Engagement, Surprise and Exploration in the Macaque Frontal Lobe

Jan Grohn

St John's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2022

Acknowledgements

First and foremost, I would like to thank my supervisors, Nils Kolling and Matthew Rushworth, for their continued help and guidance. Nils and Matthew devoted a lot of time and effort towards teaching me, and I would have not achieved any of this without their personal and professional support.

I am also incredibly grateful to the many other people who helped me complete this DPhil. The data underlying Chapter 2 was primarily collected by Urs Schüffelgen with help from Jérôme Sallet and Franz-Xaver Neubert. The data in Chapter 3 was primarily collected by Caroline Jahn with help from Jérôme Sallet, Steven Cuell and Andrew Emberton. The data from Chapters 2 and 3 also underlies Chapter 4, together with additional data collected primarily by Alessandro Bongioanni and Nima Khalighinejad, with help from Jérôme Sallet and Davide Folloni. Apart from data collection, I also greatly profited from the many discussions I had with everybody involved, in particular with Caroline, Alessandro, Nima and Jérôme but also Mark Walton, Sebastien Bouret and Lennart Verhagen.

I would also like to thank all the people that worked together with me on other projects alongside this thesis. I had lot of fun and learned a great deal from modelling risk-sensitive behaviour with Moritz Möller, Sanjay Manohar and Rafal Bogacz, working on models of reward trajectories and goal pursuit with Eleanor Holton, and on altering volatility estimates with Lilian Weber.

In addition to the many people directly involved with projects I worked on, I received a lot of support and help from the research groups I was embedded in. It was a pleasure working alongside everybody in the Rushworth group, and I particularly benefitted from discussions and help I received from Jacqueline Scholl and Miriam Klein-Flügge. I also received a lot of input from the joint meetings with the groups of Laurence Hunt and Jill O'Reilly, both during discussions of my projects and beyond.

Outside work, I received continued support throughout from my friends and family, which I am also very thankful for. Their presence and encouragement greatly enhanced my experience throughout this degree.

Finally, I would not have been able to pursue this degree without the financial help I received. I am very grateful that this thesis was made possible due to scholarships from the MRC, St John's College and the Scatcherd European Scholarship.

Abstract

In this thesis I investigate how macaque monkeys make reward-guided decisions. Specifically, I investigate behavioural, computational, and neural mechanisms for detecting and learning from surprising events, choosing whether to explore or exploit, and staying engaged. Using fMRI data of macaques performing different tasks, I identify neural activity patterns underlying these types of behaviours. I demonstrate that macaques notice reward-based surprises beyond scalar reward prediction errors: I link neural activity on the orbital surface both to obtaining a surprisingly rare amount of reward (Chapter 2), and to counterfactual prediction errors associated with an unchosen option (Chapter 3). I show that the signals related to rare reward surprise only occur when the reward-environment is sufficiently complex and incentivises learning: such surprise signals could only be detected when the underlying reward probabilities changed over the course of a session, while I find no surprise signal for static reward environments. I also demonstrate that macaques strategically modulate their tendency to explore or exploit depending on the reward-environment, and link such behaviours to different signs of reward-value signals in anterior- and mid-cingulate gyrus (Chapter 3). Finally, I identify general neural signals associated with engaging with a task rather than pausing to respond (Chapter 4). To this end, I show that a network of regions spanning perigenual anterior cingulate cortex, orbitofrontal cortex and striatum is more active when animals are engaged with a task, and causally link activity in perigenual anterior cingulate to task engagement using transcranial ultrasound stimulation. By contrast, mid- and posterior cingulate are more active when animal spontaneously disengage from a task. Using data from four distinct decision-making tasks, and controlling for all factors that were manipulated in these tasks, allows me to show that these engagement-patterns I found are task-independent and capture an intrinsic component of motivation not previously studied.

