blocks by 13.7, 8.9, 11.9, 11.6, 7.5, 6.6, 7.0, 4.7, & 2.1 ± 1.5 SE% on trials 1-9 respectively (p<.05 Mann–Whitney–Wilcoxon). For monkey 'M' the mean proportion of correct choices on trials 1-4 in a reversal block is lower than in other blocks by 16.6, 17, 11.0, & 2.0 ±.8 % respectively. The difficulty in learning the correct stimulus-feature response mapping in reversal blocks compared to other blocks is expected given the conflict with the previously learned association and the fact that the block change is uncued. However, both monkeys are still able to rapidly adapt to the new reward-association suggesting that their attentional learning processes are flexible.

To validate quantitatively that the colour dimension was the only feature used by animals to perform the task, we used a logistic regression analysis (see details in **Materials & Methods**). Sorting task features according to their subsequent predictive power for reward outcome through individual trials confirmed that stimulus colours were maximally explanatory of monkeys' behavior, while non-colour features had no systematic influence on the performance (**Fig. 2**).

Evidence for an optimal internal representation in the learning of feature values.

Having shown that animals were able to link choice outcomes (reward obtained from up and downward saccades) to the feature that determined attentional selection, three questions arise: Firstly, is there an optimal internal representation used to solve



	Model Types	Final Optimized Score	Optimization Measures			Parameters	
			SSD Mechanism	SSD Performance	Mod. Prop. unexplained	α	β
M	FB-RL	**0.122**	**0.001**	**0.025**	0.341	0.39	2.2
	NS-RL	0.185	0.035	0.185	**0.335**	0.53	15
S	FB-RL	**0.169**	**0.004**	0.076	**0.427**	0.24	1.8
	NS-RL	0.264	0.287	**0.071**	0.433	0.3	6.5

Figure 3. Reinforcement learning (RL) model schemes and results. A) Feature-based RL tracks across trials (..., n-1, n) reward-dependent values (Q) only for the relevant stimulus features (colours, in the example: g = green, r = red) and non-linearly transforms them into choice probabilities (P) of attentional selection through a softmax function. B) Non-selective RL works the same but tracks values for all task features, denoted as Colour, Space (stimulus location), Motion (rotation direction) and Time (order of each stimulus rotation onset). C) The performance of the optimized feature-based RL is better than the performance of the optimized non-selective RL for both monkeys (left x-axis: monkey M, right x-axis: monkey S). D) Optimization scores and parameters for feature based (FB) RL and non-selective (NS) RL for monkey M (blue shaded) and monkey S (red shaded). Multiple scores were used to explicitly account for different aspects of monkeys' behavior and to directly test the predictive power of each RL mechanism (see text and Materials & Methods for details). Lower scores denote better model prediction. Bold font highlights best scores relative to the alternative model. SSD denotes normalized sum of squared differences in [0,1].

the task? Secondly, how are internal representations of feature values updated after experiencing outcomes? Thirdly, is covert attentional selection fully described according

to value based mechanisms or are there other non-value based influences that also affect attentional performance systematically?

To identify the computational processes most likely controlling attentional selection, we devised and compared two Rescorla-Wagner type reinforcement learning models (Glimcher, 2011)(**Fig. 3A,B**). We describe one model, with a task set restricted to the relevant feature dimension (colour) as feature-based RL, because it contains an internal model representation of only the relevant task features (**Fig. 3A**). We contrasted this model with non-selective RL, which did not include any prior knowledge about which of the available decision variables were systematically linked to reward, but rather relied on tracking values for all stimulus features that were available, including not only stimulus colour, but also location, rotation direction, and the time onset of rotations (**Fig. 3B**).

Three independent criteria were combined to evaluate RL models in order (1) to account for the dynamics of learning in the task, (2) to analyze the plausibility of a RL mechanism for explaining monkey performance, and (3) to maximize the total number of trials in which monkey and model performance matched, corrected to penalize model biases against the least frequent outcome (see **Materials & Methods**). The direct comparison of RL models according to this evaluation revealed that the feature-based RL outperformed the non-selective RL in predicting covert attentional selection, evident in a significantly better (lower) optimized compound score of model performance (Fig. 3C, feature-based vs non-selective RL, comparison across 10 model realizations: Monkey M: $p < 0.005$; monkey S: $p < 0.001$, Mann–Whitney–Wilcoxon test). The most prominent difference between models was that the stochastic selection process was

considerably more deterministic (higher beta value) in the optimized non-selective RL model compared to the feature-based RL model (**Fig. 3D**).

Despite the overall superiority of the feature-based RL model, the two models were indistinguishable in predicting the dynamics of learning within a block as inferred from the average monkey performance, and both models explained a similar proportion of animals' covert attentional selections in single trials (feature-based RL: monkey M / S:

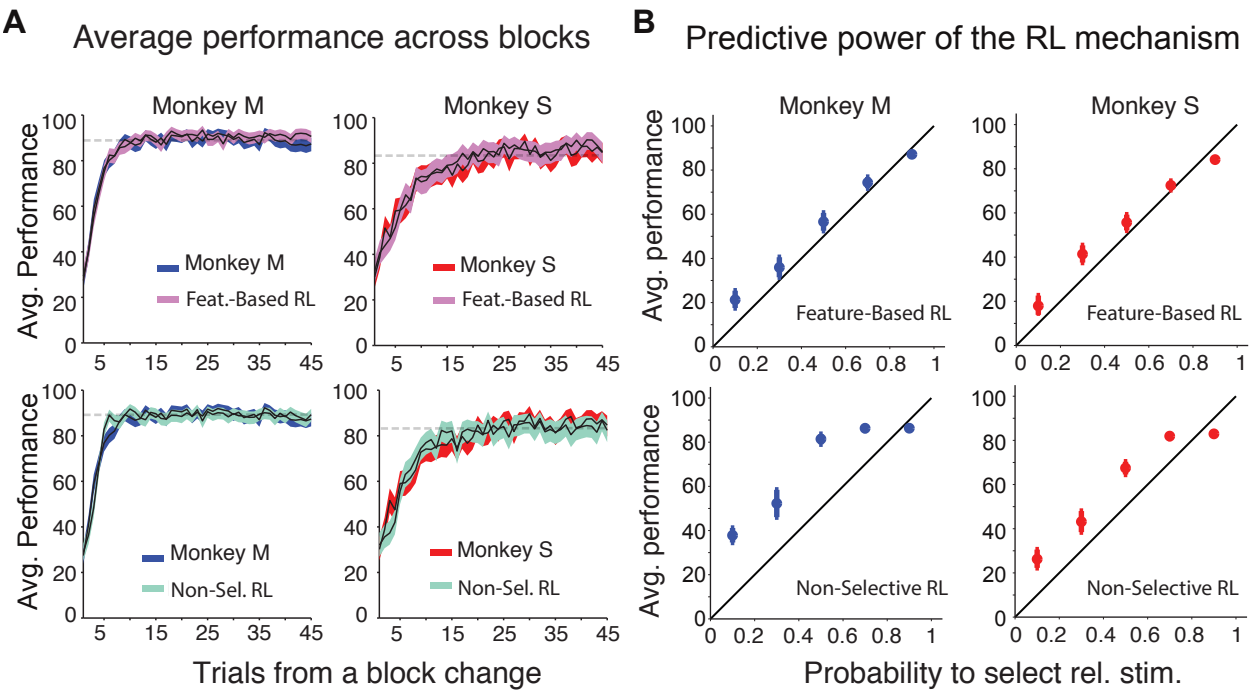


Figure 4. Performance of feature-based vs non-selective RL systems. (A) Average performance and its 95% confidence interval (shaded area) as a function of trial order within a block predicted from feature-based RL (top row) and non-selective RL (bottom row) for monkey M (left) and S (right). The normalized sum of squared differences between the performance of the monkey and models served as one model evaluation criterion, labeled SSD Performance in Figure 3D. (B) Monkey averaged performance and its 95% confidence interval (error bars) against the likelihood to select the relevant stimulus according to models. The panels show the performance of the monkeys (y-axis) corresponding to five bins that fully span the range of the probability to select the relevant stimulus. A plausible model candidate requires the model's likelihood and monkey's performance to match each other. The degree to which this happens is quantified by the normalized sum of squared differences (labeled SSD Mechanism in Fig. 3D). The panels are arranged as in (A): feature-based RL (top row), non-selective RL (bottom row), monkey M (left column), and monkey S (right column).

78.2% / 72.2%; non-selective RL: monkey M / S: 78.5% / 71.5%) (**Fig. 4A**).

The failure of non-selective RL became evident only when we compared the output of the stochastic selection process of the RL models to the selections made by the monkeys. Figure 4B illustrates that the probability to select the relevant stimulus according to the feature-based RL closely followed likelihood of correct choices made by the monkey (**Fig. 4B** upper row, the diagonal line represents a perfect match). In contrast, monkey choice likelihood deviated from the probability dictated by the non-selective RL model (**Fig. 4B**, bottom row). This result supports the suggestion that choices of the monkeys depend on prior covert attentional selection that operates on an internal representation of task relevant feature space.

Erroneous choices reveal non-value based selection biases of monkeys.

The previous analysis showed that the probability of correct attentional selections by the feature-based RL closely resembled the likelihood of correct overt choices of the monkeys on a trial-by-trial basis. This mechanism did not, however, explain why monkeys were committing non-randomly distributed errors during asymptotic performance, i.e. after they apparently had learned the reward predicting colour. The asymptotic performance of the animals corresponds in the models to the period in which feature values are close to saturation, a regime that began on average 8-16 trials after the block start (measured as the trial number needed to reach >80% performance; Monkey M: mean 8.5 trials, SEM ± 0.37 trials, Monkey S: mean 16.5 trials, SEM ± 0.84 trials).

To identify the source of these errors, we analyzed sequences of choices while at asymptotic performance and found that erroneous choices clustered together more often than expected by the performance of the feature-based RL model. Among all

erroneous choices at peak behavior (M: 10.9%, S: 14.5%), consecutive errors made up 40.8% of errors for monkey M and 34.3% for monkey S (the proportional error patterns for monkey M (S): 59.2% (65.7%) for CEC (correct-error-correct) successions, 24.2% (19.4%) for CEEC, 9.5% (9.0%) for CEEEC, etc.). Figure 5 illustrates how this error pattern deviated from a random distribution, revealing that both monkeys committed less errors in isolation and more errors in succession than expected for a stochastic

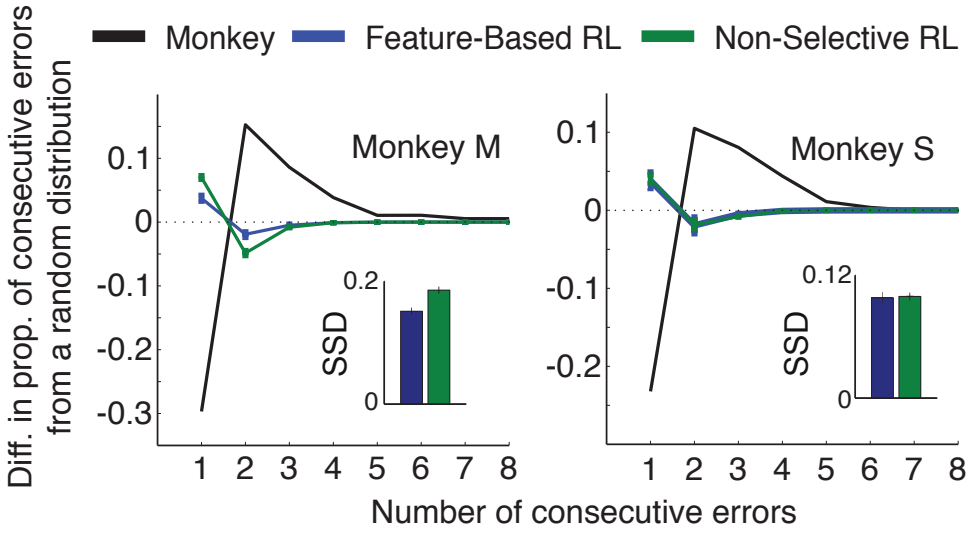


Figure 5. Failure of feature-based and non-selective RL systems to account for the pattern of consecutive errors shown by the animals during periods of asymptotic performance. The panel shows how the proportion of consecutive errors (x-axis) by monkeys (M: left, S: right) deviated from what would be expected if errors were generated by a random process (dotted line). Feature-based RL (in blue) and non-selective RL (in green) failed to capture this error pattern. The inset bar panels show the sum of squared differences (SSD) between the error pattern of monkeys and models. Errors represent SEM.

error generating process. As might be expected given its stochastic selection mechanism, the feature-based RL system (and also the non-selective RL system) failed to capture this error pattern, both generating a pattern of errors close to random (**Fig. 5**).

Value based attentional selection is weighted by non-value based selection biases.

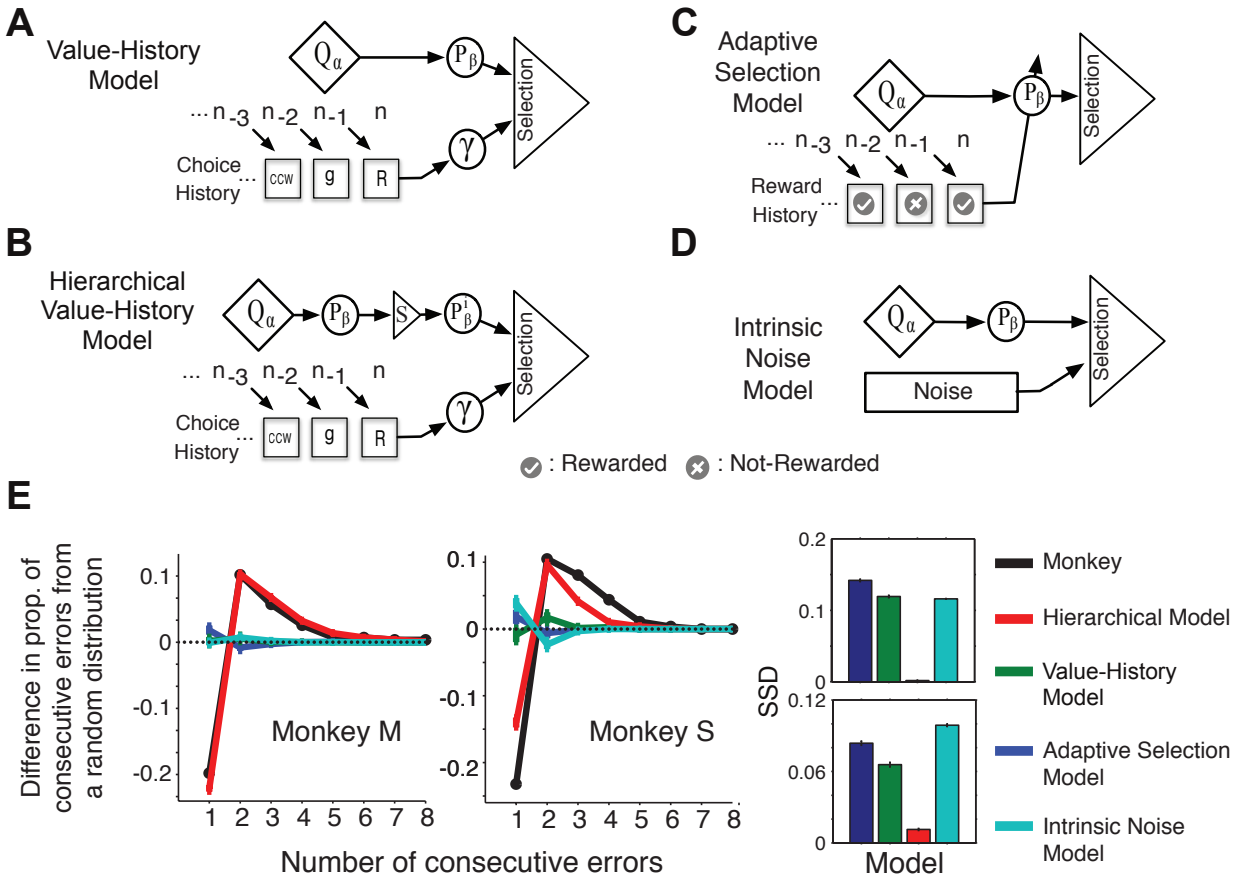


Figure 6. Schemes of extended feature-based RL systems and results from the analysis of consecutive errors in them. A) The Value History Model incorporates in the covert attentional selection an influence towards previous choice. B) The Hierarchical Value-History Model first makes a value based selection (P_I) and then weights it by the previous selection in a second selection step. C) For the Adaptive Selection Model the transformation of Q-values to choice probabilities is dynamically shifted by the reward history. D) The Intrinsic Noise Model assumes that part of monkeys performance stochasticity is independent of the value-based influences and distributes evenly among stimulus features. E) Distribution of consecutive errors for monkey M and S, for each of the extended models shown in (A-D). The left two panels show how the proportion of consecutive errors by monkeys (M: left, S: center) deviated from the proportion of errors that would be expected if errors were generated by a random process (as in Fig. 5). Only the Hierarchical model captured this property of animals' behavior. The bar plots (right) quantify the sum of squared differences (SSD) between the error pattern of monkeys (M: top, S: bottom) and models. Errors represent SEM.

The failure to account for the observed error pattern shows that feature-based RL must be complemented by additional mechanisms in order to explain the animal's

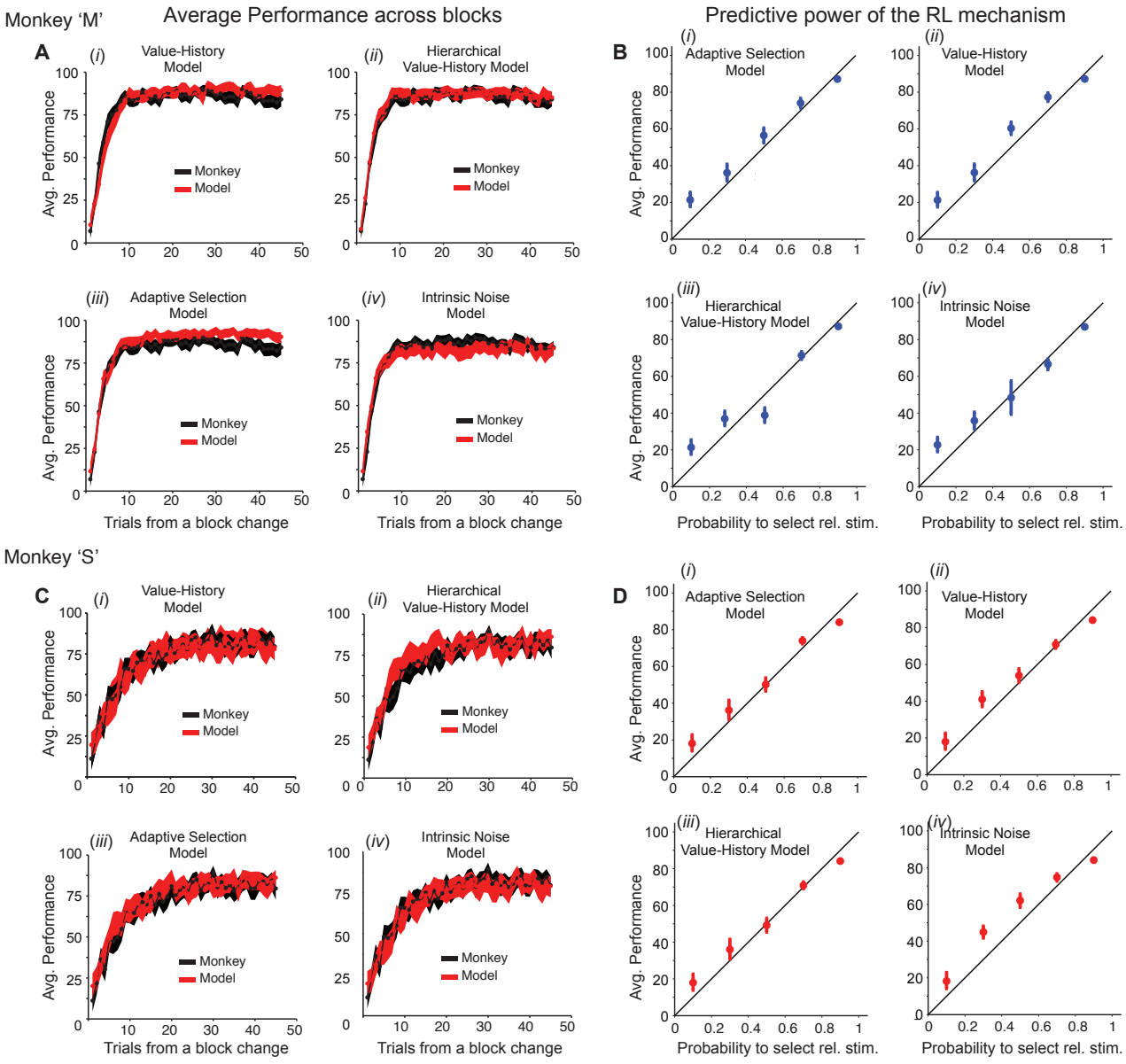


Figure 7. Average performance of monkeys and the four models that extend feature-based RL. A,B) Results from monkey M for (i) the Value-History Model, (ii) the Hierarchical Value-History Model, (iii) the Adaptive Selection Model, and (iv) the Intrinsic Noise Model, respectively. In (Ai-iv) the black shaded area shows the 95% confidence interval around the mean for the monkey, in red for each model. Panels in Bi-iv show the averaged performance and 95% confidence interval (error bars) of monkey 'M' (y-axis) corresponding to five bins that fully span the range of the probability to select the relevant stimulus. A plausible model candidate requires the model's likelihood and monkey's performance to match each other. The degree to what this happens is quantified by the sum of squared differences (SSD). C,D) Same as (A,B), for monkey S.

attentional performance pattern. We thus extended the feature-based RL system and devised four additional models, each with a distinct mechanism for explaining behavior (see **Materials & Methods**). In particular, we tested the influence of (1) a direct effect of value independent selection history onto feature specific value representations (*Value-History Model*, **Fig. 6A**), (2) a hierarchical 2-step selection process that incorporates an initial value based feature selection as well as a subsequent value independent input from selection history (*Hierarchical Value-History Model*, **Fig. 6B**), (3) a dynamic regulation of the selection stochasticity based on recent reward history (*Adaptive*

	Model Types	Final Optimized Score	Optimization Measures			Optimized Parameters					
			SSD Mechanism	SSD Performance	Corr. % unexplained	α	β	β_H	γ	P_R	μ
M	Adaptive	0.116	0.011	0.016	0.320	0.3	-	3	-	-	0.06
	Hierarchical	0.134	0.001	0.065	0.341	0.5	1.6	-	0.5	-	-
	Value-History	0.158	0.038	0.104	0.333	0.26	2.4	-	0.5	-	-
	Intrinsic Noise	0.236	0.005	0.337	0.367	0.36	1.7	-	-	0.27	-
S	Hierarchical	0.172	0.019	0.062	0.436	0.28	1.5	-	0.35	-	-
	Adaptive	0.177	0.03	0.076	0.426	0.23	-	2.4	-	-	0.12
	Value-History	0.210	0.122	0.195	0.419	0.2	1.9	-	0.4	-	-
	Intrinsic Noise	0.25	0.249	0.089	0.42	0.16	1.95	-	-	0.3	-

Table 1. The optimization scores and optimized parameters for extended models: the Adaptive Selection Model, the Hierarchical Value-History Model, the Value-History Model, and the Intrinsic Noise Model (see main text and Fig. 6). Monkey M and S results are shown in red and blue shaded cells, respectively.

Selection Model, **Fig. 6C**), and (4) in the last model we tested the influence of added noise to the system that is evenly distributed among choice features (*Intrinsic Noise Model*, **Fig. 6D**).

We optimized each of the four extended RL models separately using the compound criteria of model performance as before. Across models, the *Hierarchical Value-History Model* and the *Adaptive Selection Model* performed best (**Fig. 7**, **Table 1**).

However, given that these models used additional parameters than the basic feature-based RL (**Fig. 3A**), the improvement in explaining correct choices was at most marginal. However, in contrast to this marginal effect with respect to predicting correct choices, the prediction of erroneous choices separate model performance.In particular, predicting the pattern of consecutive errors revealed a clear advantage of the *Hierarchical Value-History Model* against feature-based RL and each of the three remaining models in both monkeys (**Fig. 6E**). Thus, the Hierarchical Value-History Model closely predicted the error patterns evident in the two monkeys. It predicted the monkey's error patterns significantly better than the Value-History Model (for monkey M (S): $p < 0.001$ (0.001) Kruskal-Wallis test), the Adaptive Selection Model (for monkey M (S): $p < 0.001$ (0.001) Kruskal-Wallis test), and the Intrinsic Noise Model (for monkey M (S): $p < 0.001$ (0.001) Kruskal-Wallis test). It is noteworthy that the prediction of error patterns was not an explicit criteria during optimization, but emerged from the sequential (2-step) selection mechanism intrinsic in the Hierarchical Value-History Model (Ahn et al., 2008; Donoso et al., 2014).