Contents

Appendices

List of Figures

1

General introduction

Contents

1.1 Thesis overview

This thesis examines several different concepts in the field of reward-guided learning and decision making. Historically, the literature has mainly focused on learning from reward prediction errors—the difference between an expected and an obtained reward—since the influential paper published by Schultz and colleagues in 1997 [1]. However, some value-neutral features of reward might also be worth learning, and learning should be strategically modulated based on the context and goals that

matter to the decision-maker. Lastly, all learning and decision-making occurs in a motivational context. These themes will be explored more throughout the thesis.

A particular focus of the thesis is on the frontal cortex (Fig 1.1A shows parcellations of the frontal cortex for different species). The frontal cortex has received a stark increase in attention in the last 30 years with a particular emphasis on studying the human frontal cortex (Fig 1.1B). Many of the human studies on the prefrontal cortex employ functional imaging, and the surge of publications in recent years shown in Fig 1.1B can likely be attributed to the increased popularity of imaging techniques such as functional magnetic resonance imagining (fMRI), electroencephalography (EEG), and to a lesser extend magnetoencephalography (MEG). Unlike publications on humans, rats or mice, the number of published papers on the macaque frontal cortex remained relatively constant throughout the years (Fig 1.1B). The great majority of papers that are published on the macaque frontal cortex directly record extra- or intracellular neural activity. This thesis is an exception to this; all neural data reported here has been obtained using functional imagining of macaques, specially fMRI. Using fMRI allows for imaging the whole frontal cortex rather than only selected sub-regions. As such, a focus of this thesis is where functional activity is located.

Across all chapters of this thesis, the cognitive process of interest is first modelled and validated. Next, the process is linked to its neural substrate using regressions and causal manipulations. To this end, models are fitted to the behavioural data that aim to capture the effects of interest. Such models can be used to infer hidden variables that are not directly observable from behaviour. For example, while in a task with two choices it is directly observable what option a participant picks on each trial, a model allows us to infer further cognitive variables, such as an approximation of the value the participant assigns to each option. We can infer value from choice data by making assumptions such as that higher valued options are chosen more frequently, and that this frequency is modulated by the exact difference in value between the options. Cleverly designed task and models allow for inferring complex hidden variables such as, for example, macaques' metamemory (the self-monitoring

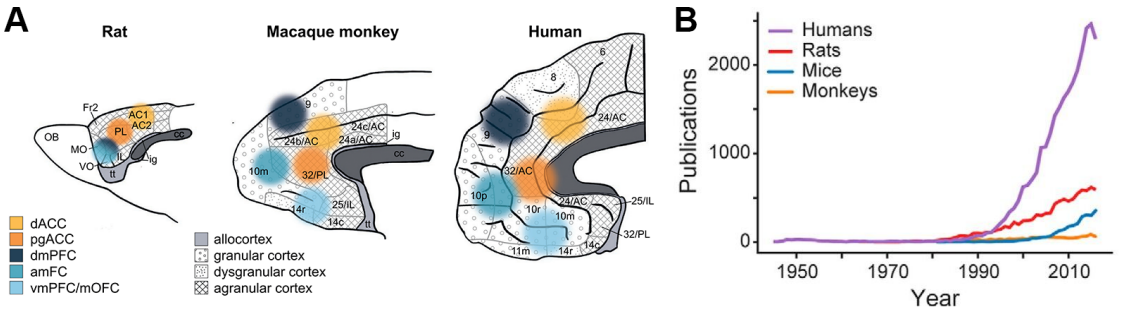


Figure 1.1: **Subregions of the frontal cortex and publications.** (**A**) Locations of functional regions in the frontal discussed in the text in the rat (left) macaque (middle) and human (right) brain. Dorsal anterior cingulate cortex (dACC), perigenual anterior cingulate cortex (pgACC), dorsomedial prefrontal cortex (dmPFC), anterior medial frontal cortex (amFC), and ventromedial prefrontal cortex / medial orbitofrontal cortex (vmPFC/mOFC) are shown, superimposed on cytoarchitectonic maps. While homologues between all regions exist for macaques and humans, this is not the case for rodents, as indicated by the colour scheme. Panel adapted from [2]. (**B**) Numbers of yearly publications on the prefrontal cortex for humans (purple line), rats (red line), mice (blue line), and monkeys (orange line). Panel adapted from [3].

of their own memories) [4], or how macaques compute the trajectory of a target they pursue [5]. Having identified such hidden variables, we can then correlate their trial-by-trial changes against neural activity recorded using fMRI. This allows us to identify regions of the brain with activity patterns that resemble our variable of interest. For example, we might find a region that shows increased activity when the value of a chosen option is high, and decreased activity when the value is low.

In Chapter 2 I discuss different notions of surprise, and show their neural substrates in the macaque brain. Yet another kind of surprise is integral to the results shown in Chapter 3. The importance of surprise in animal learning has been discovered early on [6], and the neural substrates of when an animal is surprisingly rewarded, or an expected reward is surprisingly omitted, are among the best-known results in the field [1]. The key idea behind surprise is that the brain makes predictions about the world, and then tracks if, when and how these predictions are violated. This notion of surprise is far more general than noticing surprising rewards: it has been used to describe most processes the brain is involved in, ranging from sensory processing of vision [7], somatosensation [8], and hearing [9] to associative

learning [10], executing actions [11] or reading [12].

Learning from surprising events is crucial to guide future behaviour. To make good choices, animals have to flexibly adapt their behaviour not only to recent violations of their expectation but also to contextual cues. In Chapter 3 I show how macaques modulate a particular kind of behaviour—whether to stick with the current best option or explore an alternative—depending on the context of the choice they make. I also show how value is differently represented in the brain, depending on the choice context.

Finally, in Chapter 4 I examine a more basic cognitive function. Here I examine the neural correlates of engaging with a task. If offered the choice between two options, a human or animal might instead choose to not respond at all, i.e. choose neither of the options. While such behaviour very rarely occurs for human participants studied in the laboratory (unless the task is exceedingly hard), it is a frequent occurrence when testing animals. While normally considered a nuisance, I use these pauses to build cognitive models to quantify the level of engagement on each trial, and to identify which parts of the brain are more active when the animals are engaged, or more active when they are disengaged. Finally, I show that altering neural activity can change the engagement of animals with a task, which provides a causal link between task engagement and neural activity.

The remainder of this chapter introduces some of the concepts underlying the studies in Chapters 2 - 4. It also links different processes to different areas in the brain. I will first discuss exploitation/exploration trade-offs as such studies lay a foundation for how we think about decision-making under uncertainty and in the context of learning. This is followed by the neural correlates of exploration and exploitation. I then give a brief summary of how learning can be understood using formal algorithmic models, such as reinforcement learning models. I show how such models can quantify how humans and animals match their expectations with outcomes and update their beliefs accordingly. I then discuss structural approaches to learning, which allow for learning more features of outcomes than just their

average value. Finally, I discuss basic engagement with a task as a perquisite for the types of behaviours discussed previously.

1.2 The exploitation/exploration trade-off

1.2.1 Behavioural advantages of exploration

The kinds of environments that animals encounter throughout their lives often require making complex decisions to survive: As environments are ever-changing, uncertain, and stochastic, animals need to adapt their behaviour accordingly [13, 14]. One example of such a decision is whether to continue harvesting food or water in the current location, or to try out other locations instead. While the current location might seem like the richest option available based on evidence accumulated in the recent past, continuing to harvest from it might prove detrimental because the animal would lose out on gathering information about alternative options. Harvesting from an alternative might be more advantageous, and thus knowing about it can better guide future decisions. In other words, there is an opportunity cost to exploitation, as unexplored alternatives in a changing world might offer higher average returns in the long run.

This trade-off between exploiting what seems to be the currently richest option or exploring alternatives has been studied extensively: for example, it has been shown that the frequency with which environments fluctuates between high or low values—i.e. how stable or volatile an environment is—should change how frequently animals explore such environments [15]. Whether animal behaviour aligns with such model predictions has been tested in numerous species, including pigeons [16], hummingbirds [17], and chipmunks [18].