Analysis of Value Independent Selection Biases.

Biases for stimulus features indicate limits for RL model predictions. Even though our intrinsic noise model failed to improve the predictive abilities of the model-based RL framework, the distribution of errors observed in both monkeys at asymptotic behavior (**Fig. 5** and **6E**) suggests that there is some non-value based influence in behavior,

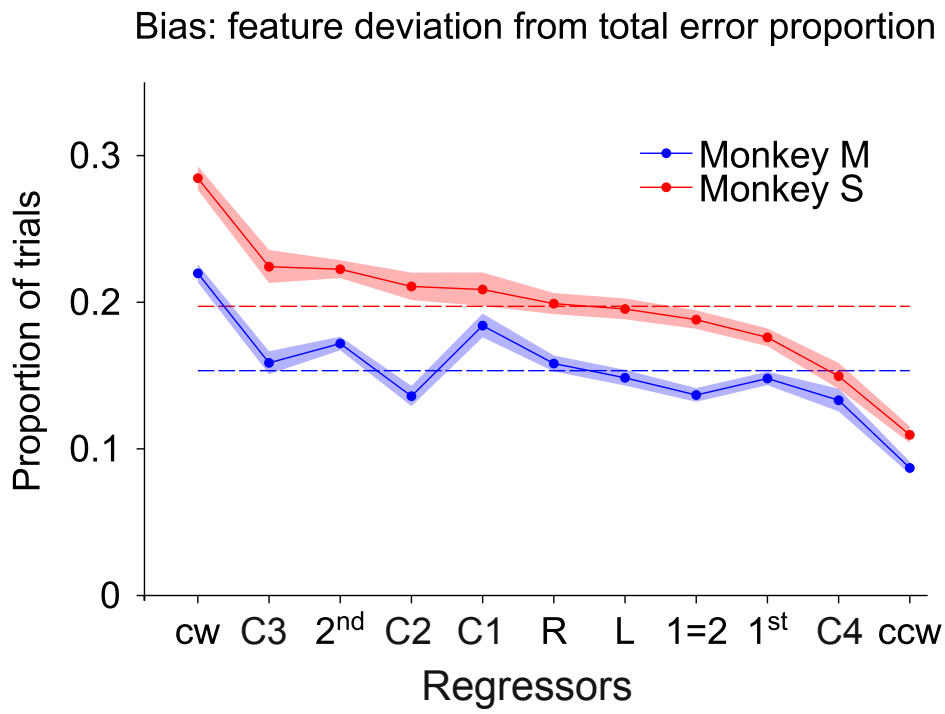


Figure 8. Non-value based feature biases measured as the proportion of errors associated with each particular stimulus feature. Monkey S (red dashed line) presents an overall larger proportion of errors than Monkey M (red dashed line), and this pattern is systematic for each feature pair (solid lines), In principle, this could be due to non-value based feature biases or explained by exploratory behavior. Given that colours are the features that predict attentional selection and eventually the behavioral choice (Fig. 2 and Fig. 3), it might be expected that exploratory behavior would mainly stay within this dimension. However, we see that the proportion of errors is distributed among all features, with no clear preference for colours. Features are sorted from the highest to the lowest feature bias according to monkey S. Shaded areas represent 95% confidence intervals.

either at the point of stimuli selection or somewhere else in the decision-making process (see below). To explore the role of possible biases relating to the selection of stimulus features that are independent of recent choice history, we ranked stimulus features according to the proportion in which they were associated to unrewarded trials (**Fig. 8**). We found that both monkeys demonstrated an almost even distribution of errors across most of the task features indicating a similar likelihood to make a choice across task features independent of the task features local associated value. Thus, errors due to

non-value based attentional biases did not represent a dominant behavioral strategy systematically used by the animals (**Fig. 2**).

2.5 Discussion

We have shown that the learning of feature based attentional selection in macaque monkeys can be predicted by models of reinforcement learning with value based selection mechanisms acting on a restricted feature space. Value based learning

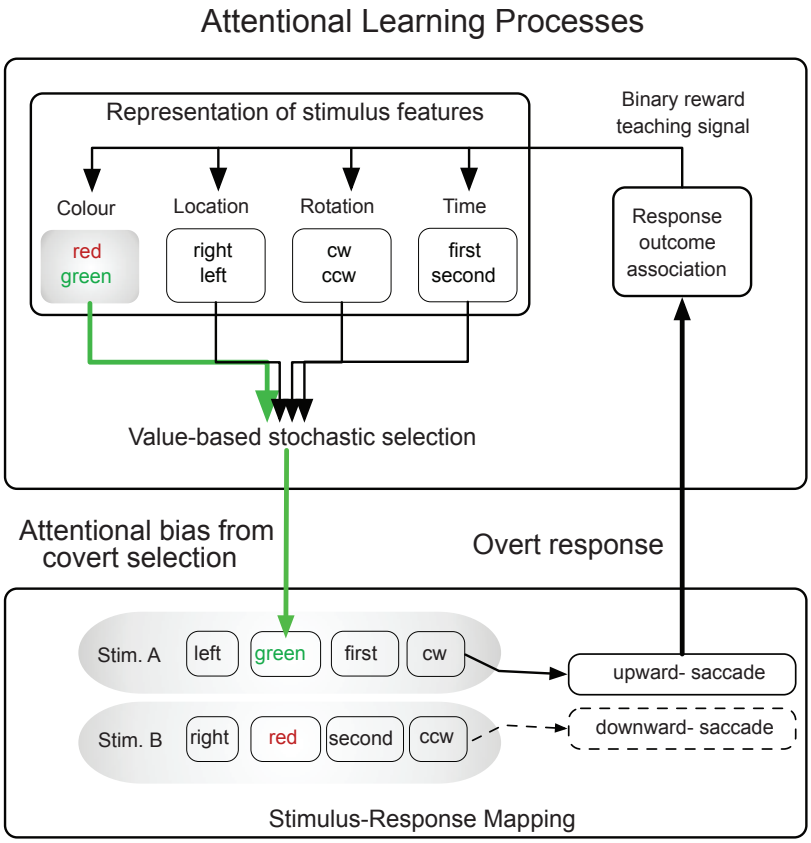


Figure 9. Separable processes underlying learning of attentional selection. Attentional selection relies on a covert decision making process that is suitable to evaluate all stimulus features (i.e. colour, location, etc.), but that after practice prioritizes the subset of relevant stimulus features that systematically link to reward (e.g., colour dimension). In each trial, a particular covert attentional selection of a stimulus (green stimulus in the example) is established by value-based competition among elements in the task set representation. Values are then updated according to the response outcome. Note that the specific response outcome critically depends on a proper attentional selection to bias the relevant sensory processing (e.g. rotation discrimination) and, as a result, trigger an adequate sensory-response mapping (e.g., upwards saccadic response to report a clockwise rotation).

explained the animals' behavior better when the updating of value representations was

restricted to the feature dimension that was task relevant (colour), and did not consider those feature dimensions (location, rotation direction, and time of rotation) that were not systematically linked to rewarding outcomes (in contrast to non-selective RL). This finding provides quantitative evidence in non-human primates that attentional selection can act on a task specific representation of relevant features. Such feature representation can be formally described as internal state model within the RL framework. Implicit in this formulation is that attention is realized as a stochastic covert selection acting on feature specific value predictions (**Fig. 9**). A second main finding of our study is that the process of value based attentional selection had to be complemented by an additional value independent selection process to account for non-random influences of selection history on the pattern of erroneous choices.

Internal value predictions of task relevant features provide the reference for attention.

We found that value-based learning of attention is not naive with respect to features in the environment that are systematically linked to reward. When an animal received a rewarding outcome, this reward was linked in individual trials to a choice on a particular rotation direction (clockwise or counterclockwise), a particular time onset (e.g. first or second), a particular location (left or right) and colour (e.g. blue) of the stimulus. If all these task variables were considered equivalently while updating value representations, the non-selective RL controller would have outperformed the feature-based learner, as local correlations of non-relevant features with reward outcomes would have impacted on monkeys' behavior. Instead, these multiple features were not treated equally in the credit assignment process (**Fig. 4B**). The updating of values was better described as being selective to prioritized task-specific representations. This finding highlights the

idea that a key component of flexible attention lies in the evaluation process of how causal sources of outcomes are identified and credited for producing the outcome. This empirically derived conclusion supports previous modeling studies that implicate attentional selection signals as critical gating signals for plasticity and learning of task relevant sensory features (Alexander, 2007; Roelfsema and van Ooyen, 2005) (see also Roelfsema et al., 2010). In summary, our findings show that the deployment of attention can be efficiently adjusted according to feature-reward associations. We should note that we could not model the origins of the segmentation of task relevant variables in the current dataset that was limited to later stages of task learning. However, we believe that it is an important future task to extend the feature-based RL model to include the learning of a segmentation between task relevant and irrelevant features by processes using either meta-learning mechanisms (Gershman and Niv, 2010; Ardid et al., 2014), or e.g. by adding an independent slow learning process that tracks input statistics and derives policies from it (Legenstein et al., 2010).

Attentional flexibility versus stickiness

After steep learning, the performance of monkeys did not reach optimality, but rather animals continued to make wrong, unrewarded choices in 10-15% of all trials during a period where expected values for stimulus colour were at a constant high level. We found that this 10-15% failure rate can be traced back to three identifiable sources that are informative about the processes controlling attention. The largest proportion of errors was accounted for by the softmax stochastic selection process (through the β parameter) that imposes a non-zero probability to select the stimulus features with the lowest values. This aspect is important because it supports the notion that attention can

be conceptualized as a stochastic selection process similar to conceptualization of overt (motor) choice (Rangel and Hare, 2010; Gottlieb, 2012).

A second source of errors in our task are feature biases of the animals that are independent of fluctuations in value predictions and reflect 'default' tendencies of animals choices (see **Fig. 9**), even though the animals could not (and did not) systematically deploy such simple strategies to solve our task (Shteingart and Loewenstein, 2014) (**Fig. 2**). The third source of erroneous performance referred explicitly to the pattern of errors that deviated from a purely stochastic process once in an asymptotic regime, with an evident tendency to repeat erroneous (unrewarded) choices (**Fig. 5**, **Fig. 6**). Both animals showed this deviation from random error generation, resembling perseveration tendencies and habit intrusions known from the motor domain. However, the repeated errors in our task referred to repetitions of the attentional selection (i.e. based on colour) from the previous trial. Only a single model was able to capture this error pattern by means of a sequential (2-step) process that complemented the basic value-based selection with a second selection process that pushed the final overt choice towards the previous selection.

Such a weighting of a current trial's value based selection is in fact an efficient strategy when the previous selection was rewarded; hence, repeating the same attentional selection is in such a condition a strategy that reduces effort and costs (Shenhav et al., 2013). However, when the previous trial's covert choice was an error and led to no reward, weighting the current value-based selection towards the non-rewarded previous covert choice is detrimental and incurs costs. This cost of committing two, or more consecutive errors represented a substantial sub-proportion of error trials

(34–41% out of the 10–15% total number of errors in the task), which may relate to the actual cost the animals are able to tolerate in the control of attention, given the effort it would take for them to improve performance. This interpretation is consistent with a recent proposal that quantifies the expected value of (attentional) control by estimating the (sum of anticipated) pay-offs against the costs to establish sufficiently strong control to obtain such pay-offs (Shenhav et al., 2013). According to this interpretation, the cost of committing errors in our task is traded against the level of effort (i.e. strength of control) that would be required to improve performance (number of rewarded trials). In particular in our task, improving performance requires constant updating of feature value representations and covert stimulus selection. We can thus speculate that the hypothesized quantity about the expected control intensity is related to the γ parameter in our Hierarchical Value-History model, which is adjusted to each monkey's tradeoff between effort and pay-off. The lower this parameter the higher is the effort to receive more value-based pay-offs, and on the contrary, the higher the parameter the larger is the attentional stickiness and the tendency to perseverate on previous attentional selections.

Implications for models of attention: value and non-value based processes.

The success in explaining actual attentional learning in primates with a feature-based RL mechanism that is weighted with an attentional stickiness process has further implications for theories of attention. Firstly, the results suggest that the valuation system plays a key role in determining what features selective attention is shifted towards independently of the saliency of those features (Navalpakkam et al., 2010; et al., 2011; Tatler et al., 2011; Chelazzi et al., 2013). Value representations in the RL

framework are predictions of stimulus values (predictions of outcomes), demonstrating that the covert control of visual attention can be understood from a predictive coding perspective such that feature-value predictions resemble reward-value predictions in the domain of overt goal-directed behavior, decision making and planning (Dehaene and Changeux, 2000; Seymour and McClure, 2008; Wilson and Niv, 2011; van der Meer et al., 2012). This conclusion resonates well with studies documenting the influence of expectations for visual perception and perceptual inferences (Summerfield and Egner, 2009; Seriès and Seitz, 2013), the influence of secondary reward associations to modify basic visual search efficiency (Anderson et al., 2011), and a growing literature documenting the influence of actual attentional experiences to shape reward memories and attentional priorities through learning mechanisms (Libera and Chelazzi, 2009; Awh et al., 2012; Chelazzi et al., 2013; Gottlieb et al., 2014).

Secondly, attention in our task also depends on a second process that weights the value based selection based on repeating the selection of previous trials irrespective of whether that selection was rewarded or unrewarded. Such a reward-insensitive mechanism is particularly useful in probabilistic choice contexts where the lack of reward at one occasion can be a mere stochastic event that is better ignored to maximize reward intake in the long run (Lau and Glimcher, 2005). In our task with a deterministic reward schedule within blocks of trials, the weighting of the current choice towards previous choices is reminiscent of (1) previous trial effects in stimulus-response learning tasks (Fecteau and Munoz, 2003), and shares similarity with (2) habitual stimulus response control (Dolan and Dayan, 2013), (3) habit intrusions (de Wit et al., 2012), (4) behavioral perseverations and stickiness (Huys et al., 2011; Dayan and

Berridge, 2014), as well as with (5) inter-trial priming and repetition memory effects (Kristjánsson, 2006; Kristjánsson and Campana, 2010; Awh et al., 2012; Anderson, 2013). All these listed effects are empirical demonstrations of the apparent influence of a memory of recent choices and attentional selections on current attentional performance. Whether these various history and memory effects serve as primary controllers of attentional selections or should better be conceived of as modulators of attention will be a question for future research. Our findings are more supportive of the former suggestion, revealing that selection history influences attentional performance in such a dominant way that it should be considered a separate control process underlying attentional selection, which complements value based control.

Taken together, we have illustrated a formal framework of attentional selection in non-human primates that provides explicit and testable hypotheses about the specific subprocesses underlying attentional control. Our hierarchical reinforcement learning model specifies these three main attentional subprocesses as (1) the feature specific learning of value predictions, (2) the stochastic value-based selection process, and (3) a non-value based memory bias that drives the system towards previously selected information. We speculate that the very structures implicated in stimulus valuation, reinforcement learning, and decision making are key structures in controlling the focus of visual attention. Each of these processes is possibly associated with separable neuronal circuits in the primate prefrontal, striatal and medial temporal lobe systems. Circuits within prefrontal regions presumably include the lateral prefrontal cortex, an area that may not have an anatomical and functional analog in the non-primate brain (Passingham et al., 2012). Our study in non-human primates could thus become a

versatile starting point to understand how multiple choice systems and subprocesses underlying stimulus selection interact to determine the target of covert attention in primates.

2.6 References

Ahn W-Y, Busemeyer J, Wagenmakers E-J, Stout J (2008) Comparison of Decision Learning Models Using the Generalization Criterion Method. Cognitive Sc: A Multidisciplinary J 32:1376–1402.

Alexander WH (2007) Shifting Attention Using a Temporal Difference Prediction Error and High-Dimensional Input. Adaptive Behavior 15:121–133.

Anderson BA (2013) A value-driven mechanism of attentional selection. Journal of Vision 13:7–7.

Anderson BA, Laurent PA, Yantis S (2011) Value-driven attentional capture. PNAS 108:10367–10371.

Ardid S, Balcarras M, Womelsdorf T (2014) “Adaptive learning” as a mechanistic candidate for reaching optimal task-set representations flexibly. BMC Neuroscience.

Ardid S, Wang X-J (2013) A tweaking principle for executive control: neuronal circuit mechanism for rule-based task switching and conflict resolution. J Neurosci 33:19504–19517.

Asaad WF, Eskandar EN (2008) Achieving behavioral control with millisecond resolution in a high-level programming environment. J Neurosci Methods 173:235–240.

Awh E, Belopolsky AV, Theeuwes J (2012) Top-down versus bottom-up attentional control: a failed theoretical dichotomy. Trends in Cognitive Sciences 16:437–443.

Cai X, Padoa-Schioppa C (2014) Contributions of orbitofrontal and lateral prefrontal cortices to economic choice and the good-to-action transformation. Neuron 81:1140–1151.

Chelazzi L, Perlato A, Santandrea E, Libera Della C (2013) Rewards teach visual selective attention. Vision Res 85:58–72.

Dayan P, Berridge KC (2014) Model-based and model-free Pavlovian reward learning: revaluation, revision, and revelation. Cognitive, Affective, & Behavioral Neuroscience 14:473–492.

Dayan P, Kakade S, Montague PR (2000) Learning and selective attention. Nat Neurosci 3:1218–1223.

de Wit S, Watson P, Harsay HA, Cohen MX, van de Vijver I, Ridderinkhof KR (2012) Corticostriatal connectivity underlies individual differences in the balance between habitual and goal-directed action control. J Neurosci 32:12066–12075.

Dehaene S, Changeux JP (2000) Reward-dependent learning in neuronal networks for planning and decision making. Prog Brain Res 126:217–229.

Dolan RJ, Dayan P (2013) Goals and habits in the brain. Neuron 80:312–325.

Donoso M, Collins AGE, Koechlin E (2014) Human cognition. Foundations of human reasoning in the prefrontal cortex. Science 344:1481–1486.

Fecteau JH, Munoz DP (2003) Exploring the consequences of the previous trial. Nat Rev Neurosci 4:435–443.

Gershman SJ, Niv Y (2010) Learning latent structure: carving nature at its joints. Current Opinion in Neurobiology 20:251–256.

Glimcher PW (2011) Colloquium Paper: Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. Proceedings of the National Academy of Sciences 108:15647–15654.

Gottlieb J (2012) Attention, learning, and the value of information. Neuron 76:281–295.

Gottlieb J, Hayhoe M, Hikosaka O, Rangel A (2014) Attention, reward, and information seeking. J Neurosci 34:15497–15504.

Hare TA, Schultz W, Camerer CF, O'Doherty JP, Rangel A (2011) Transformation of stimulus value signals into motor commands during simple choice. Proceedings of the National Academy of Sciences 108:18120–18125.

Huys QJM, Cools R, Gölzer M, Friedel E, Heinz A, Dolan RJ, Dayan P (2011) Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. PLoS Comput Biol 7:e1002028.

Kaping D, Vinck M, Hutchison RM, Everling S, Womelsdorf T (2011) Specific Contributions of Ventromedial, Anterior Cingulate, and Lateral Prefrontal Cortex for Attentional Selection and Stimulus Valuation Behrens T, ed. PLoS Biol 9:e1001224.

Kennerley SW, Behrens TEJ, Wallis JD (2011) Double dissociation of value computations in orbitofrontal and anterior cingulate neurons. Nature Publishing Group 14:1581–1589.

Kristjánsson Á (2006) Rapid learning in attention shifts: A review. Visual Cognition 13:324–362.

Kristjánsson Á, Campana G (2010) Where perception meets memory: a review of repetition priming in visual search tasks. Atten Percept Psychophys 72:5–18.

Kruschke JK, Hullinger RA (2010) Evolution of attention in learning Schmajuk N, ed. Computational models of conditioning.

Lau B, Glimcher PW (2005) Dyanmic response-by-response models of matching behavior in Rhesus monkeys. Journal of the experimental analysis of ….

Lau B, Glimcher PW (2008) Value representations in the primate striatum during matching behavior. Neuron 58:451–463.

Legenstein R, Wilbert N, Wiskott L (2010) Reinforcement learning on slow features of high-dimensional input streams. PLoS Comput Biol 6.

Libera Della C, Chelazzi L (2009) Learning to attend and to ignore is a matter of gains and losses. Psychological Science 20:778–784.

Luk C-H, Wallis JD (2013) Choice coding in frontal cortex during stimulus-guided or action-guided decision-making. J Neurosci 33:1864–1871.

Navalpakkam V, Koch C, Rangel A, Perona P (2010) Optimal reward harvesting in complex perceptual environments. PNAS 107:5232–5237.

Padoa-Schioppa C (2011) Neurobiology of economic choice: a good-based model. Annu Rev Neurosci 34:333–359.

Passingham RE, Passingham RE, Wise SP (2012) The Neurobiology of the Prefrontal Cortex. Oxford University Press.

Peck CJ, Jangraw DC, Suzuki M, Efem R, Gottlieb J (2009) Reward modulates attention independently of action value in posterior parietal cortex. J Neurosci 29:11182–11191.

Peck CJ, Lau B, Salzman CD (2013) The primate amygdala combines information about space and value. Nat Neurosci 16:340–348.

Rangel A, Clithero JA (2012) Value normalization in decision making: theory and evidence. Current Opinion in Neurobiology 22:970–981.

Rangel A, Hare T (2010) Neural computations associated with goal-directed choice. Current Opinion in Neurobiology 20:262–270.

Roelfsema PR, van Ooyen A (2005) Attention-gated reinforcement learning of internal representations for classification. Neural Computation 17:2176–2214.

Roelfsema PR, van Ooyen A, Watanabe T (2010) Perceptual learning rules based on reinforcers and attention. Trends in Cognitive Sciences 14:64–71.

Rushworth MFS, Behrens TEJ (2008) Choice, uncertainty and value in prefrontal and cingulate cortex. Nat Neurosci 11:389–397.

Rushworth MFS, Noonan MP, Boorman ED, Walton ME, Behrens TE (2011) Frontal Cortex and Reward-Guided Learning and Decision-Making. Neuron 70:1054–1069.

Seriès P, Seitz AR (2013) Learning what to expect (in visual perception). frontiers in Human Neuroscience 7:668.

Seymour B, McClure SM (2008) Anchors, scales and the relative coding of value in the brain. Current Opinion in Neurobiology 18:173–178.

Shenhav A, Botvinick MM, Cohen JD (2013) The expected value of control: an integrative theory of anterior cingulate cortex function. Neuron 79:217–240.

Shteingart H, Loewenstein Y (2014) Reinforcement learning and human behavior. Current Opinion in Neurobiology 25:93–98.

Sugrue LP, Corrado GS, Newsome WT (2004) Matching behavior and the representation of value in the parietal cortex. Science 304:1782–1787.

Summerfield C, Egner T (2009) Expectation (and attention) in visual cognition. Trends in Cognitive Sciences 13:403–409.

Tatler BW, Hayhoe MM, Land MF, Ballard DH (2011) Eye guidance in natural vision: reinterpreting salience. Journal of Vision 11:5.

Tsotsos JK (2011) A Computational Perspective on Visual Attention. MIT Press.

van der Meer M, Kurth-Nelson Z, Redish AD (2012) Information processing in decision-making systems. Neuroscientist 18:342–359.

Wilson RC, Niv Y (2011) Inferring relevance in a changing world. frontiers in Human Neuroscience 5:189.

Wunderlich K, Rangel A, O'Doherty JP (2010) Economic choices can be made using only stimulus values. PNAS 107:15005–15010.

Chapter 3.

A flexible mechanism of rule selection enables rapid feature-based reinforcement learning in new environments

to be submitted to **Frontiers in Human Decision Neuroscience**

Matthew Balcarras and Thilo Womelsdorf

Department of Biology, Centre for Vision Research, York University, 4700 Keele Street, Toronto, Ontario, M6J 1P3, Canada.