Exploration-exploitation trade-offs have also been investigated in numerous other disciplines, with early examples in control engineering, where a system under uncertainty needs to both be optimally estimated and controlled [19]. Other disciplines that tackled this problem include economics [20] and business management [21]. To analyse and simplify the mathematical problem of trading off exploration and exploitation, the problem has often been framed as a gambler having to choose

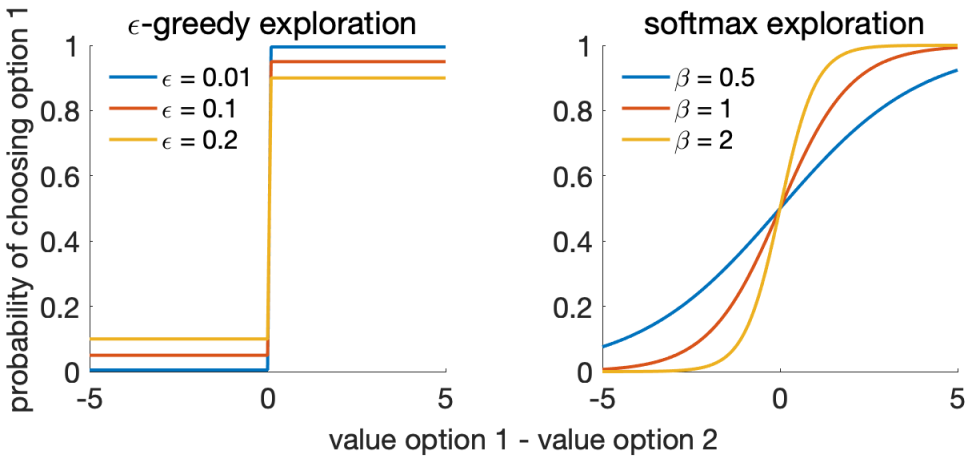


Figure 1.2: **Different types of exploration.** Probability of choosing an option in a two-alternatives choice scenario as a function of the value difference between the options. Two different ways to model exploration are shown. The left panel shows ϵ-greedy exploration. Here, the option with higher value is chosen in $1-\epsilon$ percent of cases, while a random option is chosen in ϵ percent of cases. The right panel shows softmax exploration. Here, the probability of choosing option 1 is given by $\frac{e^{\beta(value\ \ option\ \ 1)}}{e^{\beta(value\ \ option\ \ 1)}+\ e^{\beta(value\ \ option\ \ 2)}}$, where the parameter β controls the slope of the function. A function with a larger β thus leads to less exploration, as choices are determined more by the value difference.

between different slot machines (also called multi-armed bandits): should the gambler stick with the option that currently seems best, based on the recent history of rewards, or try out other slot machines to potentially discover an even better one? Several potential solutions to this problem have been proposed such as occasionally picking a random option and exploiting otherwise—called ϵ-greedy exploration [22] (Fig 1.2 left)—or assigning probabilities to options according to their relative expected value, and then choosing options with these probabilities—called softmax exploration [23] (Fig 1.2 right). Solving multi-armed bandit problem has become an integral part of the theory of computational reinforcement learning [24].

Multi-armed bandit tasks have been extensively used to study human and animal choice behaviour. It has been shown that humans use softmax rather than ϵ-greedy exploration [25], while ϵ-greedy exploration remains the standard way to model exploration in machine learning applications [24]. However, later work also showed that in addition to softmax exploration, random trialwise variability in

updating can also promote exploration through the learning process [26]. These types of exploration (ϵ-greedy, softmax, and noisy learning of action values) are undirected—exploration happens due to randomness in choice or updates, and thus a better option might be discovered by chance. In addition, humans also employ directed exploration, where informative alternatives are deliberately sampled [27]. This can be modelled as adding a bonus to the value of more uncertain options.