Correspondence: Dr. Thilo Womelsdorf, Mr. Matthew Balcarras.
York University, Department of Biology, 4700 Keele Street,
Toronto ON M3J 1P3, Canada

3.1 Abstract

Learning in a new environment is influenced by prior learning and experience. Correctly applying a rule that maps a context to stimuli, actions, and outcomes enables faster learning and better outcomes compared to relying on strategies for learning that are ignorant of task structure. However, it is often difficult to know when and how to apply learned rules in new contexts. In our study we explored how subjects employ different strategies for learning the relationship between stimulus features and positive outcomes in a probabilistic task context. We test the hypothesis that naive subjects will show enhanced learning of feature specific reward associations by switching to the use of abstract rule that associates stimuli by feature type and restricts selections to that dimension. To test this hypothesis we designed a decision making task where subjects receive probabilistic feedback following choices between pairs of stimuli. In the task, trials are grouped in two contexts by blocks, where in one type of block there is no unique relationship between a specific feature dimension (stimulus shape or colour) and positive outcomes, and following an un-cued transition, alternating blocks have outcomes that are linked to either stimulus shape or colour. Two-thirds of subjects (n=22/32) exhibited behaviour that was best fit by a hierarchical feature-rule model. Supporting the prediction of the model mechanism these subjects showed significantly enhanced performance in feature-reward blocks, and rapidly switched their choice strategy to using abstract feature rules when reward contingencies changed. Choice behaviour of other subjects (n=10/32) was fit by a range of alternative reinforcement

learning models representing strategies that do not benefit from applying previously learned rules. We show that untrained subjects are capable of flexibly shifting between behavioural rules by leveraging simple model-free reinforcement learning and context-specific selections to drive responses.

3.2 Introduction

Successful behavior in new environments benefits from leveraging learning from previous experience in the form of abstract rules - the mapping of contexts, stimuli, actions and outcomes - even though it is often difficult to know which rule is relevant to the current context (Miller, 2000; Gershman et al., 2010a; Buschman et al., 2012; Chumbley et al., 2012; Collins and Frank, 2013; Collins et al., 2014). One of the hallmarks of human behavior is that in new environments with unknown relationships between stimuli and outcomes, subjects generalize from previous experiences (Seger and Peterson, 2013; Collins et al., 2014), even when expectations about the value of stimuli for predicting reward may not be beneficial (Anderson and Yantis, 2013; Shteingart and Loewenstein, 2014). Fortunately, there is significant continuity across our every-day decision making contexts that enables positive transfer of previously learned rules, and in fact, humans work very hard to pattern our living and working environments in such a way as to provide continuity with contextual cues indicating the relevant rule to apply (Collins et al., 2014). For example, objects coloured bright red often indicate emergency response equipment, and materials and objects with specific shapes, like hexagons, indicate specific information about appropriate responses, like stopping your vehicle. However people do not always apply rules when it is beneficial to

do so. This could be because it is unclear which rule to apply or that an appropriate rule for this context has not been learned.

In this study we set out to test if naive and uncued subjects will spontaneously apply a flexible rule for learning stimulus-feature reward associations and how this behaviour can be captured in formal reinforcement learning frameworks. In particular, we explored how subjects leverage an abstract rule that maps stimulus colour and shape, independently of each other, to choice outcomes in order to improve the local learning of associations between stimuli and feedback. We hypothesized that untrained subjects exploit previous learning by spontaneously assuming that the feature dimensions of shape and colour would be relevant for solving the task and that this would translate into improved performance through a contextually structured selection process.

It is not clear how to formalize the flexible application of behavioural rules in the reinforcement learning (RL) model framework. One solution is to adapt hierarchical RL methods (Collins and Frank, 2013). There is considerable similarity between applying pre-learned rules and hierarchical learning strategies. Structuring stimulus selection hierarchically incorporates expectations about the relevance of stimuli in terms of initiation conditions, the conditions under which an alternate selection sequence is triggered (Botvinick et al., 2009; Badre and Frank, 2012; Botvinick, 2012). Previous work on hierarchical RL has focused on the benefits of temporally abstract actions, where instead of selecting from among available primitive actions, the model can select a behavioral subroutine that employs a sequence of actions. Extending this approach, we developed a model that hierarchically structures the stimulus selection process

among competing values for stimulus features. In the default scenario, basic model-free RL learns the expected value of features of visual stimuli and stochastically selects among the values of available stimulus features to receive outcomes (Donoso et al., 2014). Following the hypothesis that subjects have learned from pre-task experience that the feature categories of shape and colour are often relevant for local learning, the model compares the total expected value for stimulus features of each type, and when the difference between these total group values crosses a threshold an alternate selection process begins and stimulus selection acts only on the learned value of the relevant subset of features, i.e. the feature type (shape or colour) that is greatest (see **Materials & Methods**). The threshold is an independent model parameter fit to each subject that reflects the confidence of the model in determining a feature-value context. We believe that this adaptation of hierarchical RL represents a simple and intuitive framework for capturing the natural learning processes of untrained subjects in an operant learning environment, and provides testable implications for future research into the neural underpinnings of these processes.

3.3 Materials & Methods

Task design

All experimental procedures were approved by York University's Ethics Review Board. Thirty seven participants from the York University community participated in the experiment (age range 19-35, 26/11 male/female), and all gave their informed consent (see **Appendix B**). Participants were offered an incentive for participating in the form of a gift card valued at $10 CAD. Participants performed the experiment on a touch

sensitive Sony Vaio laptop running Windows 8, and Matlab (The Mathworks Inc.) with the Psychophysics toolbox (www.mathworks.com; www.psychtoolbox.org) and custom written Matlab scripts controlling the experiment. The laptop had a 15" capacitive touch sensitive monitor with a resolution of 1920x1080 pixels and a refresh rate of 60 Hz. Stimuli were placed at 4.6 degree from the central fixation point. The laptop was positioned comfortably, ~50-70cm, in front of subjects to ease their holding and touching responses. The temporal resolution of the touchscreen responses were in the order of 997 milliseconds (± 26msec. SEM). At the start of the experiment, participants were instructed to use the index finger of their dominant hand to touch one of the two presented stimuli, then use the same finger to hold the spacebar to receive feedback, and to make choices that maximized the number of positive feedbacks.

A trial began with the presentation of a small cross in the centre of the screen (**Fig. 1A**). After 300-600 milliseconds two stimuli appeared in two of three possible positions. The location of stimuli was randomly chosen from canonical locations equidistant from each other and the central cross. After another 1500 milliseconds the central cross was removed and subjects were free to select a stimulus. If subjects selected a target before the removal of the fixation cross, the stimuli were removed and a message was displayed reminding the subject to wait for the removal of the cross. This message was displayed for a waiting period of 500 milliseconds before a new trial began. Following the selection of a stimulus, the stimuli were removed and a message appeared on the screen informing subjects to hold the spacebar in order to receive feedback. Feedback was not given until the spacebar was depressed for 1000 milliseconds, and was either a gold star in the middle of the screen or a message saying 'sorry' when the schedule

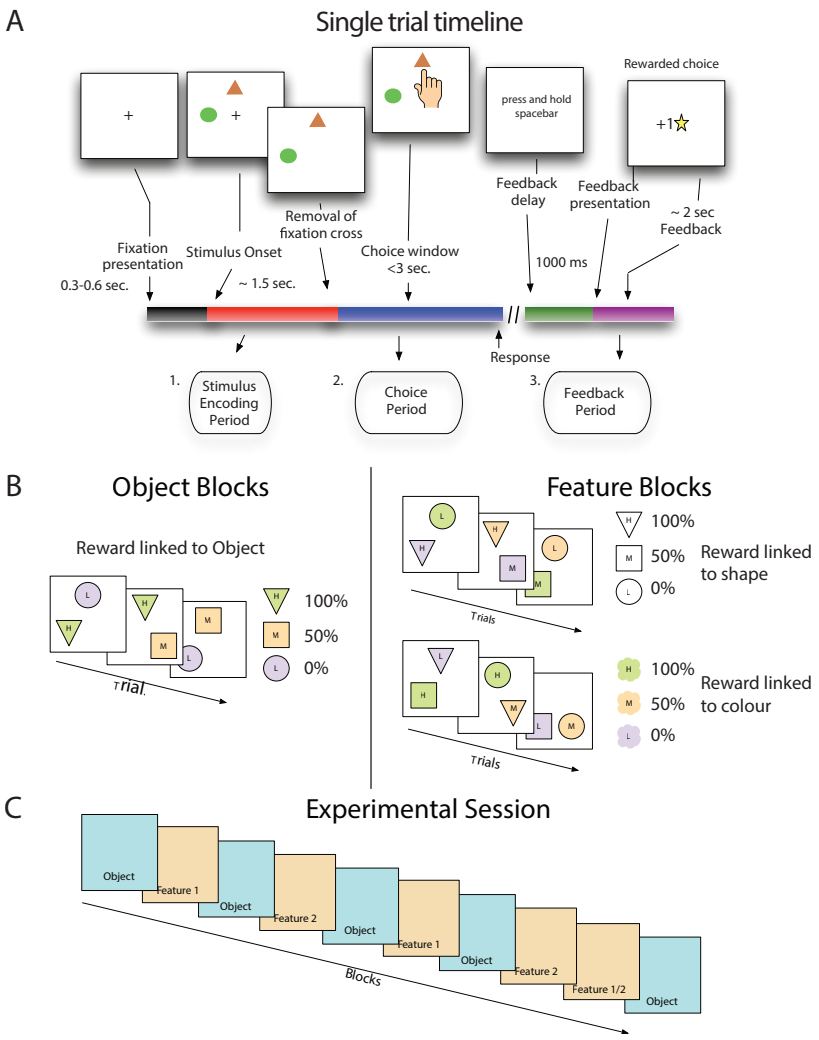


Figure 1. Stimulus value learning task. A) Subjects learned by trial and error that stimuli and stimulus features are linked to the likelihood of receiving positive outcomes. In the displays, the red 'x' denotes the chosen stimulus of subjects. The yellow starts on top of each panel indicated the feedback for correctly chosen stimuli. The right panel vertical summarizes the choice outcomes for trials shown on the left to illustrate the subjects putative internal state for determining selections on future trials. B) Stimulus reward associations were structured either such that fixed pairs of colours and shapes (in sets of three) had a probabilistic relationship with reward (object blocks) or such that stimulus features were not fixed to each other and only one feature type (either shape or colour) was linked to reward. C) For the first eight blocks, feature blocks followed object blocks using the same set of shapes and colours as the preceding object block, but with new feature-reward associations. The last pair of blocks flipped this pattern where a feature reward block (either type 1 or 2, randomly selected) precedes an object block.

associated with that stimulus determined it was either a rewarded or an unrewarded trial (see below). Gold stars awarded to the subject accumulated at the bottom of the screen, indicating to the subjects their performance thus far. After the last trial of the

session was completed, a screen was displayed which thanked the subject for participation and provided a final count of gold stars received.

Subjects made choices on stimuli that were combinations of shapes and colours. Each object block began with a new set of three shapes and three colours drawn from a set of six, and all stimuli for that block were made from combinations of these three shapes and colours (**Fig. 1B**). In object blocks, shape-colour pairs remain fixed throughout the block so that there is only thee unique stimuli appearing in the block. Feature blocks that followed object blocks used the same set of shapes and colours that appeared in the previous block, but now stimuli could be composed of any combination of colour and shape, so that there were nine possible unique stimuli appearing in the block.

The task included a hidden probabilistic reward schedule that assigned a probability of positive outcome on each trial to the two available stimuli (**Fig. 1B** - right panels). In object blocks each stimuli, a unique colour-shape pair, is assigned a probability of positive outcome, with one being 0%, one 50% and one 100%. In feature blocks, outcome probabilities are associated with a specific feature dimension, either shape or colour. In a colour-feature block, one colour is predictive of positive outcomes 0%, one 50%, and one 100%. Shape-feature blocks work the same as colour feature blocks except that probabilities are linked to stimulus shape instead of colour. In feature blocks, the non-relevant feature is only spuriously related to outcomes because of the randomized relationship between colours and shapes in these blocks. Receiving a positive outcome for a choice on colour A and shape B in a shape-feature block will not tell you anything about the likelihood of receiving a positive outcome on the next trial

where colour A appears. In both feature and object blocks, stimulus location was selected randomly and was never predictive of outcomes.

Subjects performed a stereotypical sequence of blocks (**Fig. 1C**). An experimental session began with an object block, followed by a feature block, where the relevant feature was selected at random, then another object block and feature block, where this feature is the alternate one from the first feature block. This sequence continued until the ninth block which reversed the object-feature order, and the relevant feature was randomly selected, with the final block being an object block.

Blocks ranged in length from 15-25 trials with the number of trials in a block determined by a performance criterion. If the subject had made 11 out of the first 15 choices correctly then the block ended at trial 15. Trials continued until either 80% of the last 10 trials were correct or the subject reached 25 trials. Average block length across subjects was 17.8 (SE ±2.1). In total subjects performed 7106 trials, of which 3964 trials that were from blocks showing learning were included in analysis.

Behavioral Data Analysis.

Data Analysis was done with custom written Matlab scripts (The Mathworks Inc.). Learning in a block was determined following the method of Wilson and Niv (Wilson and Niv, 2011), whereby if the slope of the average performance line from the beginning to the end of the block was positive and was above chance performance (50% correct) at the end, the block was considered to show learning.

Correct choices were determined by the selection of the stimulus with the higher probability of a positive outcome, independently of whether a positive outcome was received. On trials where the 100% likely stimulus appeared, it was always the correct

stimulus to select, even if selecting the 50% likely stimulus produced a positive outcome. Likewise on trials where the 0% likely stimulus appeared, it was always the incorrect stimulus to select, even if selecting the 50% likely stimulus did not produce a positive outcome.

Reaction times were quantified from the time when the fixation cross was removed until the time when the screen was touched. If the subject touched the screen before the fixation cross was removed the trial was declared an 'early response' and was not included in further analysis.

The generalized linear model regression was performed by using 1) the block type or 2) the block number (from 1-10), against the mean proportion of correct choices in the whole block of trials, including those blocks that did not show learning overall, in order to determine if block types or repeated exposure to the task over time was predictive of performance. This regression produced a coefficient with a corresponding p-value indicating whether the beta-coefficient has a significant predictive relationship with the average performance.

RL model algorithms.

In the basic Q-Learning Rescorla Wagner RL model (QL Basic), the value of any predictor of reward (stimulus feature, Q_i) is updated on the next time step (trial) from its previous value through the scaled reward-prediction error: The difference between the binary reward outcome (R, either 0 or 1) and the predictor itself (Skvortsova et al., 2014). The scaling factor (α) represents the learning rate:

$$Q_i(t+1) = Q_i(t) + \alpha[R(t) - Q_i(t)] \quad \text{(eq. 1)}$$

Other than the QL Basic model, all other models implemented a generalization of outcome information across all Q values. Thus, all stimulus features associated with the selected stimulus updated their value according to equation 1. Stimulus features associated with the other, non-selected stimulus were updated according to:

$$Q_i\,(t+1) = Q_i(t) + \alpha[1 - R(t) - Q_i(t)] \qquad \text{(eq. 2)}$$

The second model, QL Gen, extended QL Basic with generalization of outcome information across all Qs for features appearing on that trial and no other changes. In the third model, QL Decay, feature values were updated when they were associated with the selected stimulus features in the same way as QL Basic and QL Gen, but all non-selected features had their associated values decay as a function of time governed by the rate of decay (τ) according to:

$$Q_i\,(t+1) = Q_i(t) + \alpha[1 - R(t) - Q_i(t)] * \tau \qquad \text{(eq. 3)}$$

The fourth model, QL GainLoss, employed the same framework as QL Gen, but applied a different learning rate to positive and negative outcomes - α_G vs α_L.

$$Q_i\,(t+1) = Q_i(t) + \alpha_G[R(t) - Q_i(t)] \qquad \text{(eq. 4)}$$

$$Q_i\,(t+1) = Q_i(t) + \alpha_L[1 - R(t) - Q_i(t)] \qquad \text{(eq. 5)}$$

Stimulus feature values for all non-HRL models were non-linearly transformed into choice probabilities according to the Boltzmann equation:

$$P_i(t) = {}_e\beta Qi(t) / \Sigma_e \beta Qj(t) \quad \text{(eq. 6)}$$

where β represents the inverse temperature and establishes the strength of the non-linearity.

The *Flexible Rule Selection model* (FR_Sel) employs a selection function that is an adaptation of the standard Boltzman formulation. Rather than all available Qs competing for final selection via participating as possible choice probabilities, FR_Sel compares Q values across features by feature type, calculating the difference between the sum of total values for each type. When the difference between the total value for one feature type relative to the other types moves past a threshold (λ), only that set of values is used to compute choice probabilities according to the equations below:

$$P_i(t) = {}_e\beta Qi(t) / \Sigma_e \beta Qsel(t) \quad \text{(eq. 7)}$$

where Q_{sel} is the set of Qs such that:

$$Q_{sel} > \Sigma\, Q_{others} + \lambda \quad \text{(eq. 8)}$$

Model Optimization.

Models were optimized by performing a grid search across the total parameter space for each free parameter, attempting to minimize the ordinary least square distance between the probability associated with selecting the correct stimulus and the observed likelihood of selecting the correct stimulus (Donoso et al., 2014)(Bergstra, 2012). On

each trial the model was given the choice made by a subject and transformed that into values according the learning rate(s) of that model iteration. Values were converted into choice probabilities according the Boltzman equation and the value of β (Glimcher, 2011). The mean probability associated with the correct choice was calculated for each trial from the block start across all blocks. Values for free parameters were selected that minimized the distance between this mean probability and the mean likelihood of the subject making a correct choice.

To ensure that we fit the models to the most systematic behaviour, we bootstrapped 80% of the data from each subject 100 times for each set of parameter values, and calculated the mean ordinary least squares (OLS) score across these 100 iterations. Bootstrapping is a known method of estimating the variance of model performance (Zucchini, 2000). To confirm that optimized model results reflect systematic trends in the data and to correct for model complexity we performed a cross-validation of the model predicted data for each parameter set. Data was split in half by random selection and repeated ten times for each parameter set to ensure that results were consistent independent of data sampling. Using the Wilcoxon-Mann-Whitney test, we found that for each parameter set and each model across all subjects, there was no significant difference in score between data groups ($p>0.05$) compared to the bootstrapped results.

We did not use statistical methods for model comparison, such as the Akaike or Bayesian Information Criterion, because 1) other studies have shown that using OLS is equally capable of identifying the best model (Donoso et al., 2014), and 2) we fit the models to subject performance split by block type, which essentially creates two

datasets, and information criterion scores are not comparable across datasets (Zucchini, 2000).

3.4 Results

Behavior.

We show that average choice behavior across subjects is best explained by a reinforcement learning model that identifies the current task context and then applies a selection rule that associates stimuli by feature type and restricts stimulus selection to the relevant (i.e context specific) stimulus feature. In each trial subjects were required to make choices between two stimuli on a touchscreen and use visual feedback to learn the value of stimulus features for predicting positive outcomes (**Fig. 1a**). The likelihood of reward (a gold star displayed on the screen) was determined by a reward schedule for each stimulus or stimulus feature and was either 0, 50, or 100% and stayed constant per block, but changed without overt cue between blocks. Blocks lasted between fifteen and twenty-five trials depending on performance and a random jitter added to block length (See **Materials & Methods**). Our task represents different contexts in different blocks by changing the relationship between stimulus features and feedback (**Fig. 1b**). In half of the blocks, a set of shapes and colours were randomly shuffled to create three stimuli, so that on successive trials stimulus shape and colour are combined to form objects that maintain a continuous relationship throughout the block and likelihood of reward is attached to each fixed object. Only two stimuli are presented on each trial and the location of each stimulus is randomly selected from three possible spots. In the other half of the blocks stimulus shape and colour are combined in a continuously

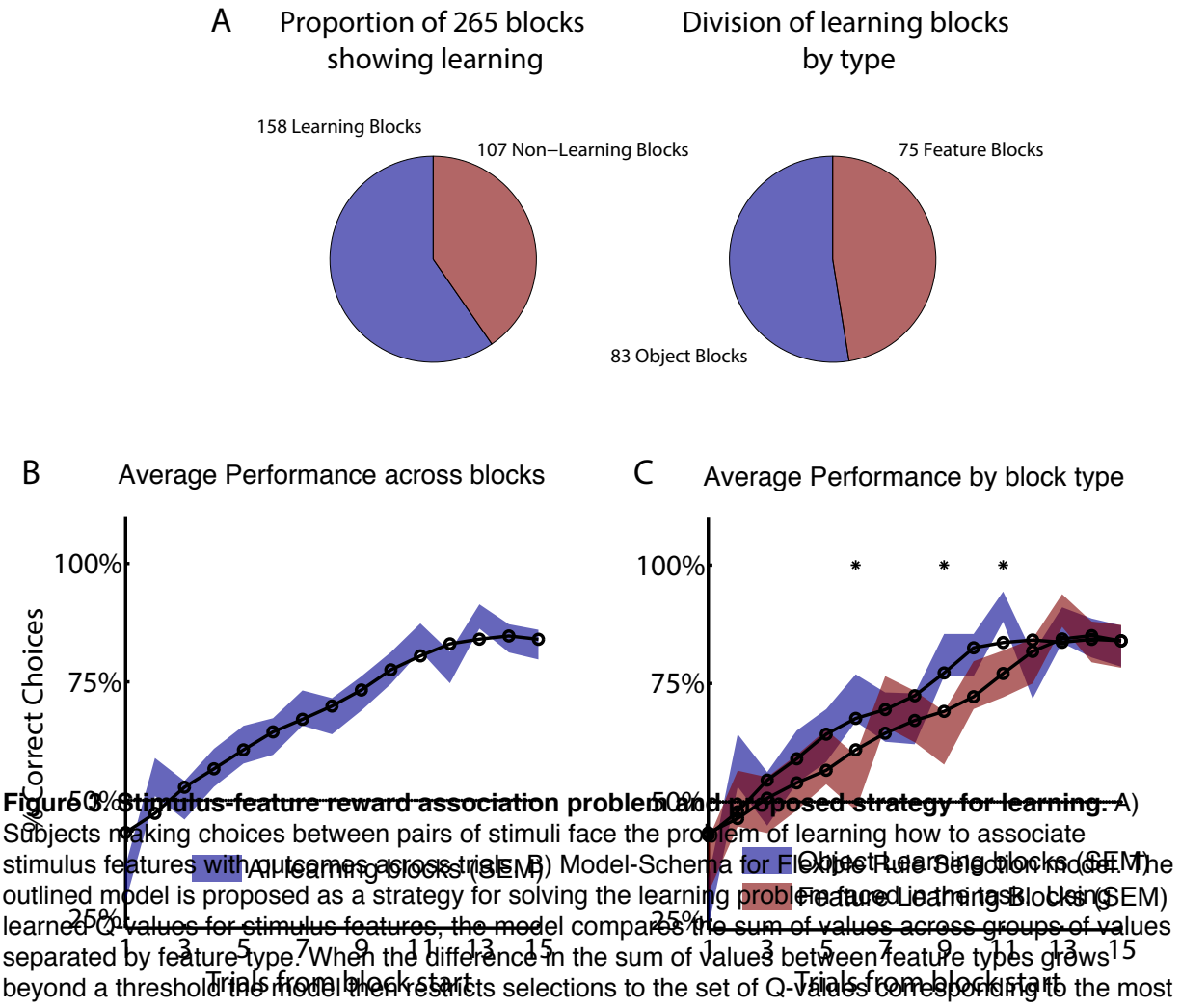


Figure 3. Stimulus-feature reward association problem and proposed strategy for learning. A) Subjects making choices between pairs of stimuli face the problem of learning how to associate stimulus features with outcomes across trials. B) Model-Schema for Flexible Rule Selection model. The outlined model is proposed as a strategy for solving the learning problem faced in this task. Using learned Q-values for stimulus features, the model compares the sum of values across groups of values separated by feature type. When the difference in the sum of values between feature types grows beyond a threshold the model then restricts selections to the set of Q-values corresponding to the most valuable feature type.

Figure 2. Learning by block type and across blocks. A) Proportion of learning blocks (left) and the distribution of learned feature and learned object blocks (right panel). Learning was identified by computing the slope of average performance in a block (see **Materials & Methods**) B) Average performance (shading indicates SEM) across all learning blocks shows a consistent increase in the proportion of correct responses for both block types. The stars denote trials with statistically significant differences in performance between block types (Mann-Whitney-Wilcoxon, $p < 0.05$). The 50% line shows chance level performance. C) Splitting average performance by block type shows that

random fashion with the likelihood of reward only being linked to one feature dimension (Gershman et al., 2010b). Different contexts were then alternated over ten blocks in a session (**Fig. 1c**).