To show such an effect, Wilson and colleagues [27] devised a task in which participants had to choose between two bandits. On the first four choices, participants were given outcomes associated with either slot-machine but could not select which machine was chosen themselves ('forced-choice trials'; Fig 1.3A). Afterwards, they were either given one (horizon 1) or six (horizon 6) consecutive free choices between the same two slot machines ('free-choice trials'; Fig 1.3A). The task was designed in such a way that one slot machine was, on average, better than the other. Wilson and colleagues reasoned that exploration should only lead to overall more reward in the horizon 6 case. In the horizon 1 case, the slot machines can only be played once, and a decision-maker should thus simply choose the option with the highest expected value. In other words, while choosing the slot machine with lower expected value might reveal that its expected value is higher than expected, this is irrelevant for future choices as the machines can only be played once. By contrast, in the horizon 6 case the machines can be played 5 more times, and thus any additional information about their expected values can be used on subsequent choices (Fig 1.3B).

Wilson and colleagues' findings showed that their participants use both undirected and directed exploration in their task. Their analysis showed that on first-choices in both the horizon 1 and horizon 6 case (shown in orange in Fig 1.3A), participants choice-patterns resembled softmax exploration (Fig 1.3C), which is consistent with undirected exploration. Additionally, if there were unequal amounts of initial information given about the bandits (i.e. three forced choices for one bandit and one for the other; Fig 1.3B), then people assigned a bonus to choosing the more informative option (the softmax for horizon 6 is shifted

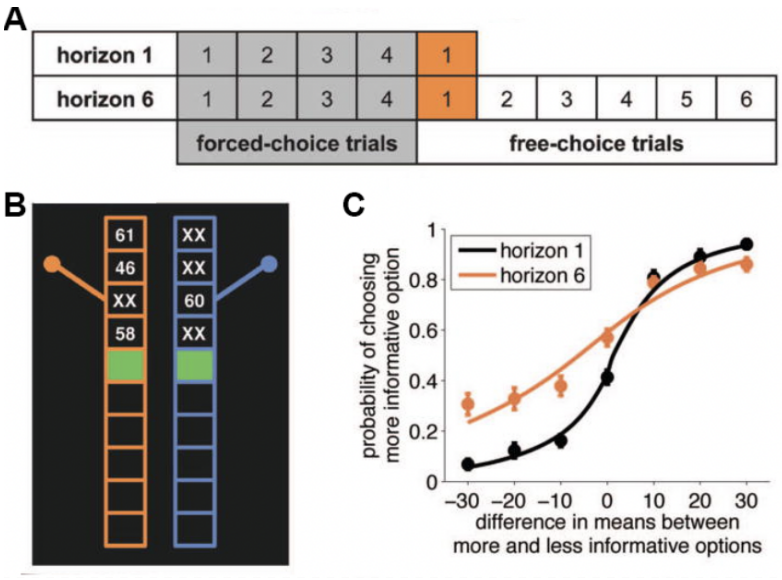


Figure 1.3: **Horizon task by Wilson and colleagues**. (**A**) Participants first see the outcome of four forced-choice trials, before making either one (horizon 1) or six (horizon 6) choices between the same two slot-machines. (**B**) Slot machines for the horizon 6 case. Slot machines differ in their average reward. (**C**) Probability of choosing the more informative option by difference in the expected reward of the options, if options were chosen unequally on forced-choice trials. While undirected softmax exploration is used in both the horizon 1 and horizon 6 case, participants also use directed exploration by assigning a bonus to the more informative option (i.e. the option that was chosen less on forced-choice trials). Figure adapted from [27].

upwards compared to the softmax for horizon 1; Fig 1.3C), which indicates that they are using directed exploration.