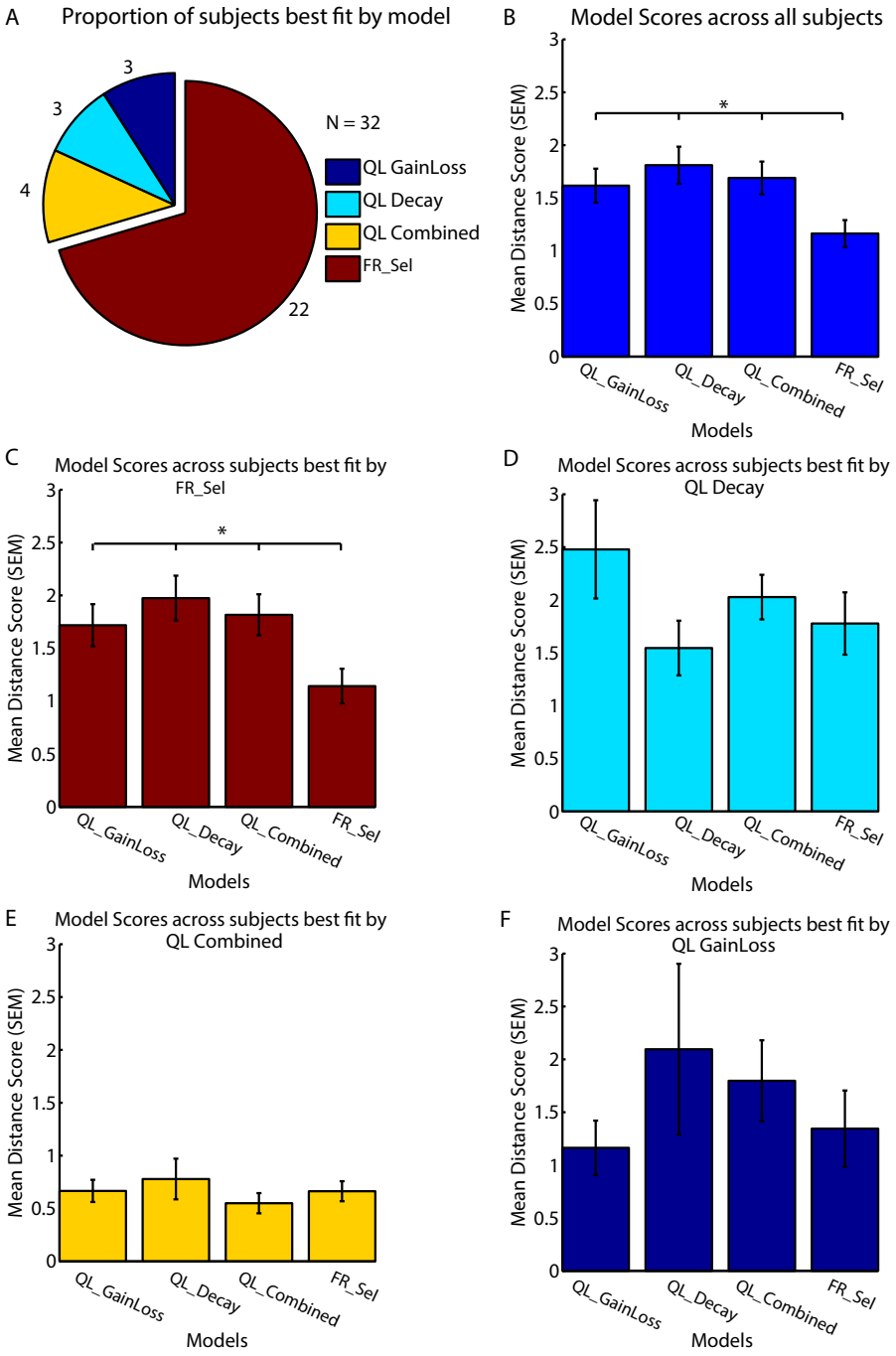
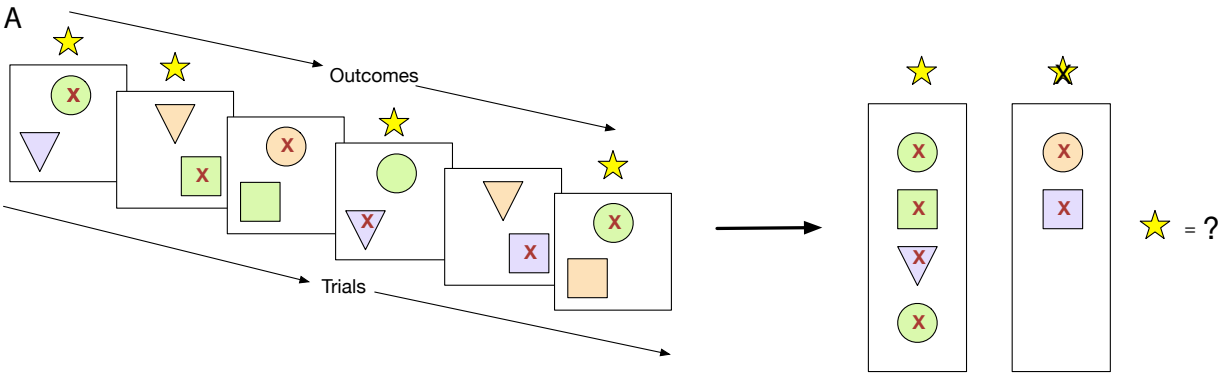
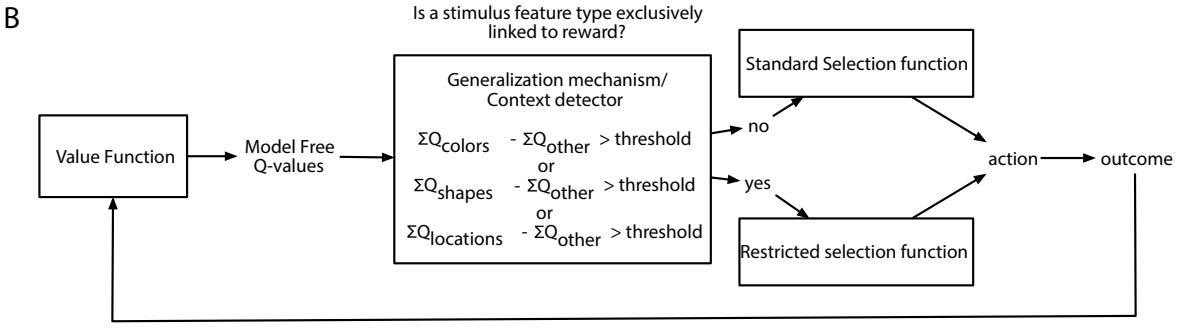


Figure 4. Model performance across subjects. A) All models were fit to each subject individually, with 69% (n=22/32) best fit by the flexible rule selection model (FR_Sel). B) The average ordinary least square score (y-axis) across all subjects for the FR_Sel model was significantly better than all other models for the subjects it fit best. (The star denotes p <0.05, Mann-Whitney-Wilcoxon test, for each pairwise comparison with FR_Sel). C) For subjects best fit by the FR_Sel model, fits with the alternative models were significantly worse. D) For the 11% (n=4) subjects best fit by the QL Decay model, the FR_Sel model provided the second best fit. E) The 11% (n=4) subjects best fit by the QL Combined model showed particular low ordinary least square scores across models. F) For the 5% (n=2) subjects best fit by the QL GainLoss model, the FR_Sel model provided the second best fit.





Subjects (n=37, all right handed, 26 male/11 female) were able to successfully use feedback to learn the correct stimulus outcome association in a majority of blocks. Using a simple criterion for learning in a block with constant feature-reward associations (See **Materials & Methods**, Wilson and Niv, 2011), we found that the majority of subjects (n=32/37) showed learning in on average 158/265 (59.6%) blocks of trials (**Fig. 2a**). Five subjects performed at chance or showed no increase in performance and were excluded from further analysis. Of the 158 blocks in which subjects learned the reward associated rule, 52.5% (83/158) were object blocks. In feature blocks, where only shapes or colours are linked to reward probabilities, subjects showed learning in 47.5% (75/158) of blocks (**Fig. 2a**).

Across subjects and all blocks that showed learning, the proportion of correct choices reached a peak of 82.9% (SE ±.03) on trial 15 (**Fig. 2b**). When average

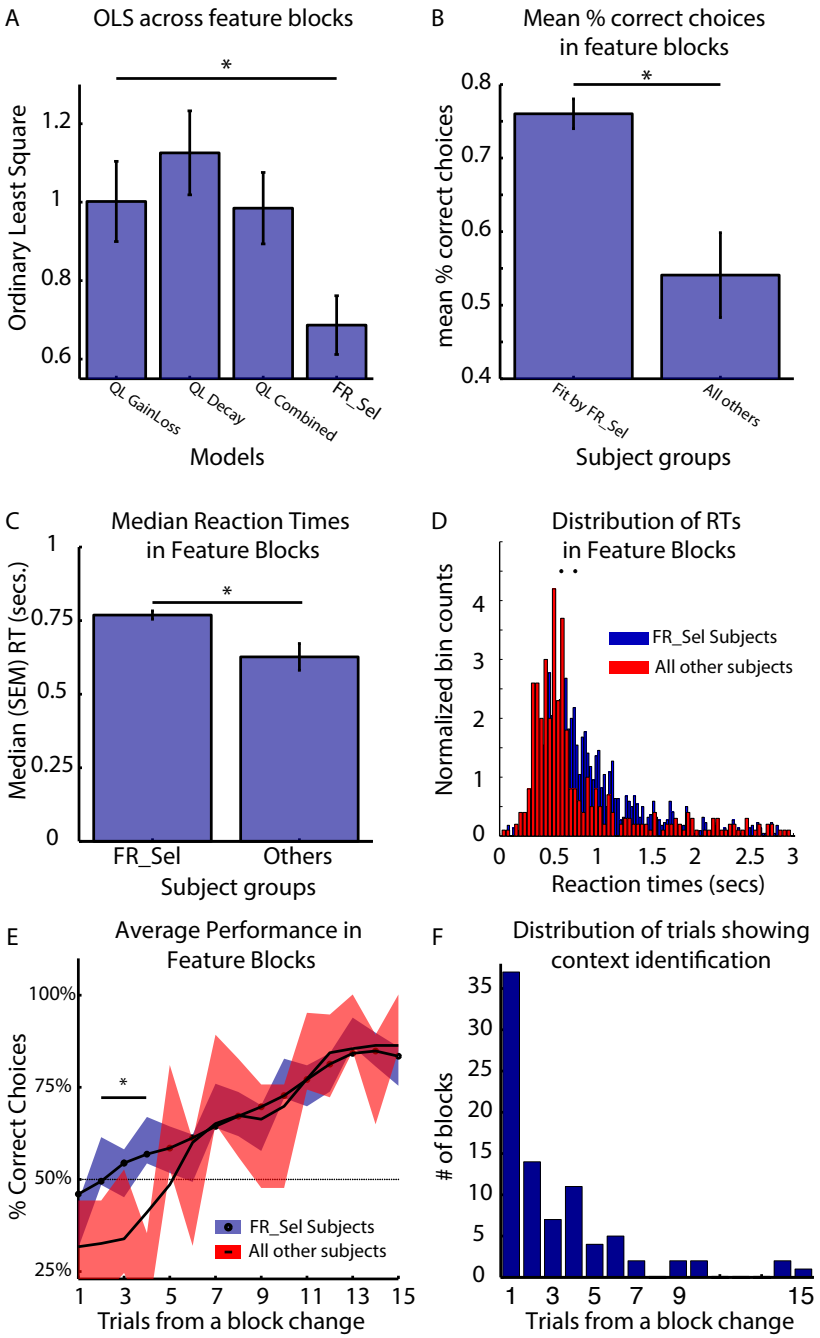


Figure 6. FR_Sel subjects outperform other subjects, react slower and learn faster. A) For subjects best fit by the FR_Sel model ('FR_Sel subjects'), FR_Sel model predictions in feature blocks was significantly better than all other models. (B) FR_Sel subjects (n=22) make on average more correct choices than subjects (n=10) with choice performance that was best fit by other models. C-D) FR_Sel subjects have significantly slower reaction times in feature blocks compared to all other subjects. (E) FR_Sel subjects show faster learning in feature blocks, having a significantly higher learning rate on early trials in a block compared to other subjects ($p<0.05$ Mann-Whitney-Wilcoxon). (F) Early responsiveness to the context of feature blocks is predicted by context identification mechanism of the FR_Sel model. The distribution of trials in feature blocks where the model identified the feature specific context is heavily weighted to the first five trials.

performance was split by block type we found that subjects were significantly better in object blocks at trials early in the block compared to feature blocks (**Fig. 2b**). On trials

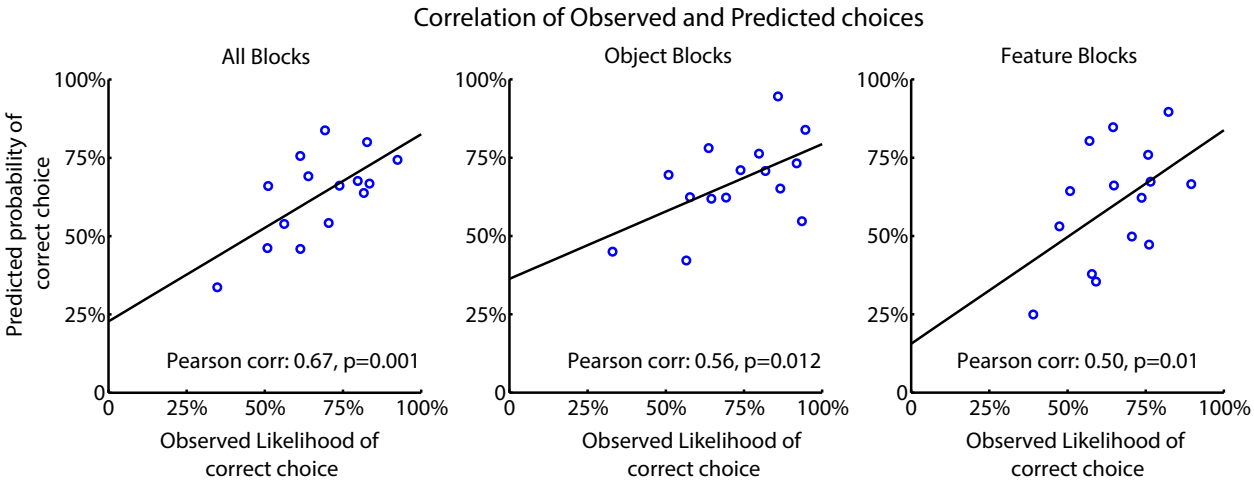


Figure 5. Model performance for FR_Sel across best fit subjects. The Pearson correlation was calculated between the mean observed choice likelihood and the predicted probability of making a correct choice based on the FR_Sel model. The FR_Sel model produces significantly correlated predictions across all blocks (left), as well as for object blocks (middle) and feature blocks (right panel).

nine and eleven, average performance in object blocks significantly exceeded that in feature blocks by 17.5, and 14.22% respectively ($p<0.05$ Mann-Whitney-Wilcoxon). However, at the end of the block, subjects performed equally well in object and feature blocks with identical proportions of 82.89% (SE ±.044) - 82.86% (SE ± .045) correct choices, respectively at trials 12-15 in a block.

To test the hypothesis that performance in a block of trials is related to a learning mechanism that is sensitive to block type, we performed a generalized linear model regression of the proportion of correct responses in a block against the block type sequence, i.e Object block 1, Feature Block 1, Object 2, etc. We found that this produced a significant ($p<0.05$) regression coefficient, allowing us to reject the null hypothesis, which indicates that performance is linked to the block type sequence. For

comparison, we also regressed the raw block order in a session (Block 1, 2, 3, etc.) against the proportion of correct responses in a block, and we found that this did not result in a significant coefficient ($p>0.05$), indicating the performance in a block is not simply a function of time or increased exposure to the task .

Models.

We considered a range of different learning strategies that could be deployed to solve the task through reinforcement learning mechanisms. Each of these strategies was quantified by a separate Q-Learning model (Rescorla, 1976; Cavanagh et al., 2010; Skvortsova et al., 2014) with different functionality representing different assumptions about: 1) the salience of positive versus negative feedback (QL GainLoss) (**Materials & Methods**, eqs. 4 & 5)(Gehring and Willoughby, 2002), 2) the impact of time and working memory capacity on learned values (QL Decay) (**Materials & Methods**, eq. 3) (Seymour et al., 2012; Skvortsova et al., 2014), 3) the generalization of outcome information across stimulus values (QL Gen) (**Materials & Methods**, eq. 2), and 4) the relevance of subsets of feature values for action selection (Flexible Rule Selection - FR_Sel) (**Materials & Methods**, eq. 7 & 8). The FR_Sel model was developed in order to capture the hypothesis that subjects would use Q-values for stimulus features to identify a rule for learning, in this case a rule that associates outcomes across trials by feature type and uses the difference in value between types to restrict selection to the most valuable type (**Fig. 3**).

In addition we tested three further models that were combinations of models 1-3.

All models were fit to subject data that showed learning by grid search across the entire parameter space (**Table 1,** see **Materials & Methods**)(Cavanagh et al., 2010;

Donoso et al., 2014; Skvortsova et al., 2014). Using the best fit parameters, we then bootstrapped 80% of the data 100 times to perform parametric statistics over the total OLS scores (see **Materials & Methods**). Because the average performance of the subjects differed across blocks separated by type, we calculated the OLS not only pooled across all blocks, but also for feature blocks and object blocks separately, which produced a final result for each optimized model in three dimensional OLS space.

Model Name	Number of parameters	mean (sem) alpha (alphaG)	mean (sem) alphaL	mean (sem) beta	mean (sem) threshold	mean (sem) decay (tau)	mean (sem) OLS distance across all subjects	number of subjects best fit by model	mean (sem) OLS distance across best fit subjects
QL Basic	2	.5523 (.0562)	n/a	.0051 (5.3445e-04)	n/a	n/a	5.5029 (.3397)	0	n/a
QL Generalized	2	.4795 (.0515)	n/a	.0053 (5.4342e-04)	n/a	n/a	6.0816 (.3943)	0	n/a
QL Decay	3	.5318 (.0541)	n/a	.0061 (5.3445e-04)	n/a	1.0401 (.0099)	1.8104 (.1754)	3	1.5453 (.2591)
QL GainLoss	3	.4568 (.0437)	.3614 (.0451)	.0047 (5.0114e-04)	n/a	n/a	1.6171 (.1610)	3	1.163 (.2578)
QL Combined	4	.3727 (.0360)	.3841 (.0550)	.0053 (5.2840e-04)	n/a	1.0918 (.0152)	1.6899 (.1544)	4	.5497 (.0957)
FR_Sel	4	.3727 (.046)	.3727 (.0520)	.0578 (0058)	.2441 (.0355)	n/a	1.1638 (.1276)	22	1.142 (.1644)
FR_Update	4	.2386 (.0233)	.4568 (.057)	.0365 (.0053)	.01 (7.9363e-19)	n/a	6.0995 (.3982)	0	n/a

Table 1. Model Names and best fitting parameter values along with measures of fit for individual subjects. Highlighted row shows scores and parameter values for the Flexible Rule-Selection (FR_Sel) model, which fit 22/32 subjects significantly better than any other subject (Mann-Whitney-Wilcoxon, $p<0.05$).

Models were compared by calculating the euclidean distance between the combined OLS score and the ideal score of zero. Four of the seven models considered had at least one subject that was best fit by the model, but the significant majority of subjects (68.75%, 22/32) was fit best by the Flexible Rule Selection (FR_Sel) model (pairwise comparison of bootstrapped OLS scores between all model pairs, Mann-Whitney-Wilcoxon ranksum $p<0.05$) (**Fig. 4a**).

The mean score for the FR_Sel model was significantly better than all other models averaged across all subjects (Mann-Whitney-Wilcoxon ranksum $p<0.05$) and was significantly better than all other models for those subjects that were best fit by the model when tested independently (Mann-Whitney-Wilcoxon ranksum $p<0.05$)(**Fig.4b,c**). We then quantified how the FR_Sel model scored for subjects whose choices were best

fit by one of the other models in order to infer whether these subjects may have used entirely distinct learning strategies, or whether the FR_Sel choice mechanism was still a versatile explanation for these subjects. As shown in Figure 4d,e,f we found that the FR_Sel model consistently provided the second best explanation for learning choice probabilities in those subjects best fit by the QL Decay model (n=4 subjects, 11%), the QL Combined model (n=4 subjects, 11%), and the QL GainLoss model (n=2 subjects, 5%) (**Fig. 4d,e,f**).

We tested how the choice probabilities produced by the FR_Sel model predicted the observed likelihood of subjects' correct choices. For this analysis we selected those subjects best fit by the model and computed the Pearson correlation of the average choice probabilities for the model and subjects for all trials, and for trials from feature type and object blocks separately (**Fig. 5**).

The FR_Sel model has a significant correlation with the observed data in all block groups ($r=0.6655$, $p=0.001$; $r=0.56$, $p=0.012$; $r=0.50$, $p=0.01$; Pearson correlation) showing that its computed choice probabilities are predictive of average subject behavior.

The key functional difference between the FR_Sel model and all other models is its ability to flexibly change selection strategies by restricting selection among Q-values to a specific feature domain when the history of choices provided sufficient information about feature type specific gains and losses (see **Materials & Methods**, **Fig. 3**). Accordingly, subjects using the FR_Sel mechanism (as inferred from being best fit by the FR_Sel model), should show improved performance particularly when transitioning into feature blocks over subjects utilizing other selection strategies (as inferred from

being best fit by one of the other models). In support of this suggestion, we found that the FR_Sel model produced the best fit to subject data in feature blocks across all those subjects best fit by FR_Sel according to their overall OLS score (**Fig. 6a**).

Calculating the mean percent correct choices in feature blocks shows that FR_Sel subjects outperformed all other subjects ($p<0.05$, Mann-Whitney-Wilcoxon) (**Fig. 6b**). They also showed significantly slower reaction times ($p<0.05$, Mann-Whitney-Wilcoxon) (**Fig. 6c**). Across subjects the median reaction times of subjects did not correlate with the mean % correct choices of subjects ($r=0.044$, $p=0.366$). Examining the dynamics of subject performance in feature blocks also showed a significant difference across subject groups. FR_Sel subjects show more rapid learning early in the block, with significantly better performance until trial five (for trial numbers two to five, $p<0.05$, Mann-Whitney-Wilcoxon) (**Fig. 6e**). Faster learning early in the block is another implication of the functionality of the FR_Sel model. Rule deployment specifies context specific selection processes, and this context specific selection, ie. selection that is restricted to a specific feature domain, is triggered when the difference in values between feature types crossed a threshold. For all subjects best fit by the FR_Sel model this threshold value was quite low (0.21), indicating that very few trials were required to separate values between feature types. With the model generalizing outcome information across chosen and unchosen features, summed values across feature types rapidly diverge. We identified the trials in feature blocks when the FR_Sel model triggered feature specific selection, and plotted the distribution of these trials across blocks (**Fig. 6f**). The model identifies the relevant feature type in the current context rapidly with an average (median) of 2 (SE ± 0.4) trials and with the majority of blocks

being identified within the first five trials, which is consistent with the rapid learning early in the block observed in the subject performance.

3.5 Discussion

In this study we tested subjects on their ability to flexibly apply a previously learned abstract rule, respond to uncued context changes, and learn stimulus-feature outcome associations. We developed a set of predictive behavioural models using the reinforcement learning framework, which allowed us to fit the choices of each subject to a unique model, separating subjects that utilize advantageous rule-driven behaviour from those that do not. We found that two-thirds of subjects (n=22/32), who were untrained on the task and naive to its design, utilized a strategy for learning that reflected the application of a pre-learned abstract rule relating the association of stimulus feature dimensions to positive outcomes. Importantly, the subjects best fit by the hierarchical rule model were also the subjects that performed the best in more difficult feature blocks, and displayed a significantly slower reaction time on choices in those blocks. Previous studies exploring rule learning and rule driven behavior have focused on either how simple rules are learned via reinforcement, or on how rules can be learned and generalized for application in new contexts. Our study extends this work by quantifying how successful subjects who are naive to the task spontaneously utilize pre-learned task rules to learn in a novel task context.

Rule learning and switching has been studied extensively, typically with a framework such as the Wisconsin Card Sorting Task (WCST) and its analogues (Grant and Berg, 1948; Milner, 1963; Wallis et al., 2001; Buckley et al., 2009; Badre et al.,

2010; Mian et al., 2014). In the WCST, four key cards provide the subject with different cues about potential sorting principles across three perceptual dimensions - colour, shape and number. Subjects attempt to correctly sort the 128 responses cards one at a time, according to the unknown rule, via feedback in the form of binary outcome information (correct vs. incorrect). In our study, we were interested in the flexible application of a more abstract rule, where the rule is informative of a general principle but does not specify the final mapping of a stimulus feature to outcomes, as in the WCST (Bengtsson et al., 2009; Collins and Frank, 2013). Similar to the WCST, the optimal rule to apply in feature blocks is to associate outcomes across trials with a specific feature dimension of the stimulus, however, in our task subjects applying this rule must still learn the specific likelihoods of reward associated with the set of stimulus features presented in that block. For example, after identifying the current context (block) as a colour-relevant block, the subject must then learn the rank ordering, or the relative likelihoods of reward, associated with the three colours that appear in that context. Whereas in the WCST, once the stimulus feature sorting rule is learned subjects only need to maintain this rule until it is switched (Stuss et al., 2000; Buckley et al., 2009; Nyhus and Barceló, 2009).

Recently there has been some exploration of how rules are learned and generalized to new contexts. Collins et. al. (Collins and Frank, 2013) have shown that subjects are capable of learning rules for task set organization and generalize these rules into new contexts, even when applying a particular rule is not beneficial. Our results are consistent with these findings, with the difference being that those subjects in our study who spontaneously displayed beneficial rule-guided behavior learned the

abstract rule prior to the task. Collins et. al also used hierarchical RL methods to quantify the computational processes associated with abstract rule learning and generalization. Similarly to their approach and that of others in the field (Badre and Frank, 2012; Botvinick, 2012; Donoso et al., 2014), we developed a hierarchical rule selection system that relies on simple model free learning of expected outcomes for stimuli and stimulus features. A model-based approach did not seem appropriate here as rewards were assigned to stimuli stochastically, and transitions between block types were jittered and uncued, all of which prevented subjects from anticipating the likelihood of transitions between states (trials and block types), which is a key functionality of model-based systems.

Many studies of human decision making analyze patterns of choice behaviour that collapses subjects into a single unit. This is often done in order to perform analyses of neural activity that averages results across subjects (Cavanagh et al., 2010; Helfinstein et al., 2014; Rudorf and Hare, 2014). While this approach has the benefit of increasing the statistical power of certain techniques it is insensitive to inter-subject variability. Analyzing and fitting models to the systematic behaviour of each subject, as we have done here, permits the identification of those patterns of choices that are related to the unique strategy of each subject. It is no surprise to experimentalists that human subjects bring a range of pre-task experiences and expectations to bear on the the experimental problem (Shteingart and Loewenstein, 2014), but this is notoriously difficult to account for, and is often just ignored. By using a range of models, each with an associated set of conceptual assumptions, we are able to separate subjects by their flexible application of adaptive rules. While we found that subjects best fit by our flexible

rule-selection model where also separable from other subjects according to overall block performance and reactions times, we do not have a hypothesis about why reaction times for these subjects are significantly slower than for other subjects in feature blocks. Further work in this area is needed to clarify the relationship between advantageous use of a flexible rule and reaction times.

Neural activity associated with rule-driven behaviour has been found in the prefrontal cortex of humans and non-human primates (Miller, 2000; Buschman et al., 2012; Bengtsson et al., 2009). Based on the similarity of our results to that of previous studies (Gershman and Niv, 2010; Collins et al., 2014), we would expect to see activity profiles in FR_Sel subjects similar to that found in Collins et. al. (Collins et al., 2014) and Cavanagh et. al. (Cavanagh et al., 2010). Because our computational approach produces trial by trial, and subject by subject, estimates of expected values for stimulus features, as well as estimates of trial onsets for rule deployment, simultaneous recording of neural activity in human subjects performing our task would enable sensitive and specific insights into the networks underlying rule deployment and feature value learning. Single trial regression analysis are an underused but powerful tool for investigating the neural mechanisms underlying computational processes implicated in human learning because they compensate for inter-subject variability. Further work on the problem of learning and the ongoing influence of prior learning would likely link lateral PFC areas known to be involved with rule learning and switching to ventromedial PFC areas known to be involved with estimates of stimulus and action values (Wallis et al., 2001; Bengtsson et al., 2009; Buckley et al., 2009; Badre et al., 2010; Wunderlich et al., 2010; Gershman et al., 2010a; Mian et al., 2014; Rudorf and Hare, 2014).

3.6 References

Anderson BA, Yantis S (2013) Persistence of value-driven attentional capture. Journal of Experimental Psychology: Human Perception & Performance 39:6–9.

Badre D, Frank MJ (2012) Mechanisms of Hierarchical Reinforcement Learning in Cortico-Striatal Circuits 2: Evidence from fMRI. Cerebral Cortex 22:527–536.

Badre D, Kayser AS, D'Esposito M (2010) Frontal cortex and the discovery of abstract action rules. Neuron 66:315–326.

Bengtsson SL, Haynes J-D, Sakai K, Buckley MJ, Passingham RE (2009) The representation of abstract task rules in the human prefrontal cortex. Cerebral Cortex 19:1929–1936.

Bergstra, James, and Yoshua Bengio (2012) Random search for hyper-parameter optimization. The Journal of Machine Learning Research 13.1: 281-305.

Botvinick MM (2012) Hierarchical reinforcement learning and decision making. Current Opinion in Neurobiology 22:956–962

Botvinick MM, Niv Y, Barto AC (2009) Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. Cognition 113:262–280.

Buckley MJ, Mansouri FA, Hoda H, Mahboubi M, Browning PGF, Kwok SC, Phillips A, Tanaka K (2009) Dissociable components of rule-guided behavior depend on distinct medial and prefrontal regions. Science 325:52–58.

Buschman TJ, Denovellis EL, Diogo C, Bullock D, Miller EK (2012) Synchronous Oscillatory Neural Ensembles for Rules in the Prefrontal Cortex. Neuron 76:838–846.

Cavanagh JF, Frank MJ, Klein TJ, Allen JJB (2010) Frontal theta links prediction errors to behavioral adaptation in reinforcement learning. NeuroImage 49:3198–3209.

Chumbley JR, Flandin G, Bach DR, Daunizeau J, Fehr E, Dolan RJ, Friston KJ (2012) Learning and generalization under ambiguity: an fMRI study. PLoS Comput Biol 8:e1002346.

Collins AGE, Cavanagh JF, Frank MJ (2014) Human EEG uncovers latent generalizable rule structure during learning. J Neurosci 34:4677–4685.

Collins AGE, Frank MJ (2013) Cognitive control over learning: creating, clustering, and generalizing task-set structure. Psychological Review 120:190–229.

Donoso M, Collins AGE, Koechlin E (2014) Human cognition. Foundations of human reasoning in the prefrontal cortex. Science 344:1481–1486.

Friston K, Schwartenbeck P, FitzGerald T, Moutoussis M, Behrens T, Dolan RJ (2014) The anatomy of choice: dopamine and decision-making. Philos Trans R Soc Lond, B, Biol Sci 369:20130481–20130481.

Gehring WJ, Willoughby AR (2002) The medial frontal cortex and the rapid processing of monetary gains and losses. Science 295:2279–2282.

Gershman SJ, Blei DM, Niv Y (2010a) Context, learning, and extinction. Psychological Review 117:197–209.

Gershman SJ, Cohen JD, Niv Y (2010b) Learning to selectively attend. 32nd Annual Conference of the Cognitive ….

Gershman SJ, Niv Y (2010) Learning latent structure: carving nature at its joints. Current Opinion in Neurobiology 20:251–256.

Grant DA, Berg E (1948) A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. Journal of Experimental Psychology 38:404–411.

Helfinstein SM, Schonberg T, Congdon E, Karlsgodt KH, Mumford JA, Sabb FW, Cannon TD, London ED, Bilder RM, Poldrack RA (2014) Predicting risky choices from brain activity patterns. PNAS.

Mian MK, Sheth SA, Patel SR, Spiliopoulos K, Eskandar EN, Williams ZM (2014) Encoding of rules by neurons in the human dorsolateral prefrontal cortex. Cerebral Cortex 24:807–816.

Miller EK (2000) The prefrontal cortex and cognitive control. Nat Rev Neurosci 1:59–65.

Milner B (1963) Effects of Different Brain Lesions on Card Sorting. Arch Neurol 9:90–100.

Nyhus E, Barceló F (2009) The Wisconsin Card Sorting Test and the cognitive assessment of prefrontal executive functions: A critical update. Brain and Cognition 71:437–451.

Rescorla RA (1976) Stimulus generalization: some predictions from a model of Pavlovian conditioning. J Exp Psychol Anim Behav Process 2:88–96.

Rudorf S, Hare TA (2014) Interactions between dorsolateral and ventromedial prefrontal cortex underlie context-dependent stimulus valuation in goal-directed choice. J Neurosci 34:15988–15996.

Seger CA, Peterson EJ (2013) Categorization = decision making + generalization. Neurosci Biobehav Rev 37:1187–1200.

Seymour B, Daw ND, Roiser JP, Dayan P, Dolan R (2012) Serotonin selectively modulates reward value in human decision-making. Journal of Neuroscience 32:5833–5842.

Shteingart H, Loewenstein Y (2014) Reinforcement learning and human behavior. Current Opinion in Neurobiology 25:93–98.

Skvortsova V, Palminteri S, Pessiglione M (2014) Learning To Minimize Efforts versus Maximizing Rewards: Computational Principles and Neural Correlates. J Neurosci 34:15621–15630.

Stuss DT, Levine B, Alexander MP, Hong J, Palumbo C, Hamer L, Murphy KJ, Izukawa D (2000) Wisconsin Card Sorting Test performance in patients with focal frontal

and posterior brain damage: effects of lesion location and test structure on separable cognitive processes. Neuropsychologia 38:388–402.

Wallis JD, Anderson KC, Miller EK (2001) Single neurons in prefrontal cortex encode abstract rules. Nature 411:953–956.

Wilson RC, Niv Y (2011) Inferring relevance in a changing world. frontiers in Human Neuroscience 5:189.

Wunderlich K, Rangel A, O'Doherty JP (2010) Economic choices can be made using only stimulus values. PNAS 107:15005–15010.

Zucchini W (2000) An Introduction to Model Selection. Journal of Mathematical Psychology 44:41–61.

Chapter 4.

Estimates of expected value of stimuli are correlated with theta band power in human ventromedial pre-frontal cortex

Matthew Balcarras, Cristiano Micheli, and Thilo Womelsdorf

Department of Biology, Centre for Vision Research, York University, 4700 Keele Street, Toronto, Ontario, M6J 1P3, Canada.

Correspondence: Dr. Thilo Womelsdorf, Mr. Matthew Balcarras.
York University, Department of Biology, 4700 Keele Street,
Toronto ON M3J 1P3, Canada

4.1 Abstract

Making a choice between two options can be difficult. One way to make a good choice between environmental stimuli is to compare the expected value of both stimuli. However, the relationship between the learning and comparison of expected values for stimuli and their neuronal representation in band-limited, rhythmic neuronal activity in the human brain is unknown. Previous studies using functional magnetic resonance imaging (fMRI) have suggested that activation in ventromedial prefrontal cortex (vmPFC) is related to the difference in value between two available stimuli. Here, we reveal in a proof-of-principle study using intracranial electrocorticography (ECoG) in human subjects that band-limited power changes in the vmPFC related to the expected values of stimuli is predicted by a reinforcement learning model for a single human subject. We found that during the formation of a decision about which of two stimuli is more likely to lead to a rewarding outcome - before any overt choice is made - the trial-by-trial difference in expected value between available stimuli was significantly correlated with normalized power in the theta frequency band (4-8Hz). This finding suggests that the circuit dynamics underlying theta band activity in vmPFC carry significant information about the ongoing prediction of stimulus values. We speculate that this finding constitutes a strong proof-of-principle that flexible learning can be be traced to those rhythmic activity signatures which have previously been implicated in long-range networks underlying value-based decision making.

4.2 Introduction

Learning how to improve choices among available options is an essential skill for biological fitness. Choices between environmental stimuli can be improved by learning associations between stimuli and outcomes in order to estimate their long-run worth or value. The Reinforcement Learning (RL) framework has proven to be a powerful tool for providing time sensitive predictions of the internal, subjective value of stimuli and the mechanisms involved in making value-based choices (Niv et al., 2005; Dayan and Daw, 2008). RL provides a formal account of how choices are improved through the incremental learning of values for choice variables by using outcomes to update expectations. By using these values to make comparisons between multiple options, the most adaptive choice can be made.

The RL framework contains separable processes that have been correlated with signals of brain activity in different brain regions (Rangel et al., 2008; Rushworth et al., 2011). Stimulus value signals have been identified in ventromedial prefrontal cortex (vmPFC) and in orbital frontal cortex (OFC) (Rangel et al., 2008; Rushworth et al., 2011; Rudorf and Hare, 2014; Skvortsova et al., 2014), and activity relating to the difference in value between available stimuli has also been identified in vmPFC (Rudorf and Hare, 2014). It is unclear however, how BOLD activations in these regions of the brain are related to the oscillatory processes known to underly inter-areal interactions and integrative brain function. Rhythmic oscillations reflect dynamic changes in the excitability of local neuronal groups and oscillations in frequency specific bands have been identified as playing different communicative and computational roles in the brain (Womelsdorf et al., 2007; Buzsáki and Watson, 2012). In our study we set out to test whether frequency specific oscillations are linked to the neuronal processing of stimulus

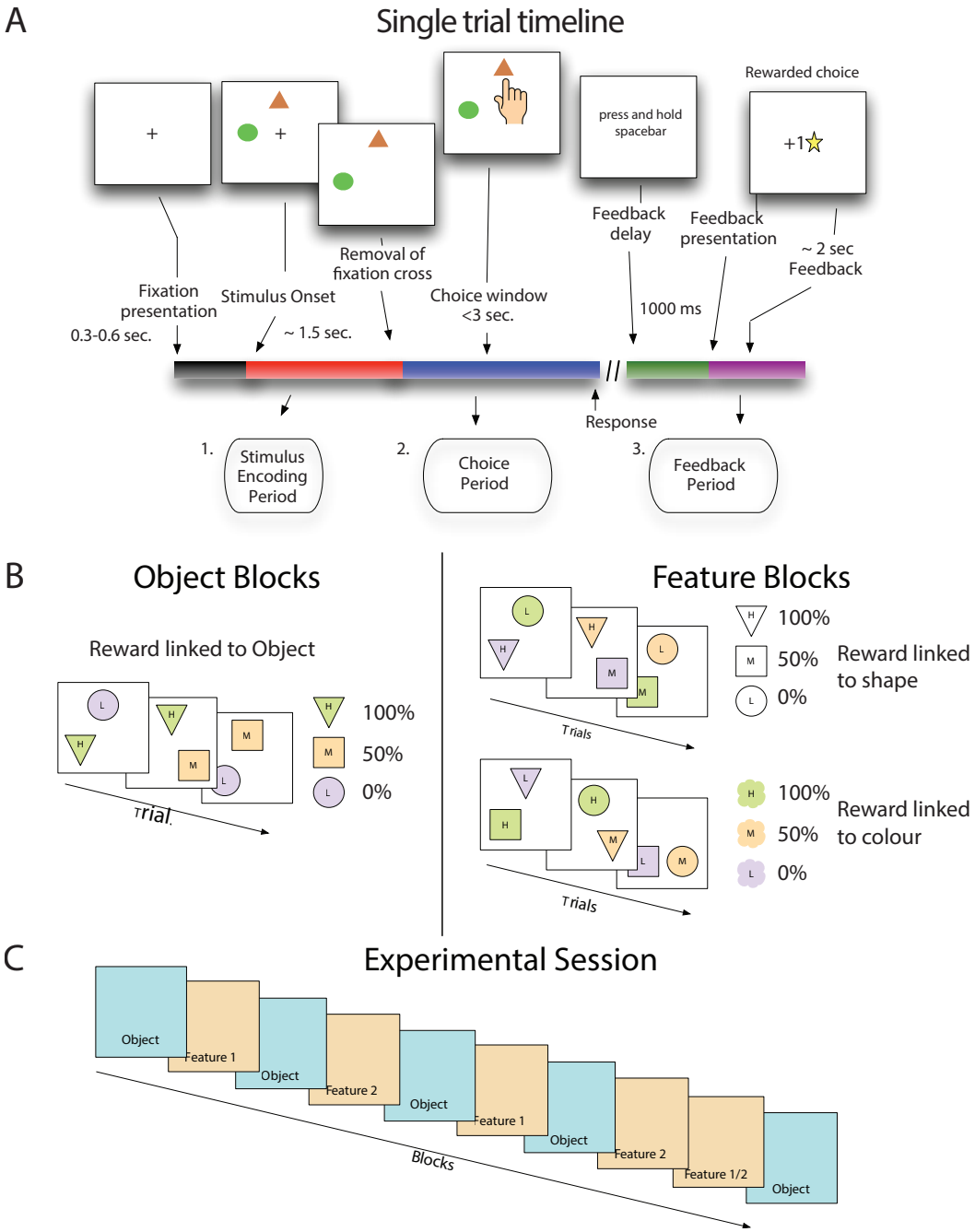


Figure 1. Stimulus value learning task. A) Subjects learned by trial and error that stimuli and stimulus features are linked to the likelihood of receiving positive outcomes. In the displays, the red 'x' denotes the chosen stimulus of subjects. The yellow starts on top of each panel indicated the feedback for correctly chosen stimuli. The right panel vertical summarizes the choice outcomes for trials shown on the left to illustrate the subjects putative internal state for determining selections on future trials. B) Stimulus reward associations were structured either such that fixed pairs of colours and shapes (in sets of three) had a probabilistic relationship with reward (object blocks) or such that stimulus features were not fixed to each other and only one feature type (either shape or colour) was linked to reward. C) For the first eight blocks, feature blocks followed object blocks using the same set of shapes and colours as the preceding object block, but with new feature-reward associations. The last pair of blocks flipped this pattern where a feature reward block (either type 1 or 2, randomly selected) precedes an object block.

value information.

Low-frequency theta band (4-8Hz) activity has been implicated in the processing of value expectations of relevant stimuli in cognitive control tasks, working memory tasks and tasks with predictable reward schedules (Başar et al., 2001; Lee et al., 2005; Tsujimoto et al., 2006; Womelsdorf et al., 2010a). Theta oscillations are highly correlated with learning (Benchenane et al., 2010) and emerge specifically at so-called decision points when subjects compare stimulus and action values to inform choices (Womelsdorf et al., 2010b). Additionally, theta oscillations are associated with stimulus selection rules (Womelsdorf et al., 2010a), carry stimulus specific information in visual cortex (Lee et al., 2005), and modulates gamma band activity as part of inter-areal coupling between memory and attentional processes (Womelsdorf and Fries, 2007; Canolty and Knight, 2010; Bosman et al., 2012). In human subjects theta activity has been reported to increase with cognitive demands and working memory load (Jensen and Tesche, 2002; Brincat and Miller, 2015). Most importantly, working memory function is related to rule-guided behaviour (Amso et al., 2014) and therefore it is likely that the neural mechanisms underlying working memory are also involved in the computations that integrate learned values into selection rules.

We designed a task where subjects need to learn information about stimulus-reward associations across time and integrate this information to select between competing options. We hypothesized that the strength of low frequency oscillations, specifically in the theta band (4-8Hz), is directly related to information about the relative difference in expected stimulus values between available stimuli. To test this hypothesis we tested a human subject with electrodes implanted subdurally at the medial wall of

the vmPFC. We then fit the subjects' behaviour with Q-Learning RL models that provided the best trial-by-trial predictions about the expected value of stimuli. We found that the difference in the predicted stimulus values, which is the key decision variable underlying adaptive choices, was correlated with theta band specific neuronal activity measured with ECoG electrodes near the vmPFC.

4.3 Materials & Methods

Task design

All experimental procedures were approved by York University's Ethics Review Board (see **Appendix - B**). Participants performed the experiment on a touch sensitive Sony Vaio laptop running Windows 8, and Matlab (The Mathworks Inc.) with the Psychophysics toolbox (www.mathworks.com; www.psychtoolbox.org) and custom written Matlab scripts controlling the experiment. The laptop had an 15" capacitive touch sensitive monitor with a resolution of 1920x1080 pixels and a refresh rate of 60 Hz. Stimuli were placed at 4.6 degrees from the central fixation point. The laptop was positioned comfortably, ~50-70cm, in front of subjects to ease their holding and touching responses. The temporal resolution of the touchscreen responses were on the order of 1000 milliseconds (± 30 msec. SEM). At the start of the experiment, participants were instructed to use the index finger of their dominant hand to touch one of the two presented stimuli, then use the same finger to hold the spacebar to receive feedback, and to make choices that maximized the number of positive feedbacks.

A trial began with the presentation of a small cross in the centre of the screen (**Fig. 1A**). After 300-600 milliseconds two stimuli appeared in two of three possible positions.

The location of stimuli was randomly chosen from canonical locations equidistant from each other and the central cross. After another 1500 milliseconds the central cross was removed and subjects were free to select a stimulus. If subjects selected a target before the removal of the fixation cross, the stimuli were removed and a message was displayed reminding the subject to wait for the removal of the cross. This message was displayed for a waiting period of 500 milliseconds before a new trial began. Following the selection of a stimulus, the stimuli were removed and a message appeared on the screen informing subjects to hold the spacebar in order to receive feedback. Feedback was not given until the spacebar was depressed for 1000 milliseconds, and was either a gold star in the middle of the screen or a message saying 'sorry' when the schedule associated with that stimulus determined it was either a rewarded or an unrewarded trial (see below). Gold stars awarded to the subject accumulated at the bottom of the screen, indicating to the subjects their performance thus far. After the last trial of the session was completed, a screen was displayed which thanked the subject for participation and provided a final count of gold stars received.

Subjects made choices on stimuli that were combinations of shapes and colours. Each object block began with a new set of three shapes and three colours drawn from a set of six, and all stimuli for that block were made from combinations of these three shapes and colours (**Fig. 1B**). In object blocks, shape-colour pairs remain fixed throughout the block so that there is only thee unique stimuli appearing in the block. Feature blocks that followed object blocks used the same set of shapes and colours that appeared in the previous block, but now stimuli could be composed of any combination

of colour and shape, so that there were nine possible unique stimuli appearing in the block.

The task included a hidden probabilistic reward schedule that assigned a probability of positive outcome on each trial to the two available stimuli (**Fig. 1B** - right panels). In object blocks each stimuli, a unique colour-shape pair, is assigned a probability of positive outcome, with one being 0%, one 50% and one 100%. In feature blocks, outcome probabilities are associated with a specific feature dimension, either shape or colour. In a colour-feature block, one colour is predictive of positive outcomes 0%, one 50%, and one 100%. Shape-feature blocks work the same as colour feature blocks except that probabilities are linked to stimulus shape instead of colour. In feature blocks, the non-relevant feature is only spuriously related to outcomes because of the randomized relationship between colours and shapes in these blocks. Receiving a positive outcome for a choice on colour A and shape B in a shape-feature block will not tell you anything about the likelihood of receiving a positive outcome on the next trial where colour A appears. In both feature and object blocks, stimulus location was selected randomly and was never predictive of outcomes.

Subjects performed a stereotypical sequence of blocks (**Fig. 1C**). An experimental session began with an object block, followed by a feature block, where the relevant feature was selected at random, then another object block and feature block, where this feature is the alternate one from the first feature block. This sequence continued until the ninth block which reversed the object-feature order, and the relevant feature was randomly selected, with the final block being an object block.

Blocks ranged in length from 15-25 trials with the number of trials in a block determined by a performance criterion. If the subject had made 11 out of the first 15 choices correctly then the block ended at trial 15. Trials continued until either 80% of the last 10 trials were correct or the subject reached 25 trials. Average block length across subjects was 17.8 (SE ±2.1). In total subjects performed 7106 trials, of which 3964 trials that were from blocks showing learning were included in analysis.

Behavioral Data Analysis.

Data Analysis was done with custom written Matlab scripts (The Mathworks Inc.). Learning in a block was determined following the method of Wilson and Niv (Wilson and Niv, 2011), whereby if the slope of the average performance line from the beginning to the end of the block was positive and was above chance performance (50% correct) at the end, the block was considered to show learning.

Correct choices were determined by the selection of the stimulus with the higher probability of a positive outcome, independently of whether a positive outcome was received. On trials where the 100% likely stimulus appeared, it was always the correct stimulus to select, even if selecting the 50% likely stimulus produced a positive outcome. Likewise on trials where the 0% likely stimulus appeared, it was always the incorrect stimulus to select, even if selecting the 50% likely stimulus did not produce a positive outcome.

Reaction times were quantified from the time when the fixation cross was removed until the time when the screen was touched. If the subject touched the screen before the fixation cross was removed the trial was declared an ‘early response’ and was not included in further analysis.

RL model algorithms.