As indicated by Wilson and colleagues, the degree to which exploration is being used can and should be strategically modulating depending on the choice context. However, strategic modulation can come in many forms, for example by modulating both directed and undirected exploration, or only either one of these forms of exploration. In Chapter 3 I show behavioural and neural data of a task similar to the one Wilson and colleagues used but adapted for macaque monkeys. While I found no evidence for directed exploration in the behaviour of the macaques we tested, I show that macaques strategically modulate the degree to which they use undirected exploration depending on the task-context. Strategic modulation of behaviour is also found in humans, which use sophisticated strategies to explore

their environments, such as repeatedly choosing novel options before then directly sampling options with the most uncertainty [28].

The idea behind both directed and undirected exploration is that discovering information about the environment enables agents to make better choices in the future. However, there is also evidence that suggests humans and animals value and seek out such information even if it does not help guide future choices in the task [29, 30]. In other words, such overall behavioural tendencies towards curiosity [31] might have a benefit evolutionary by favouring knowledge acquisition and its exploitation in the long run, but do not always require a direct benefit to the agent.

1.2.2 The neural substrates of exploration

Neurally, frontopolar cortex has been shown to be more active when humans explore [25], and to represent the relative advantage of switching to a better option [32]. By contrast, ventromedial prefrontal cortex (vmPFC; see Fig 1.1A for anatomical location) tracks the value of the option that is currently being exploited [25, 32]. However, the need to explore only arises in contexts where rewards are not static but need to be monitored and updated. Thus, exploration and learning are interconnected, both theoretically and neurally. Exploration due to noisy updates of action values correlates with activity in the dorsal anterior cingulate cortex (dACC; see Fig 1.1A for anatomical location) [26], whereas neural correlates of the changes in uncertainty used in directed exploration have been found in frontopolar cortex [33]. In line with this finding, disrupting frontopolar activity using transcranial magnetic stimulation (TMS) impaired directed but not undirected exploration [34].

It has been proposed that exploratory or exploitative brain states are regulated by the neuromodulatory locus coeruleus-norepinephrine (LC-NE) system rather than only activity changes in specific brain regions. The LC-NE system provides inputs to dAC and its change affects pupil dilation which can be used as an indirect index of LC activity [35, 36]. For instance, larger baseline pupil diameter has been associated with making exploratory choices [37], suggesting increased LC activity when exploring. NE, the associated neurotransmitter, has been linked to

undirected, value-free exploration, which has been shown to be affected when NE levels are altered, whereas there is no such effect on directed exploration [38]. These findings suggest that there are multiple different systems that regulate different kinds of exploration. The system regulating undirected random exploration can be linked to NE, which transmits a general brain-state rather than being locally contained within a specific area.

Within dACC, specific microcircuits have been linked to exploratory behaviour. For example, in one study by Tervo and colleagues [39], rats were trained on a task in which they had to accept or reject two different options that were cued by different tones. The reward probabilities of each option were set up in a way where one option was better than the other, but this could change in unpredictable and uncued ways, which required the rats to occasionally explore. When optogenetically silencing the pathway that links dACC and substantia nigra pars reticulata (SNpr), rats explored less. By contrast, optogenetically silencing a different pathway that links dACC and striatum resulted in rejecting non-rewarded options more. This finding highlights distinct computations dACC performs: it is involved in exploration via the dACC-SNpr pathway but also in persistence and effort via the dACC-striatal pathway. The function of dACC beyond its role in exploration, which is to carefully weigh up the pros and cons of complex strategies such as value-driven exploration, will be discussed in more detail later in this chapter.

1.3 Learning the value and other features of choices

1.3.1 Reinforcement learning

As mentioned before, exploration and exploitation are inherently linked to learning: Most models of exploration assume that its aim is to learn the value of alternatives to choose them more/less in the long run. Analogously, most models of learning assume (at least implicitly) that its aim is to enable better choices in the long run either directly or through generalisation.