In the basic Q-Learning Rescorla Wagner RL model (QL Basic), the value of any predictor of reward (stimulus feature, Q_i) is updated on the next time step (trial) from its previous value through the scaled reward-prediction error: The difference between the binary reward outcome (R, either 0 or 1) and the predictor itself(Skvortsova et al., 2014). The scaling factor (α) represents the learning rate:

$$Q_i(t+1) = Q_i(t) + \alpha[R(t) - Q_i(t)] \quad \text{(eq. 1)}$$

Other than the QL Basic model, all other models implemented a generalization of outcome information across all Q values. Thus, all stimulus features associated with the selected stimulus updated their value according to equation 1. Stimulus features associated with the other, non-selected stimulus were updated according to:

$$Q_i(t+1) = Q_i(t) + \alpha[1 - R(t) - Q_i(t)] \quad \text{(eq. 2)}$$

The second model, QL Gen, extended QL Basic with generalization of outcome information across all Qs for features appearing on that trial and no other changes. In the third model, QL Decay, feature values were updated when they were associated with the selected stimulus features in the same way as QL Basic and QL Gen, but all non-selected features had their associated values decay as a function of time governed by the rate of decay (τ) according to:

$$Q_i\ (t+1) = Q_i(t)\ + \alpha[1 - R(t) - Q_i(t)\ ] * \tau \qquad \text{(eq. 3)}$$

The fourth model, QL GainLoss, employed the same framework as QL Gen, but applied a different learning rate to positive and negative outcomes - α_G vs α_L.

$$Q_i\ (t+1) = Q_i(t)\ + \alpha_G[\ R(t) - Q_i(t)\ ] \qquad \text{(eq. 4)}$$

$$Q_i\ (t+1) = Q_i(t)\ + \alpha_L[1 - R(t) - Q_i(t)\ ] \qquad \text{(eq. 5)}$$

Stimulus feature values for all non-HRL models were non-linearly transformed into choice probabilities according to the Boltzmann equation:

$$P_i(t) = {}_e\beta Qi(t) / \Sigma_e \beta Qj(t) \qquad \text{(eq. 6)}$$

where β represents the inverse temperature and establishes the strength of the non-linearity.

The *Flexible Rule Selection* model (FR_Sel) employs a selection function that is an adaptation of the standard Boltzman formulation. Rather than all available Qs competing for final selection via participating as possible choice probabilities, FR_Sel compares Q values across features by feature type, calculating the difference between the sum of total values for each type. When the difference between the total value for one feature type relative to the other types moves past a threshold (λ), only that set of values is used to compute choice probabilities according to the equations below:

$$P_i(t) = {}_e\beta Qi(t) / \Sigma_e \beta Qsel(t) \quad \text{(eq. 7)}$$

where Q_{sel} is the set of Qs such that:

$$Q_{sel} > \Sigma\, Q_{others} + \lambda \quad \text{(eq. 8)}$$

Model Optimization.

Models were optimized by performing a grid search across the total parameter space for each free parameter, attempting to minimize the ordinary least square distance between the probability associated with selecting the correct stimulus and the observed likelihood of selecting the correct stimulus (Donoso et al., 2014)(Bergstra, 2012). On each trial the model was given the choice made by a subject and transformed that into values according the learning rate(s) of that model iteration. Values were converted into choice probabilities according the Boltzman equation and the value of beta (Glimcher, 2011). The mean probability associated with the correct choice was calculated for each trial from the block start across all blocks. Values for free parameters were selected that minimized the distance between this mean probability and the mean likelihood of the subject making a correct choice.

To ensure that we fit the models to the most systematic behaviour, we bootstrapped 80% of the data from each subject 100 times for each set of parameter values, and calculated the mean OLS score across these 100 iterations. Bootstrapping is a known method of estimating the variance of model performance (Zucchini, 2000).

We did not use statistical methods for model comparison, such as the Akaike or Bayesian Information Criterion, because 1) other studies have shown that using OLS is

equally capable of identifying the best model (Donoso et al., 2014), and 2) we fit the models to subject performance split by block type, which essentially creates two datasets, and information criterion scores are not comparable across datasets (Zucchini, 2000).

Data acquisition and preprocessing.

For data acquisition patient connectors were transferred to a separate research amplifier system (NeuroScan SynAmps2 data acquisition system; Compumedics, Charlotte, NC, USA). Data were acquired at 5kHz (0.3-1kHz band pass prior to digitization (SynAmps2; Compumedics Neuroscan USA Ltd., Charlotte, NC, USA). Visual stimuli were delivered Psychophysics Toolbox and Matlab on a Sony Vaio laptop with Windows 8. Eye movements were monitored with two electrodes over and near the eyes of the subjects. Electrooculogram channels were acquired at 5kHz (0.3-1kHz band pass prior to digitization.

Experimental Subject.

Our subject was affected by medically intractable epilepsy and underwent surgical implantation of subdural strip electrodes (PMT, MN, USA) to localize epileptogenic regions. A 4-contact subgaleal strip electrode (PMT, Chanhassen, MN, USA) electrode over the parietal midline and facing away from the brain was used for ground and reference. The subject was not on any medication for seizure control or pain relief during data collection. All procedures of the study followed the Good Clinical Practice procedures of Toronto Western Hospital, and were approved by the University Health Network Research Ethics Board. Informed consent was obtained prior to the recordings

(see **Appendix - C**). The subject was attentive and collaborative and had corrected-to-normal vision.

ECoG Data Analyses.

The data were analyzed offline with custom Matlab scripts using functions from the FieldTrip open-source toolbox (http://fieldtrip.fcdonders.nl/; Oostenveld et al., 2011). The multichannel experimental data were split into single channel files and resampled at 1kHz. The data were cleaned from possible line noise artifacts with a DFT filter at 60 Hz and higher harmonics (120 Hz and 180 Hz), lowpass filtered at 300 Hz with a 4th order forward-reverse Butterworth filter and baseline corrected (i.e. demeaned, see Womelsdorf et al., 2006 and http://fieldtrip.fcdonders.nl/faq/why_does_my_tfr_look_strange). We performed a visual artifact rejection using the graphical interface tool 'rejectvisual' and 'databrowser' from the Fieldtrip toolbox. Subsequently, the bad trials containing spike activity were manually excluded and, following artifact removal (see **Appendix E - Chapter 4 Supplementary Figures**), the datasets were common average re-referenced and each channel's amplitude variation was z-scored to the session's standard deviation before frequency analysis commenced. Mean and std of the z-scores are calculated for each session using the time epoch when no stimulus is visible (baseline epoch).

We selected twelve channels from the total 92 available by identifying the closest channels to previously identified ROI. The data were cleaned from possible line noise artifacts with a DFT filter at 60 Hz and higher harmonics (120 Hz and 180 Hz), lowpass filtered at 300 Hz with a 4th order forward-reverse Butterworth filter and baseline corrected.

Spectral analysis was performed by Fourier analysis applied to 0.5 sec. time windows centred on the middle of stimulus onset period. Data were tapered using dpss tapers (discrete prolate spheroidal sequences) and +/- 4 Hz frequency bandwidth (corresponding to 3 tapers). For initial inspection we ran a time frequency analysis using 0.5 sec. windows and overlapping sliding windows with a 0.1 sec. step size.

Baseline normalization.

Power spectra for the decision period on each trial were normalized relative to baseline activity - one second pre-stimulus onset - by z-scoring. Power values from during the decision period had the baseline mean subtracted, and the result was divided by the standard deviation of activity in the period.

Anatomical preprocessing and electrodes labeling.

We used computed tomography (CT) and magnetic resonance (MR) imaging scans for anatomical reconstructions following standard clinical routines. The MRs used axial 3 Tesla, 3D FSPGR sequences with TR/TE 7.88/3.06, and with 1 mm thick slices. In addition, MR images were acquired with the following sequences: axial T2 FLAIR, coronal FSTIR, axial FSTIR, sagittal T1 FLAIR. Single subject CT to MR realignment, co-registration, cortical segmentation and MNI space normalization were performed on the subject's brain anatomy using SPM8 (statistical parametric mapping; http://www.fil.ion.ucl.ac.uk/spm/software/spm8/) and custom scripts. A manual procedure was carried out to mark the electrodes in the co-registered CT space. Electrode positions were projected onto the pre-surgical MR anatomy, with the help of CTMR software and custom scripts, to partially compensate for the surgical brain displacement. The MNI template 'mni_icbm152_t1_tal_nlin_asym_09c' was used because of its compatibility

with the default anatomical template of SPM8, upon which single subject brain MNI normalization was based. The cortical sheet of the template was extracted by means of a custom Matlab script and the CTMR package (Hermes et al., 2010).

Correlation analysis.

We quantified the relationship of trial-by-trial prediction of Q-value differences by summing the values associated with all features of a stimulus on each trial, then took the difference between the two sums for the stimuli that appeared.

We extracted spectral power from 1-30 Hz across all trials and for all electrodes. We used the median of this activity on each trial to compare with Q-value differences. We computed the Pearson correlation across all trials between the difference in Q-values for available stimuli and the median power for each frequency.

4.4 Results

We devised a learning task for human subjects that provided probabilistic feedback following choices between two visual objects on a touch screen computer (see **Materials & Methods**). By using information from outcomes, either a gold star or a 'sorry', it is possible for subjects to learn the correct choice between available stimuli with different associated probabilities of positive outcome. A subject (female, 25 years of age, right handed) undergoing surgery for relief of pharmacologically resistant epilepsy performed the task five times for a total of 50 blocks and 938 trials. Using a simple criterion for learning across a block (Wilson and Niv, 2011), we determined that the subject showed learning in 37/50 blocks, with the proportion of correct choices across the block going from 41.67% (SE ± 8.1%) to 82.61% (SE ± 6.23%) by the end of the

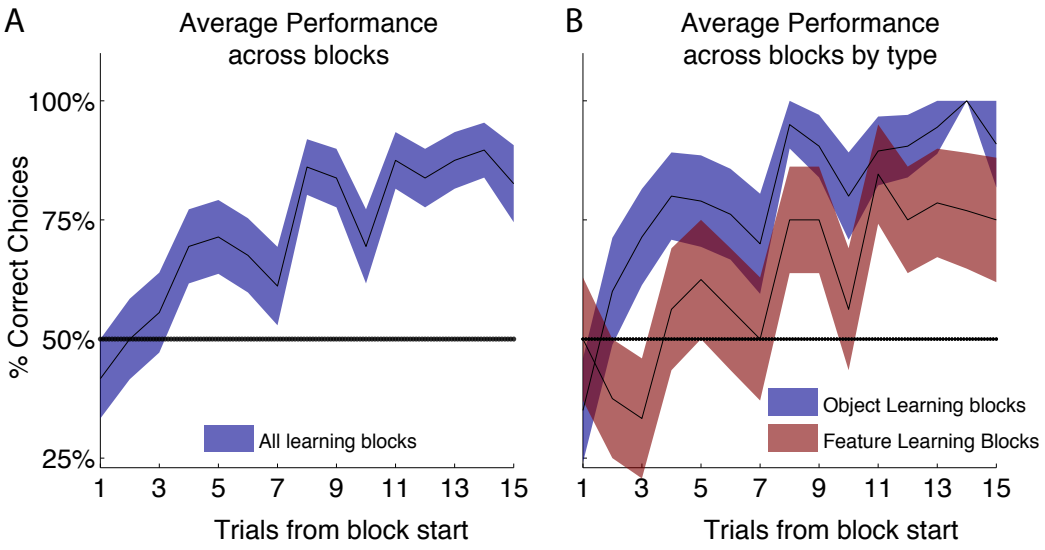


Figure 2. Learning across blocks and by block type. A) Average performance (% correct choices on each trial) across all learning blocks. B) Average performance in object blocks was consistently better than that in feature blocks. This reflects the comparative difficulty of learning under the two block type conditions.

block, for an average increase of 40.9% in correct choices (**Fig. 2a**). Dividing the performance across blocks by block type shows that there is a significant difference in performance between feature blocks and object blocks (Mann-Whitney-Wilcoxon, $p<0.05$) (**Fig. 2b**), indicating that the subject responded differentially to the relative difficulty in learning the appropriate selection rule for feature blocks (see **Materials & Methods**).

We used a number of Q-Learning RL models to capture the systematic learning behaviour of the subject and to provide predictions of the subjective value associated with stimuli and stimulus features (see **Materials & Methods**). Each model incorporated a set of parameters that reflect different influences on the learning process that is well described in the literature, such as limitations in working memory capacity (**Materials & Methods**, eq. 3) (Seymour et al., 2012; Skvortsova et al., 2014), differential learning

rates for positive and negative outcomes (**Materials & Methods**, eqs. 4 & 5)(Gehring and Willoughby, 2002), generalization (or lack thereof) of outcome information to non-selected options (**Materials & Methods**, eq. 2), feature specific selection rules (Flexible Rule Selection - FR_Sel) (**Materials & Methods**, eq. 7 & 8), and combinations of all these characteristics (QL_Combined). We fit all models to the behaviour of the subject by minimizing the ordinary least square distance between the mean probability of selecting the correct stimulus predicted by the model and the mean observed likelihood of the subject making a correct choice across both block types (Donoso et al., 2014)

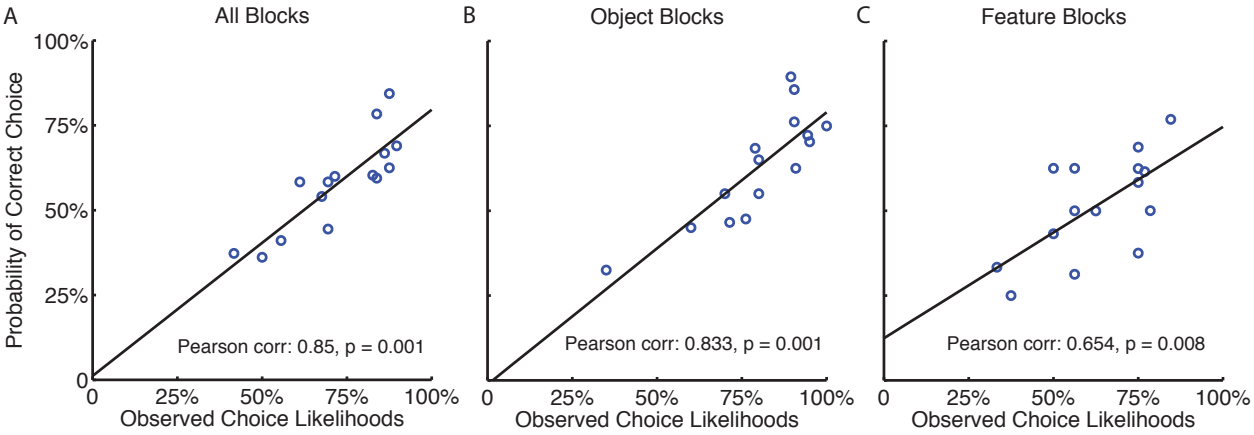


Figure 3. The probability of model values producing correct choices is highly correlated with the observed likelihood of the subject making correct choices. A) Figure shows the mean predicted choice probabilities across all blocks produced by bootstrapping the 80% of the data 100 times. Pearson correlation r=0.864, p<0.05. B) Mean predicted choice probabilities versus the mean likelihood of the subject making a correct choice across object blocks. C) Mean predicted choice probabilities versus the mean likelihood of the subject making a correct choice across feature blocks.

(see **Materials & Methods**). We bootstrapped 80% of the data 100 times while searching for optimal parameters to ensure that values selected for free parameters in each model reflect the most systematic patterns in the behaviour.

The best fitting model for this subject was a model called QL_GainLoss, which produced the highest correlation between predictions and behaviour for all models (**Fig.**

3) (see **Materials & Methods**), which has three free parameters, including two different learning rates - alpha gain: α_G (.21) and alpha loss: α_L (.38) which scale the updating of values differentially according to the valence of the outcome. The model also includes a β (0.002) parameter that governs the non-linear transformation of values into choice

Optimized Model Parameters and Scores

Model Name	Number of parameters	alpha (alphaG)	alphaL	beta	threshold	decay-tau	mean (sem) OLS distance	mean(sem) Pearson correlation
QL GainLoss	3	0.21	0.38	0.002	n/a	n/a	0.8103 (0.01)	0.864 (0.02)
QL Generalized	2	0.24	n/a	0.01	n/a	n/a	0.821 (0.064)	0.755 (0.012)
QL Decay	3	0.45	n/a	0.031	n/a	1.09	2.1 (0.015)	not significant
QL Basic	2	0.16	n/a	0.002	n/a	n/a	1.459 (0.002)	not significant
QL Combined	4	0.3	0.34	0.021	n/a	1.01	0.8658 (0.032)	0.745 (0.001)
FR_Sel	4	0.22	0.1	0.016	0.081	n/a	0.573 (0.022)	0.596 (0.011)
FR_Update	4	0.2	0.1	0.011	0.091	n/a	0.736 (0.011)	0.56 (0.01)

Table 1. Parameter values and optimization scores for RL models. QL_GainLoss significantly more correlated with subject behaviour (Mann-Whitney-Wilcoxon $p<0.05$) than all other models after being optimized to minimize the OLS distance (see Materials & Methods).

probabilities (**Table 1**). The very low value of β for this model indicates that it has a very low likelihood of exploration in regimes where values for alternatives are very different. In conjunction with a low β, the model also has a relatively low value for both α_G and α_L.

Low alpha values indicate that the model needs several positive outcomes (5+) in order for values for rewarding stimuli and features to saturate. Alpha values, the model learning rates, reflect the speed of the model in acquiring new information about reward, which can be interpreted as the confidence in the information value of experienced outcomes; alpha values close to one indicate complete confidence, that every new outcome is perfectly informative about the long-term value of a stimulus. Alpha values closer to zero indicate low confidence in new information, and values for stimuli are changed very slowly.

A recent study in human subjects has shown that the teaching signal employed by RL models, the reward prediction error (RPE), can be linked to subject reaction times as well as changes in reaction time across trials (Cavanagh et al., 2010). The RPE is the difference on each trial between the expected value of the selected stimulus and the experienced outcome. It is hypothesized that surprising outcomes (large RPE) prompt changes in choice behaviour. To explore the relationship between learning mechanisms of the QL_GainLoss model and behavioural adaptations of the subject we performed a generalized linear regression of the RPE produced by the QL_GainLoss model and the reaction times of the subject, where the reaction time was calculated as the time elapsed between the go signal (the removal of the fixation cross; see **Fig. 1**) and the point of contact with the touch screen indicating a response (**Materials & Methods**). We found that there was a weak, but significant, correlation between the RPE and the reaction time on the next trial (Pearson correlation, r=0.102, p=0.012), and there was a slightly larger correlation between RPE and the change in reaction time - computed as the difference between RT on the next trial and the RT on the previous trial (Pearson correlation r=0.16, p=0.001). This model-behavior correlation supports the previously observed correlation between model value predictions and observed choice behaviour and suggests that the model has captured systematic patterns of behaviour in this subject.

The RPE is used by the model to update Q-values, with the difference in the updated Q-values constituting the main decision variable that the subject uses to inform

her decision. We also tested whether the difference in Q-values, a key decision variable,

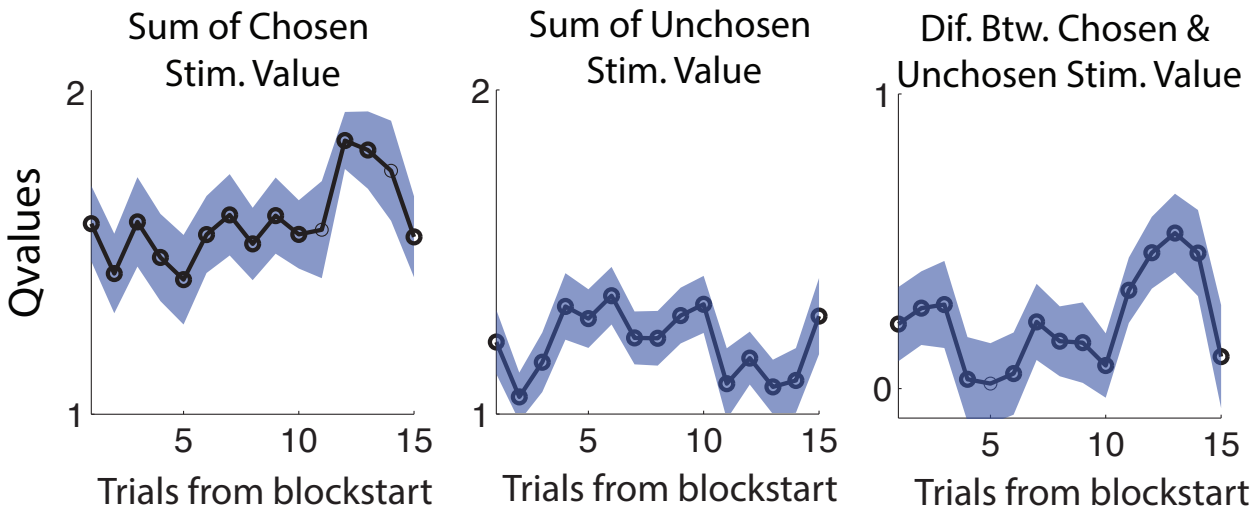


Figure 4. Dynamics of Q-values for chosen and unchosen stimuli across blocks. Q-values for chosen typically exceed that of unchosen stimuli. Trials later in a block see q-values for chosen stimuli rise compared to that of unchosen.

is correlated with reaction times. As shown in **Fig. 4** Q values and Q-value differences varied on average across trials in a block, thus allowing to correlate reaction times and neuronal activity (below) with the model predicted decision variable. We found that Q-value difference negatively correlated with reaction times on the following trial (for correct trials, r = -0.14, p = 0.003; for incorrect trials, r=-0.21, p=0.01).

We measured ECoG activity in the vmPFC of the subject while she was performing the task. Subdural, cortical surface strips were placed over frontal and parietal areas, laterally and medially in both hemispheres (**Fig. 5c - Fig. 6**). Recent work has identified MNI coordinates in vmPFC that show activation related to the value and difference in value for stimuli [3,51,-16] (Hare et al., 2011; Rudorf and Hare, 2014), [-2,44,10] (Skvortsova et al., 2014). Using an MNI template, the native space MRI electrode locations from this subject were transformed into MNI space. For analysis we

selected twelve electrodes from those implanted that were closest to the ROI identified

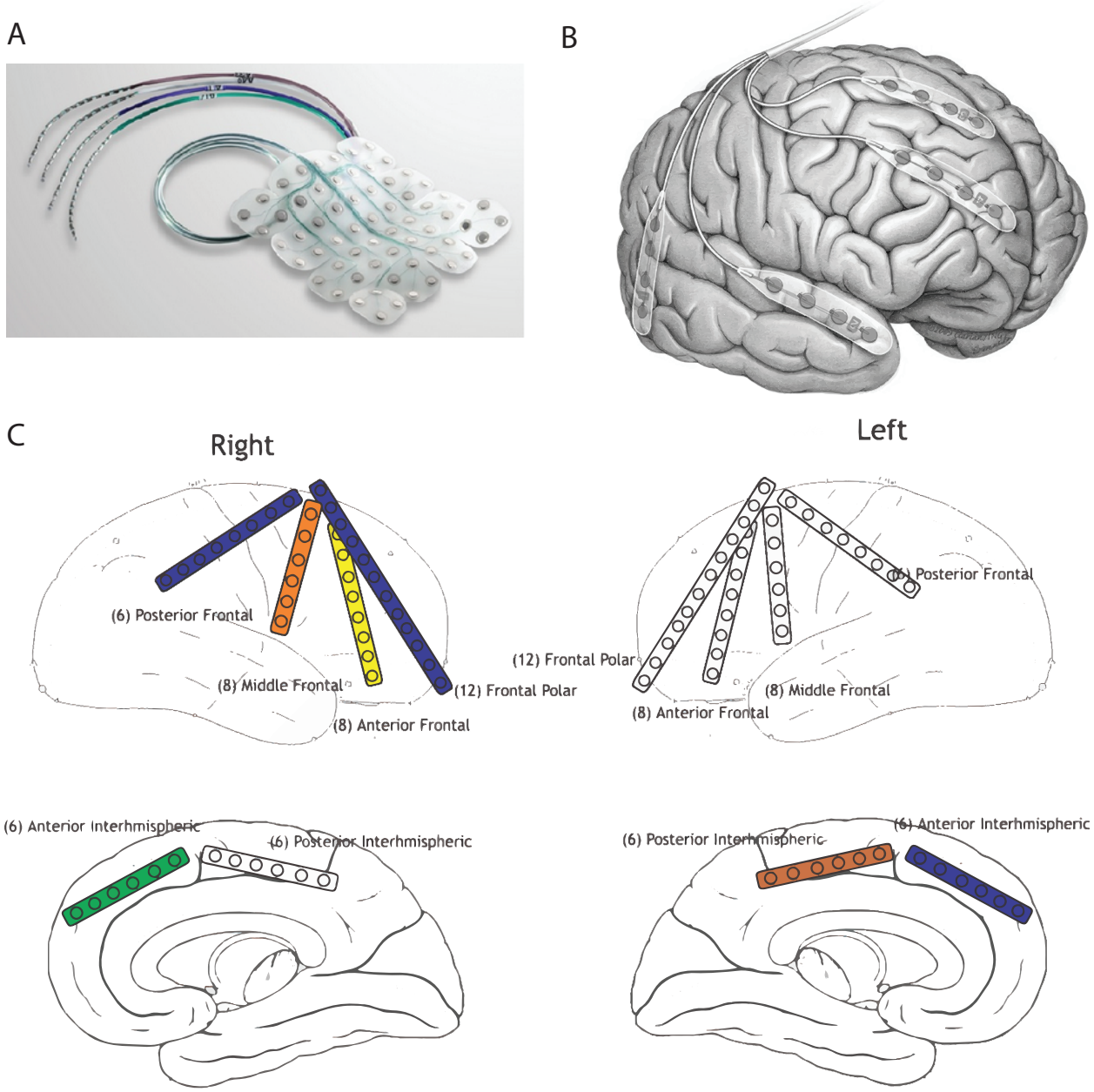


Figure 5. Subdural cortical surface electrodes and surgical placement targets for the monitoring of extracellular current and the detection of epileptic activity in the brain. A) A grid of subdural electrodes from PMT Corporation (Chanhassen, MN, USA) similar to the ones used in our patient. B) A cartoon showing how subdural strip electrodes are placed on the cortical surface. C) A sketch used by the surgeon indicating the targeted coverage areas for our subject.

in the BOLD literature as most likely to show stimulus value related activity. These

electrodes include eight medial interhemispheric locations, anterior left and right, and four on the lateral surface, two left and two right (**Fig. 6**).

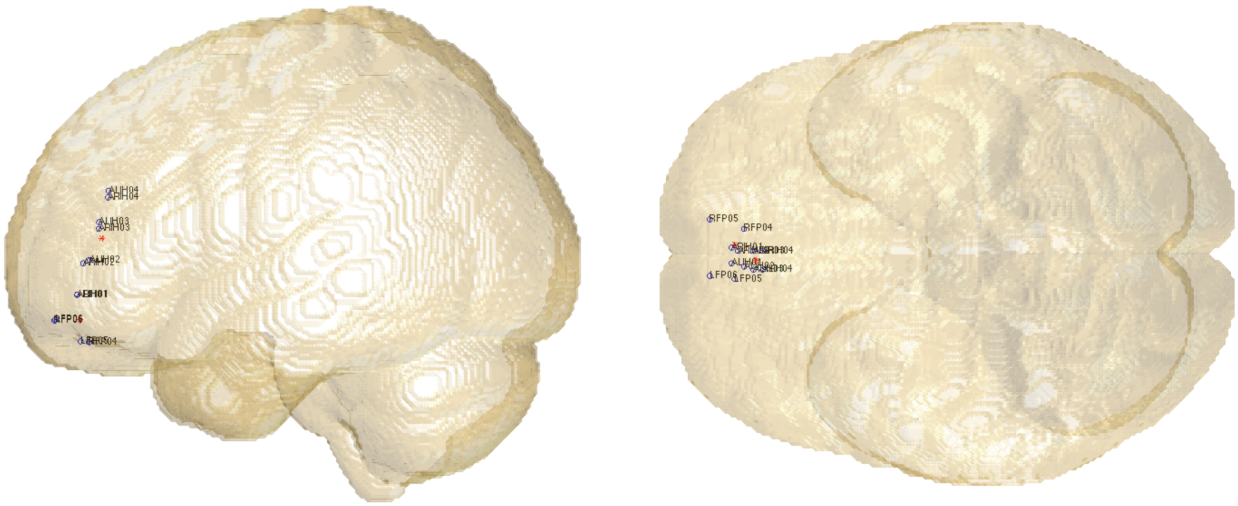


Figure 6. Selected electrode locations and labels projected onto the cortical surface in MNI space. Eight electrodes were on the medial surface, four each of both hemispheres, and four were on the lateral surface.

We performed signal preprocessing and visual artifact rejection using the Fieldtrip toolbox (www.fieldtriptoolbox.org; see **Materials & Methods**). Out of the five sessions of observed performance we selected one session that retained the majority of post-rejection electrodes and trials for further behavioural analysis and robust RL model fitting (see **Materials & Methods**). From this session, we removed four electrodes that contained excessively noisy components, as well as individual trials that reflected oculomotor activity or other transient noisy influences on the signal. We computed the Fourier spectra for each electrode from 2-60Hz averaged over 0.5-1.5 seconds post stimulus onset to confirm the absence of spurious transients in the signal during the decision period. The log power spectra showed showed a relationship to frequency that

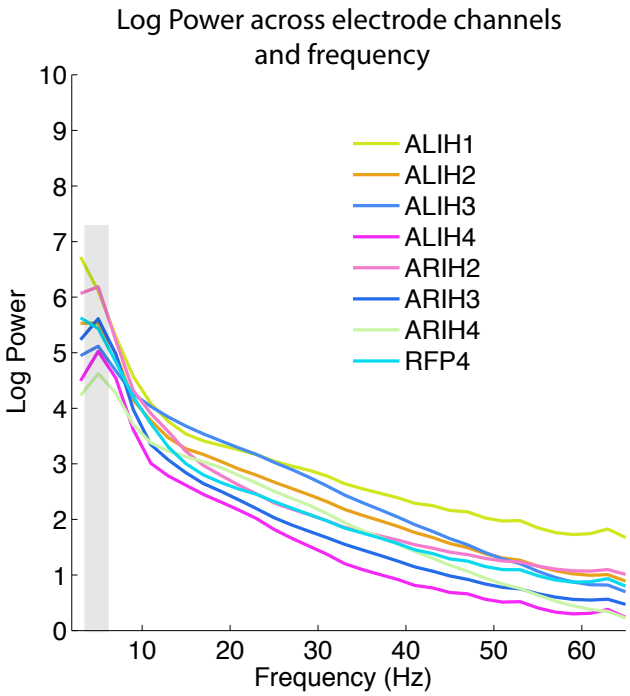


Figure 7. The logarithm of power averaged over .5-1.5 sec post stimulus onset shows an heightened response in the 4-6 Hz frequency range.

was similar to the 1/f shape that is noted in the literature (Buzsáki and Draguhn, 2004), with a notable peak in power in the 4-6 Hz frequency range relative to other frequencies (**Fig. 7**)(Buzsáki and Draguhn, 2004).

We calculated the Pearson correlation between the difference in q-values for available stimuli on each trial predicted by the QL_GainLoss model and the baseline normalized power for all frequencies between 2-30 Hz in 2 Hz increments at each time point in 100 millisecond increments between -0.5 and 1.5 seconds relative to stimulus onset (see **Materials & Methods**). We found that there were seven time points that showed a significant correlation between power and the difference in q-values and all of these were for the 4-6 Hz frequency (mean Pearson correlation r=0.54 (SE ± 0.04); $p<0.05$)(**Fig. 8a**). No other frequency bands showed a significant correlation with q-

value differences in the stimulus onset period. The difference between q-values for available choices in the stimulus onset period is the key decision variable, and the

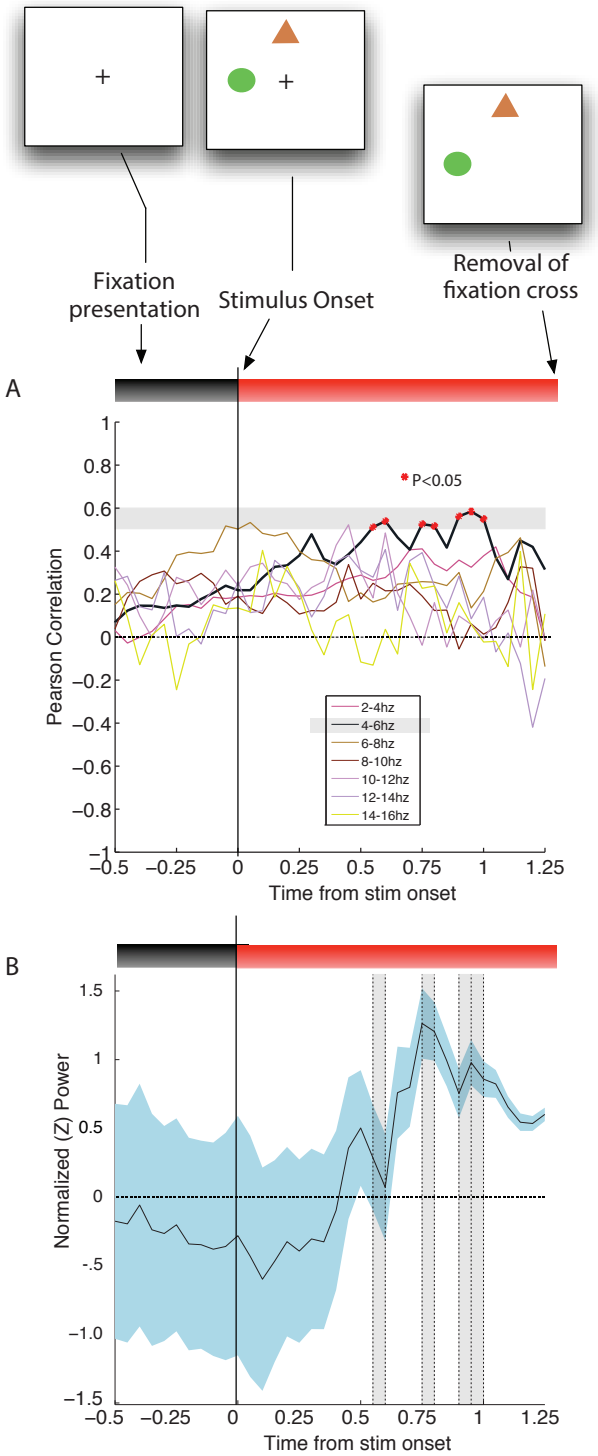


Figure 8. Normalized power in 4-6 Hz is correlated with the difference in value between available stimuli post stimulus onset.

stimulus onset period is the decision window for stimulus selection (**Fig. 8b**). A correlation between this decision variable and theta band power suggests that this band-limited, low frequency component of oscillatory activity in vmPFC may participate in the computational processes of selecting stimuli from among available options before an overt choice is made. Correcting for multiple comparisons using the Bonferonni-Holm method does not allow any p-values to remain as significant. However, further work is required to confirm whether the results shown here indicate a general trend over multiple sessions and subjects or are merely a spurious statistical anomaly.

4.5 Discussion

In our study we showed how a Q-Learning RL model can describe the choice behaviour and reaction times of a subject learning the association between stimuli and outcomes. We showed that this learning was captured by reinforcement learning (RL) mechanisms. The RL mechanisms estimated reward prediction errors that significantly correlated with the actual speed of behavioural adjustment of the subject during learning. Most importantly, the RL predicted decision variable (the difference in predicted stimulus values) correlated with the strength of trial-by-trial power fluctuation in a narrow 4-8Hz theta frequency band in electrodes overlaying the vmPFC of a single human subject. This correlation emerged shortly after stimulus onset during an epoch when the subject supposedly forms the decision which of the two stimuli to select. Taken together, this set of results provide a strong proof-of-principle that reinforcement learning processes can be tracked in band-limited oscillatory activity in those brain

regions that are implicated by slow BOLD fluctuations to convey the critical decision variables underlying behavioural adjustment in complex learning task.

Learning is a complex phenomena, just as humans are complex. Experimental explorations of learning in human subjects are influenced by the wide-ranging experience that people bring to bear on the experimental context. It is not always clear that subjects performing tasks do so with the perspective that is assumed by the experimenter (Shteingart and Loewenstein, 2014). Likewise, learning is a process that can be done better or worse, in terms of the strategies actually employed by a subject out of the total possible set of available mechanisms. It is a strength of RL approaches to understanding variable learning strategies in the brain that it is possible to adapt the RL framework to incorporate unique subject behaviour as well as different possible mechanisms involved in the learning process. We found that the average behaviour of this subject could be split into two significantly different subgroups according to the block type, and we optimized our models to fit both of the performance profiles, which is a novel strategy in model optimization.

The formal, model based analyses of behaviour performed here proved to be a powerful tool, allowing us to account for a wide range of a systematic behaviour. In our study, the selection of the best explaining overall model accounted for the significantly poorer performance in feature blocks, independently of the superior performance of the object blocks. By comparing different learning architectures we also found that the subject showed learning that was differentially influenced by outcome valence. The best fitting model incorporated different learning rates for positive and negative outcomes, which is a frequently observed behavioural phenomena (e.g. Gehring and Willoughby,

2002). Most importantly our results provide direct evidence that this outcome signalling is affecting the subject's learning performance by revealing a positive correlation of reward prediction signals and reaction times in the trials following the occurrence of the reward prediction signal. This finding documents the predictive validity of reinforcement mechanisms for the adjustment of behaviour in compl choice tasks.

In the cognitive neuroscience of decision making, reinforcement learning models have become a powerful and ubiquitous tool for exploring the hidden variables of learning in animal and human subjects (Lee et al., 2011; Rushworth et al., 2011; Vickery et al., 2011). RL is an effective tool because it has been shown to reliably predict many aspects of choice behaviour, but also because it provides specific hypotheses about the state of neural activity (Alexander and Brown, 2011; Khamassi et al., 2013). Different components of RL processes, including the representation of values, the reward prediction error teaching signal, and the selection of final choices have been shown to have associated neural activity at the level of single neurons (Schultz, 1997; Florian, 2007; Blanchard and Hayden, 2014), BOLD (Seymour et al., 2004; Daw et al., 2011; Rushworth et al., 2011; Simon and Daw, 2011; Wimmer et al., 2012), and EEG (Cavanagh et al., 2012; Collins et al., 2014; Frank et al., 2015). Because RL models can provide subject specific, and trial-by-trial, predictions about the dynamics of brain activity, they are capable of providing novel insights into the basic mechanisms of learning in the brain (Dayan and Niv, 2008; Niv, 2009). Possibly due to the difficulty in acquiring ECoG data in humans, use of RL models for analyzing intracranial EEG has rarely been reported in the literature and presents an exciting opportunity for a new level of understanding learning mechanisms in the brain (Jacobs and Kahana, 2010). Here

we have shown as a proof of principle that the trial specific predictions of estimated value for stimuli can be used to interpret the strength of frequency specific oscillations in extracellular cortical currents. This represents a natural extension of many existing methods in the literature and demonstrates that future work in this area will benefit from ongoing use of RL models for model-based analyses of intra-cranial EEG.

Theta frequency oscillations are a low-frequency rhythm in the brain that reflects the synchronized activity of local populations of neurons, and even though theta is associated with several brain functions (Benchenane et al., 2010; Womelsdorf et al., 2010b; Burke et al., 2014), such as the association shown here with stimulus value differences, it is not clear how it facilitates these roles mechanistically (Caplan et al., 2003). In our study the key decision variable of the difference in value between the two available options only becomes correlated with the power of theta oscillations more than five hundred milliseconds after onset of stimuli, which is roughly three cycles of 6 Hz activity, and this correlation continues for another five hundred milliseconds. One theory of the role of theta in cortical computations is that it coordinates interactions between neuronal groups by setting precise temporal windows for local circuit computation (Mizuseki et al., 2009). To make a value based choice between multiple stimuli, several computations need to be performed and results integrated, and three theta cycles may provide sufficient time in which to participate in this information pipeline from working memory recall to motor plan selection (Mizuseki et al., 2009). The correlation seen here could represent the result of a previous computation, which derives the difference in value from the summed value of both stimuli, being made available for further computational processing by the decision circuit (Rangel and Hare, 2010). The change

in theta power in this time window would then set a temporal window for downstream computations and motor activity to read out the relevant choice information from the appropriate neuronal groups. The theory that the correlation of theta power with stimulus value information represents the integration and coordination of stimulus value computations could relate to previous findings that associate theta with sensory-motor integration (Bland and Oddie, 2001; Greenberg et al., 2015), and the consolidation of long term learning in medial temporal lobe structures (Buzsáki, 1996).

In summary, the results of our study provide a versatile argument justifying future studies to search for the precise role of oscillatory theta band activity in human vmPFC during the the learning of abstract associations of stimulus features and objects with rewarding outcomes.

4.6 References

Alexander WH, Brown JW (2011) Medial prefrontal cortex as an action-outcome predictor. Nature Publishing Group 14:1338–1344.

Amso D, Haas S, McShane L, Badre D (2014) Working memory updating and the development of rule-guided behavior. Cognition 133:201–210.

Başar E, Başar-Eroglu C, Karakaş S, Schürmann M (2001) Gamma, alpha, delta, and theta oscillations govern cognitive processes. Int J Psychophysiol 39:241–248.

Benchenane K, Peyrache A, Khamassi M, Tierney PL, Gioanni Y, Battaglia FP, Wiener SI (2010) Coherent theta oscillations and reorganization of spike timing in the hippocampal- prefrontal network upon learning. Neuron 66:921–936.

Blanchard TC, Hayden BY (2014) Neurons in Dorsal Anterior Cingulate Cortex Signal Postdecisional Variables in a Foraging Task. Journal of Neuroscience 34:646–655.

Bland BH, Oddie SD (2001) Theta band oscillation and synchrony in the hippocampal formation and associated structures: the case for its role in sensorimotor integration. Behav Brain Res 127:119–136.

Bosman CA, Schoffelen J-M, Brunet N, Oostenveld R, Bastos AM, Womelsdorf T, Rubehn B, Stieglitz T, De Weerd P, Fries P (2012) Attentional Stimulus Selection through Selective Synchronization between Monkey Visual Areas. Neuron 75:875–888.

Brincat SL, Miller EK (2015) Frequency-specific hippocampal-prefrontal interactions during associative learning. Nat Neurosci:1–10.

Burke JF, Sharan AD, Sperling MR, Ramayya AG, Evans JJ, Healey MK, Beck EN, Davis KA, Lucas TH, Kahana MJ (2014) Theta and high-frequency activity mark spontaneous recall of episodic memories. J Neurosci 34:11355–11365.

Buzsáki G (1996) The hippocampo-neocortical dialogue. Cereb Cortex 6:81–92.

Buzsáki G, Draguhn A (2004) Neuronal oscillations in cortical networks. Science 304:1926–1929.

Buzsáki G, Watson BO (2012) Brain rhythms and neural syntax: implications for efficient coding of cognitive content and neuropsychiatric disease. Dialogues Clin Neurosci 14:345–367.

Canolty RT, Knight RT (2010) The functional role of cross-frequency coupling. Trends in Cognitive Sciences 14:506–515.

Caplan JB, Madsen JR, Schulze-Bonhage A, Aschenbrenner-Scheibe R, Newman EL, Kahana MJ (2003) Human theta oscillations related to sensorimotor integration and spatial learning. J Neurosci 23:4726–4736.

Cavanagh JF, Figueroa CM, Cohen MX, Frank MJ (2012) Frontal Theta Reflects Uncertainty and Unexpectedness during Exploration and Exploitation. Cerebral Cortex 22:2575–2586.

Cavanagh JF, Frank MJ, Klein TJ, Allen JJB (2010) Frontal theta links prediction errors to behavioral adaptation in reinforcement learning. NeuroImage 49:3198–3209.

Collins AGE, Cavanagh JF, Frank MJ (2014) Human EEG uncovers latent generalizable rule structure during learning. J Neurosci 34:4677–4685.

Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ (2011) Model-Based Influences on Humans' Choices and Striatal Prediction Errors. Neuron 69:1204–1215.

Dayan P, Daw ND (2008) Decision theory, reinforcement learning, and the brain. Cognitive, Affective, & Behavioral Neuroscience 8:429–453.

Dayan P, Niv Y (2008) Reinforcement learning: The Good, The Bad and The Ugly. Current Opinion in Neurobiology 18:185–196.

Donoso M, Collins AGE, Koechlin E (2014) Human cognition. Foundations of human reasoning in the prefrontal cortex. Science 344:1481–1486.

Florian RV (2007) Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity. Neural Computation 19:1468–1502.

Frank MJ, Gagne C, Nyhus E, Masters S, Wiecki TV, Cavanagh JF, Badre D (2015) fMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. J Neurosci 35:485–494.

Friston K, Schwartenbeck P, FitzGerald T, Moutoussis M, Behrens T, Dolan RJ (2014) The anatomy of choice: dopamine and decision-making. Philos Trans R Soc Lond, B, Biol Sci 369:20130481–20130481.

Gehring WJ, Willoughby AR (2002) The medial frontal cortex and the rapid processing of monetary gains and losses. Science 295:2279–2282.

Greenberg JA, Burke JF, Haque R, Kahana MJ, Zaghloul KA (2015) Decreases in theta and increases in high frequency activity underlie associative memory encoding. NeuroImage.

Hare TA, Schultz W, Camerer CF, O'Doherty JP, Rangel A (2011) Transformation of stimulus value signals into motor commands during simple choice. Proceedings of the National Academy of Sciences 108:18120–18125.

Hermes D, Miller KJ, Noordmans HJ, Vansteensel MJ, Ramsey NF (2010) Automated electrocorticographic electrode localization on individually rendered brain surfaces. J Neurosci Methods 185:293–298.

Jacobs J, Kahana MJ (2010) Direct brain recordings fuel advances incognitive electrophysiology. Trends in Cognitive Sciences 14:162–171.

Jensen O, Tesche CD (2002) Frontal theta activity in humans increases with memory load in a working memory task. Eur J Neurosci 15:1395–1399.

Khamassi M, Enel P, Dominey PF, Procyk E (2013) Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters. In: Progress in Brain Research, pp 441–464 Progress in Brain Research. Elsevier.

Lee D, Seo H, Jung MW (2011) Neural Basis of Reinforcement Learning and Decision Making. Annu Rev Neurosci 35:120518152625006.

Lee H, Simpson GV, Logothetis NK, Rainer G (2005) Phase locking of single neuron activity to theta oscillations during working memory in monkey extrastriate visual cortex. Neuron 45:147–156.

Mizuseki K, Sirota A, Pastalkova E, Buzsáki G (2009) Theta Oscillations Provide Temporal Windows for Local Circuit Computation in the Entorhinal-Hippocampal Loop. Neuron 64:267–280.

Niv Y (2009) Reinforcement learning in the brain. Journal of Mathematical Psychology.

Niv Y, Duff MO, Dayan P (2005) Dopamine, uncertainty and TD learning. Behav Brain Funct 1:6.

Oostenveld R, Fries P, Maris E, Schoffelen J-M (2011) FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput Intell Neurosci 2011:156869–156869.

Rangel A, Camerer C, Montague PR (2008) A framework for studying the neurobiology of value-based decision making. Nat Rev Neurosci 9:545–556.

Rangel A, Hare T (2010) Neural computations associated with goal-directed choice. Current Opinion in Neurobiology 20:262–270.

Rudorf S, Hare TA (2014) Interactions between dorsolateral and ventromedial prefrontal cortex underlie context-dependent stimulus valuation in goal-directed choice. J Neurosci 34:15988–15996.

Rushworth MFS, Noonan MP, Boorman ED, Walton ME, Behrens TE (2011) Frontal Cortex and Reward-Guided Learning and Decision-Making. Neuron 70:1054–1069.

Schultz W (1997) A Neural Substrate of Prediction and Reward. Science 275:1593–1599.

Seymour B, Daw ND, Roiser JP, Dayan P, Dolan R (2012) Serotonin selectively modulates reward value in human decision-making. Journal of Neuroscience 32:5833–5842.

Seymour B, O'Doherty JP, Dayan P, Koltzenburg M, Jones AK, Dolan RJ, Friston KJ, Frackowiak RS (2004) Temporal difference models describe higher-order learning in humans. Nature 429:664–667.

Shteingart H, Loewenstein Y (2014) Reinforcement learning and human behavior. Current Opinion in Neurobiology 25:93–98.

Simon DA, Daw ND (2011) Neural correlates of forward planning in a spatial decision task in humans. J Neurosci 31:5526–5539.

Skvortsova V, Palminteri S, Pessiglione M (2014) Learning To Minimize Efforts versus Maximizing Rewards: Computational Principles and Neural Correlates. J Neurosci 34:15621–15630.

Tsujimoto T, Shimazu H, Isomura Y (2006) Direct recording of theta oscillations in primate prefrontal and anterior cingulate cortices. Journal of Neurophysiology 95:2987–3000.

Vickery TJ, Chun MM, Lee D (2011) Ubiquity and specificity of reinforcement signals throughout the human brain. Neuron 72:166–177.

Wilson RC, Niv Y (2011) Inferring relevance in a changing world. frontiers in Human Neuroscience 5:189.

Wimmer GE, Daw ND, Shohamy D (2012) Generalization of value in reinforcement learning by humans. Eur J Neurosci 35:1092–1104.

Womelsdorf T, Anton-Erxleben K, Pieper F, Treue S (2006) Dynamic shifts of visual receptive fields in cortical area MT by spatial attention. Nat Neurosci 9:1156–1160.

Womelsdorf T, Fries P (2007) The role of neuronal synchronization in selective attention. Current Opinion in Neurobiology 17:154–160.

Womelsdorf T, Johnston K, Vinck M, Everling S (2010a) Theta-activity in anterior cingulate cortex predicts task rules and their adjustments following errors. PNAS 107:5248–5253.

Womelsdorf T, Schoffelen JM, Oostenveld R, Singer W, Desimone R, Engel AK, Fries P (2007) Modulation of Neuronal Interactions Through Neuronal Synchronization. Science 316:1609–1612.

Womelsdorf T, Vinck M, Leung LS, Everling S (2010b) Selective Theta-Synchronization of Choice-Relevant Information Subserves Goal-Directed Behavior. frontiers in Human Neuroscience 4:1–13.

Chapter 5

5.1 - Summary and future work

In this thesis I have shown that Reinforcement Learning models are capable of capturing both the internal, covert attentional selection of a stimulus feature, and the flexible use of rules used by naive subjects for learning in a novel environment. Through three separate projects, involving both animal models and naive human subjects, we demonstrated the power of RL models to describe the computational mechanisms of both the learning of values and the deployment of learned values for attentional control. Additionally, using the predictions produced by an RL model, we also demonstrated that using RL models for analysis of neural activity can be extended to the domain of oscillatory activity. We showed that trial-by-trial and subject specific analyses of frequency specific ECoG activity can be performed using the RL framework, which permits new and exciting insights into the basic functions of the human brain.

The first important conclusion that can be drawn from the results shown here is, as was previously discussed in chapter 2, covert attentional selection can be realized as stochastic selection acting on specific value predictions, or in other words that top-down sources of attentional control are likely co-extensive with systems of value-based learning. The results we have shown fit with the model of attentional selection we propose (see **Chapter 2, Fig. 9**), whereby in an incentive-driven learning environment, control of attention is driven by mechanisms that track expected value for stimulus

features. This result adds to previous findings that connect value learning and attentional control, but here we clarify that it is not just overt choice behaviour that is explicable in economic terms but also covert internal shifts of attentional focus.

Another implication of the work shown in this thesis is seen in the overlap between projects and the limitations of economic explanations of learning in incentive-based tasks. In chapter 2 we show how covert attentional selection can be realized as stochastic selection acting on specific value predictions, but also that value predictions are not capable of explaining all systematic behaviour. The best explanation of monkey attentional selection required the incorporation of an 'attentional stickiness' component, where the monkey prefers to attend to the feature previously attended to regardless of its associated value. This is an example of the influence of multiple systems on the decision making process, and despite the ability of the pure value based RL model to explain the majority of monkey choice behaviour, it represents a limitation of economic approaches to entirely predict incentive based choice behaviour (Gottlieb et. al., 2014). Similarly, in chapter 3 we show that the best account of choice behaviour in a majority of subjects requires the inclusion of a rule that speeds learning in specific circumstances, and we show that this rule is not learned during task performance. Learning, for 22/32 subjects, was not entirely local and possibly involved input from long-term memory systems. Both results from chapter 2 and 3 demonstrate the complex nature of learning, and selection among, values for covert and overt choice behaviour. To fully describe the processes underlying incentive-driven learning it is necessary to account for the input of multiple systems and sources of information, even in experimental settings (Gottlieb, et. al. 2014; Mongillio et. al. 2014). Again, it is a strength of the computational approach to

cognitive neuroscience that multiple influences of cognitive control, both value-based and non-value based, can be incorporated into a model framework such that specific predictions about their role can be tested (Frank & Badre, 2015). As we showed in the range of models tested against observed behaviour in all projects, many potential sources of input into learning and stimulus selection can be evaluated using the RL framework.

There are several natural extensions of the results shown in this work. In the same way that the q-values produced by an RL model were used to interpret the extra-cellular currents recorded from the cortical surface in chapter 4, the RL model results from chapter 2 are already being utilized for analyzing single cell electrophysiology that was recorded simultaneously while the data analyzed here was collected. As has been mentioned previously, the usefulness of RL models for understanding the neuronal basis of learning and attentional control in the brain comes from their production of very specific predictions about the dynamics of brain activity relative to observable behaviour. Single-unit activity in the macaque brain was simultaneously recorded in three separate locations, lateral prefrontal cortex, anterior cingulate cortex and the hippocampus, and there are many unresolved issues in how these areas dynamically interact during the ongoing learning of values for covert attentional selection. It is expected that future work with the results shown in chapter 2 and the corresponding electrophysiology will produce unique insights into the neuronal circuits underlying multiple choice systems and the subprocesses of stimulus selection.

There are many questions about the role of oscillatory activity in the cortex in relation to learning values for stimulus selection that are unanswered in this work. The

ECoG dataset collected and analyzed here (chapter 4) still has the ability to speak to questions about the role of other areas in frontal cortex and the dynamics of inter-areal interaction for learning and selecting stimuli. The work performed here explored the functional role of low-frequency activity from a subset of available electrode channels. It is expected that further investigations with this data into the role of higher frequency power, cross-frequency interactions, as well as anatomically distinct areas, will yield more insights into the larger pre-frontal network involved in learning and stimulus selection.

The use of RL models to provide computational level predictions about the relationship between behaviour, cognitive function, and neuronal activity is becoming more common and more widespread. In addition to the kind of work shown here, RL models are capable of providing new and unique insights into the pathologies underlying dysfunctional learning and attentional systems in the brain (Stopper and Floresco, 2015; Maia and Frank, 2011). As shown in pilot work exploring the cellular and molecular processes affected by neuro-psychiatric drugs (supported by this author, see **Appendix A** - Additional research contributions), RL models can be used to provide highly specific hypothesis about the relationship of neuro-active compounds to large scale functional networks, something that is currently not well understood. Likewise, RL models are now being used to analyze the choice behaviour of people with certain neuropathologies and mental health disorders as a means of connecting known deficits in learning and decision making to the dysfunction of local circuits associated with the sub-processes of RL (Maia and Frank, 2011; Montague et al., 2012). This nascent research area is known as ‘Computational Psychiatry’ (Montague et al., 2012).

Computational Psychiatry represents one of the most exciting extensions for the use of RL models for understanding the brain and providing new, more effective, interventions. Because RL models contain separable processes with mechanisms that can be adapted to describe the performance of individual subjects, they can provide unique and powerful insights into healthy and unhealthy brains by exploiting their potential for 'model-based' analysis (Frank & Badre, 2015; Mars et. al. 2012). It is hoped that as research using ECoG in human patients with epilepsy continues, such as the work shown here, the insights into the computational mechanisms of learning in the brain will also yield insights into the dysfunctional processes of epilepsy.

5.2 References

Collins AGE, Brown JK, Gold JM, Waltz JA, Frank MJ (2014) Working memory contributions to reinforcement learning impairments in schizophrenia. J Neurosci 34:13747–13756.

Frank MJ, Badre D (2015) How cognitive theory guides neuroscience. Cognition 135:14–20.

Gottlieb, J., Hayhoe, M., Hikosaka, O., & Rangel, A. (2014). Attention, reward, and information seeking. J Neurosci 34:15497–15504.

Maia TV, Frank MJ (2011) From reinforcement learning models to psychiatric and neurological disorders. Nature Publishing Group 14:154–162.

Mars RB, Shea NJ, Kolling N, Rushworth MFS (2012) Model-based analyses: Promises, pitfalls, and example applications to the study of cognitive control. The Quarterly Journal of Experimental Psychology 65:252–267.

Mongillo G, Shteingart H, Loewenstein Y (2014) The Misbehavior of Reinforcement Learning. Proc IEEE 102:528–541.

Montague PR, Dolan RJ, Friston KJ, Dayan P (2012) Computational psychiatry. Trends in Cognitive Sciences 16:72–80.

Stopper CM, Floresco SB (2015) Dopaminergic circuitry and risk/reward decision making: implications for schizophrenia. Schizophr Bull 41:9–14.

Appendix A - Additional research contributions

Ardid, S., **M. Balcarras**, & T. Womelsdorf. (2014). “Adaptive learning” as a mechanistic candidate for reaching optimal task-set representations flexibly. BMC Neuroscience.

Hassani, S.A., M. Oemisch, **M. Balcarras**, S. Westendorff, and T. Womelsdorf. (2015). Alpha-2A Noradrenergic Activation improves behavioural flexibility during Feature-based Reversal Learning. Society for Neuroscience annual conference; Chicago: Nov. 9-13.

Appendix B - Consent form A - York Community members

Date: March 3, 2014

Study Name: Learning how to value visual stimuli: Utilizing the reinforcement learning framework to understand the emergence of selective attentional control.

Researchers: Matthew Balcarras
PhD Candidate,
York University, Department of Biology
Faculty of Science and Engineering
4700 Keele Street

Prof. Thilo Womelsdorf
York University, Department of Biology
Faculty of Science and Engineering
4700 Keele Street
Toronto, Ontario, M3J 1P3

Purpose of the Research: The purpose of the study is an improved knowledge of human perception and decision making. Your participation will help to understand the brain mechanisms underlying these higher human cognitive functions. In particular, we investigate how decisions about visual experiences are formed and test in the experiments different aspects that influence the efficiency of human decision making and our ability to selectively focus our attention on specific visual objects. For example, your performance will be compared between tasks, which differ only in the focus of your visual attention. Differences in performance between these task conditions allow conclusions about the influence of selective attention to the processing of visual information.
A detailed understanding of these functions is an important prerequisite for helping patients suffering from specific visual, attentional and learning disturbances. The participation in the tests per se does not yield a direct health benefit, but rather will inform clinical researchers in improving health benefits.

What You Will Be Asked to Do in the Research: You will be asked to participate in an interactive computer based decision-making experiment, where we will track your choices to test your decision making processes. In this experiment, pairs of stimuli will be presented on a computer screen. You will sit on a chair in front of this screen, using a normal posture and position. When indicated by a visual cue, you will select one of the two stimuli by touching it with your finger on the touchscreen. After making a selection you will be asked to press the keyboard 'spacebar' to receive feedback about your choice. Feedback is given in the form of either a 'happy face' pictogram in the center of the screen or with the text 'sorry'. A single presentation of stimuli and the corresponding choice and feedback is called a trial, and a total test consists of approximately 150-200 trials.

A session typically takes one hour and includes the repetition of the test. This hour includes a break in between tests. You will set the pace as you start every test by keystroke. Typically, a study consists of several sessions. It is very important for us, that you finish a study completely. But you are free to interrupt the measurements at any time.

Risks and Discomforts: We do not foresee any risks or discomfort from your participation in the research.

Benefits of the Research and Benefits to You: The research will inform our general understanding of human decision making. We will incorporate the results from these tests into psychological and neuroscientific theories of decision making that are being developed to understand the brain mechanisms that underlie human decision making. You have no immediate benefit from these tests, but you can choose after completion of the tests to learn about the specific scientific questions that we tested and thereby obtain knowledge about the current scientific reasoning about human decision making.

Voluntary Participation: Your participation in the study is completely voluntary and you may choose to stop participating at any time. You have the right to not answer any specific questions. If you would decide not to volunteer this will not have any influence on the nature of your relationship with the researchers involved or with York University either now, or in the future. Participants will receive a $10 gift card redeemable at either Starbucks or Tim Horton's.

Withdrawal from the Study: You can stop participating in the study at any time, for any reason, if you so decide. Your decision to stop participating, or to refuse to answer particular questions, will not affect your relationship with the researchers, York University, or any other group associated with this project. If you choose not to continue, you will still receive a gift card for $10 to either Starbucks or Tim Horton's. In the event you withdraw from the study, all associated data collected will be immediately destroyed wherever possible.

Confidentiality: All information you supply during the research will be held in confidence and unless you specifically indicate your consent, your name will not appear in any report or publication of the research. Names of participants are removed for data processing and analysis, and are not associated with results in any way. The data will be archived on a secure server with encrypted password protection contained in the supervisor's office. The data will be securely stored for a period of at least three years after which it will be archived by the research supervisor on secure servers under his control. Confidentiality will be provided to the fullest extent possible by law.

Questions About the Research? If you have questions about the research in general or about your role in the study, please feel free to contact the lead researcher, Matthew Balcarras by email or the head of the laboratory conducting the research, Dr. Thilo Womelsdorf, by e-mail. This research has been reviewed and approved by the Human Participants Review Sub-Committee, York University's Ethics Review Board and conforms to the standards of the Canadian Tri-Council Research Ethics guidelines. If you have any questions about this process, or about your rights as a participant in the study, please contact the Sr. Manager & Policy Advisor for the Office of Research Ethics.

Legal Rights and Signatures:

I, _____, consent to participate in the study on attention and decision making conducted by Matthew Balcarras, Dr. Thilo Womelsdorf and his lab's researchers. I have understood the nature of this project and wish to participate. I am not waiving any of my legal rights by signing this form. My signature below indicates my consent.

Signature_____ **Date**_____
Participant

Signature_____ **Date**_____
Principal Investigator

Appendix C - Informed Consent form B - UHN

University Health Network

CONSENT TO PARTICIPATE IN A RESEARCH STUDY

Study Title Neuropsychological and Neurophysiological Testing in Functional Neurosurgery Patients

Principle Investigator

Taufik Valiante MD PhD
Dept of Neurosurgery
Toronto Western Hospital
University Health Network
[redacted]

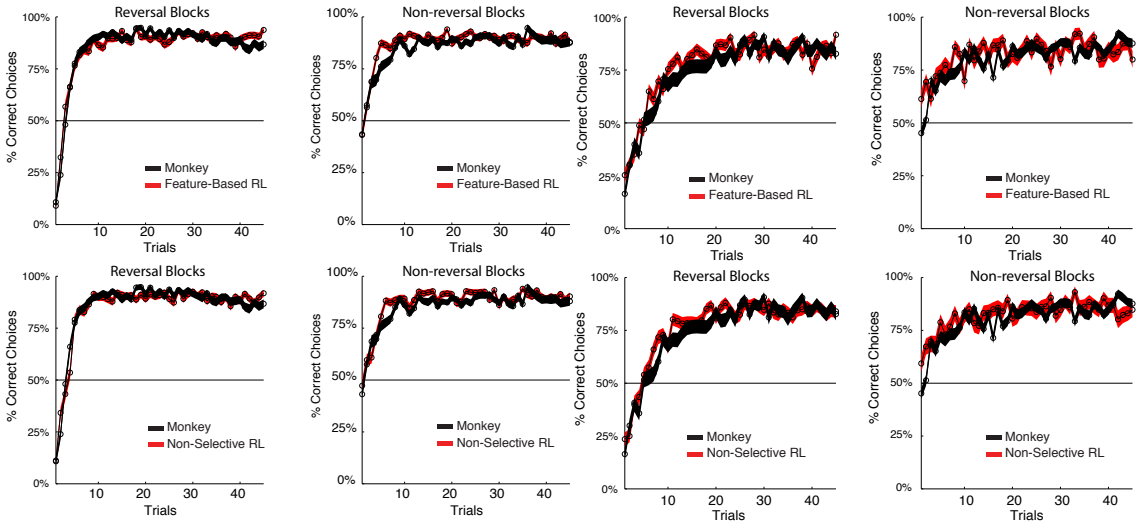
Co-Investigators

Dr. Andres M. Lozano, MD, PhD Dept of Neurosurgery Toronto Western Hospital University Health Network [redacted]	Dr. Nir Lipsman, MD, PhD Dept of Neurosurgery Toronto Western Hospital University Health Network [redacted]
Dr. Christopher Honey PhD Department of Psychology [redacted] Toronto, ON M5S 3G3 Canada	Dr. Thilo Womelsdorf PhD York University, Department of Biology, Faculty of Science, 4700 Keele Street [redacted]
Dr. Kari Hoffman PhD Centre for Vision Research Depts of Psychology, Biology York University [redacted]	

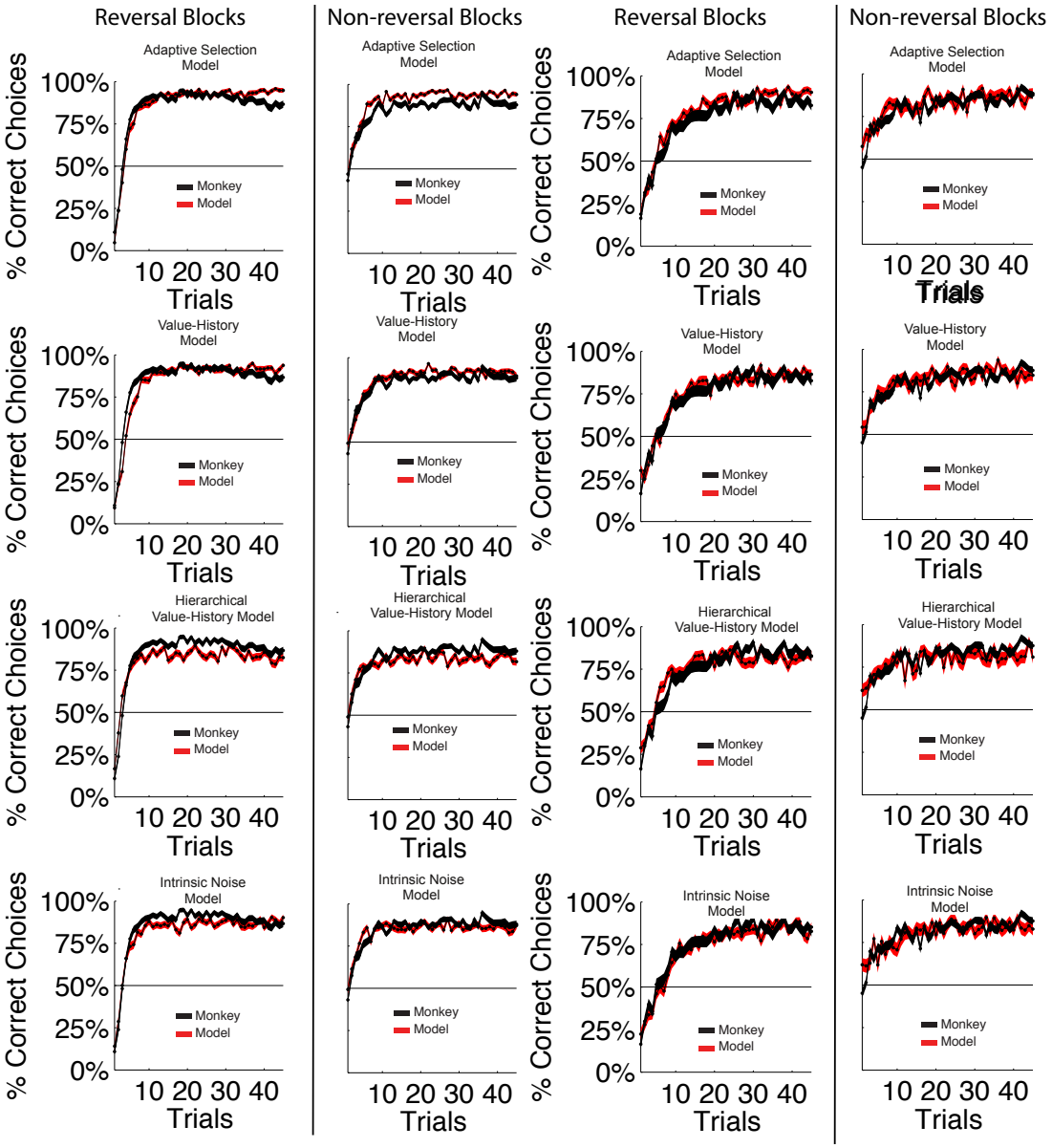
Funding Source No funding is required for this study

24 Hour Pager Number Dr. Nir Lipsman
Neurosurgery Resident, Study Co-ordinator
[redacted]

Appendix D - Chapter 2 Supplementary figures

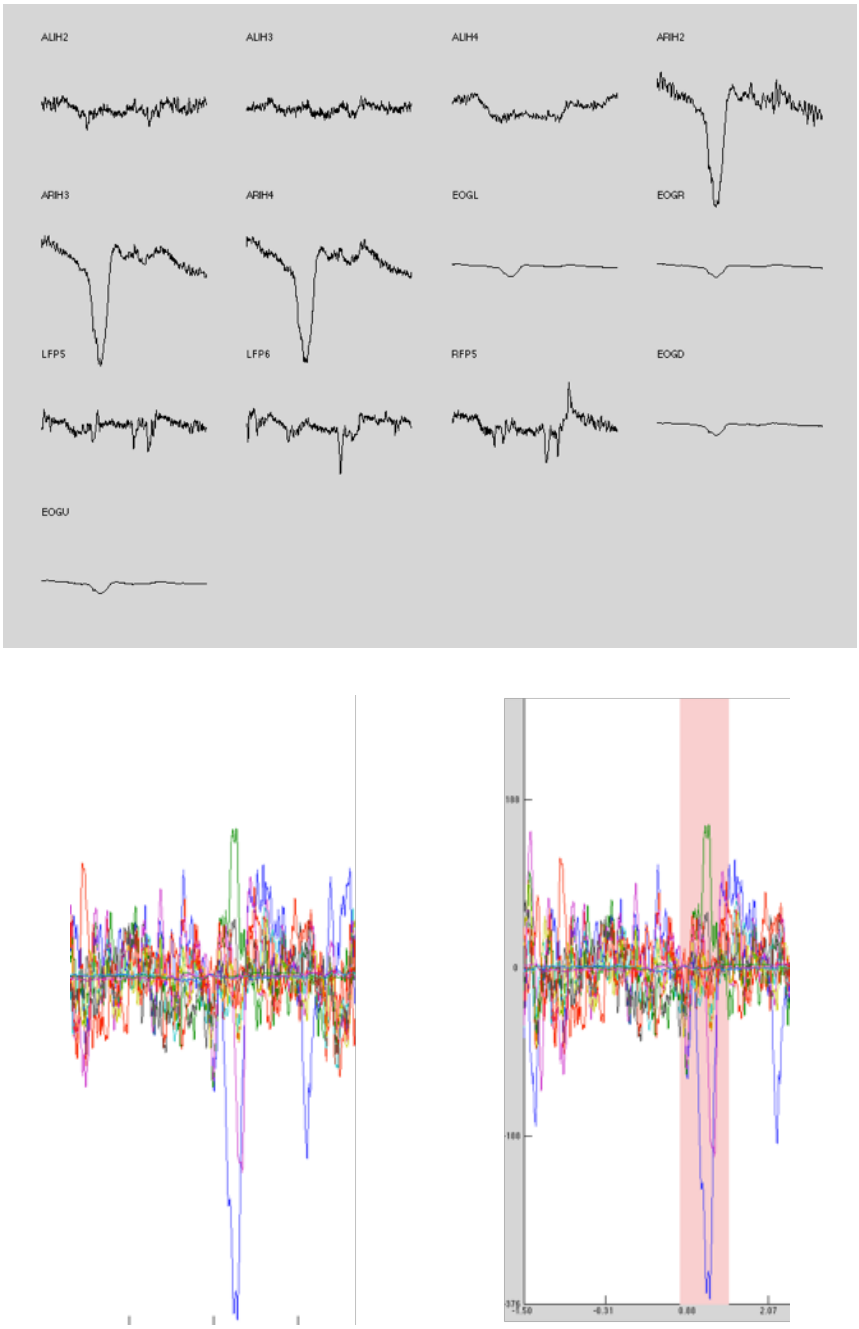


Supplementary Figure 1. Average performance curves for monkey 'M' - left panels, and monkey 'S' - right panels, versus Feature-Based RL and non-selective RL for blocks following a colour-reward reversal, and non-reversal blocks.



Supplementary Figure 2. Average performance curves for monkey 'M' - left panels, and monkey 'S' - right panels, versus alternative models for blocks following a colour-reward reversal, and non-reversal blocks.

Appendix E - Chapter 4 Supplementary figures



Supplementary Figure 1. An example of transient activity in EEG signal rejected from analysis due to noise introduced from external sources. Figure shows a screenshot from the graphic interface of the 'databrowser' function in the FieldTrip toolbox. Individual traces show the raw signal recorded on a single trial in nine EEG electrodes and four EOG electrodes. Significant transient activity indicates non-relevant noise in the EEG signal.

Appendix F - Collaborative contributions to this work

In chapter two, the task was designed by Dr. Thilo Womelsdorf, and the monkey data was collected by Dr. Womelsdorf and Dr. Daniel Kaping. Logistic regression analyses related to possible selection biases was performed by Dr. Salva Ardid. Dr. Ardid also contributed comments and supervisory direction to modelling and behavioural analysis.

In chapter three, data collection was aided by undergraduate student, Omar Abid.

In chapter four, electrode implantation was performed by neurosurgeon Dr. Taufik Valiante, and electrode location reconstruction was performed by Dr. Cristiano Micheli.