A simple but immensely popular model of how animals and humans learn is that for each option, an internal value representation is incrementally updated to match

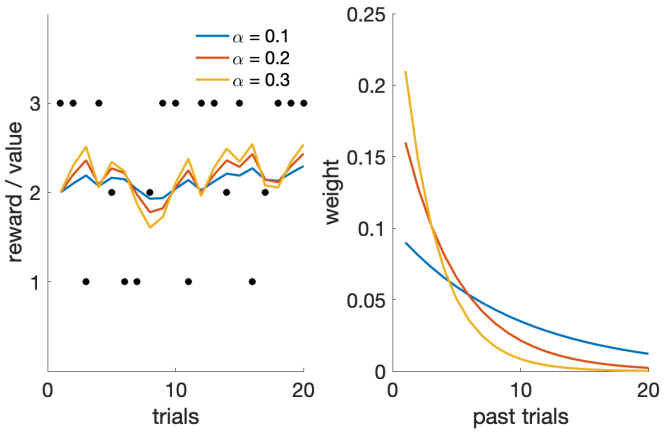


Figure 1.4: **Rescorla-Wagner learning.** The left panel shows value updates for three different hypothetical learning rates. On each trial, reward (black dots) is delivered and the value is updated proportional to the prediction error. The right panel shows the weights different learning rates apply to recently experienced rewards. Compared to a lower learning rate, a higher learning rates weights recent trials more strongly and past trials less strongly.

the true value of that option. This process encapsulated in the Rescorla-Wagner learning rule [40], which states that the change in the value v is proportional to the difference between the experienced reward and the current value estimate (called a prediction error δ). The proportionality factor that determines the size of the update is called a learning rate α, which is assumed to lie between 0 (no learning) and 1 (only learn from the most recent reward). As such, the full learning rule is $\Delta v = \alpha\delta$. Fig 1.4 left shows this learning rule in action for different values of α: on each trial, a reward of either 1, 2, or 3 is delivered, and the value estimate is updated. For smaller values of α the resulting updates are also smaller, and the value thus stays closer to the overall mean, whereas for larger α's the updates are larger, and the value estimate is more strongly influenced by the recent history. This can also be seen in Fig 1.4 right, where the influence of past rewards on the current value estimate is shown. A smaller α puts comparatively more weight on past and less weight on recent trials compared to a larger α.

Learning optimally in an environment constitutes finding an α that is suitable for the environment. In a static environment like in Fig 1.4, it makes sense to initially start with a somewhat high α to adjust the learned value to the range of rewards,

and then to use a low α afterwards that tracks the (constant) mean of the reward distribution. This allows the agent to not mistakenly over-learn from a reward away from the mean as the weight but to integrate rewards over a long time-horizon. By contrast, if the mean the rewards are sampled from changes throughout the task, then a higher α is appropriate as this allows the agent to adjust its value expectation to these changes. Here, a low α would mistakenly put an emphasis on rewards in the distant past that were sampled from a distribution with a mean that has now changed and thus should not be used to guide choices in the present.

As such, learning in a noisy environment (and finding the best α) amounts to determining the source of misestimation/prediction error: to what degree is the noise due to sampling from some constant mean - suggesting the trialwise fluctuations and errors should mostly be ignored (sometimes called 'stochasticity') - or due to a genuine change in the mean, meaning the expectation should be updated rapidly (sometimes this meta-estimate of genuine changeability of an environment is called 'volatility' estimation) [41, 42].

1.3.2 Neural correlates of prediction errors and

It has been famously shown that dopamine neurons encode reward prediction errors [1, 43], which unified prediction error-based theories of animal learning (such as the Rescorla-Wagner rule) and theoretical results in artificial intelligence with discoveries in neuroscience. It has been found that dopaminergic neurons in the ventral tegmental area (VTA) homogeneously encode reward prediction errors [44] but these prediction errors vary in their corresponding learning rates for positive and negative prediction errors, which allows the system as a whole to represent the full reward distribution [45]. Representing reward by more than just its average value (as done, for example, by the Rescorla-Wagner learning rule), comes with a range of benefits, such as enabling more robust learning and allowing a system to make risk-sensitive choices [46, 47]. While the most direct evidence for reward prediction errors in VTA comes from electrophysiological recordings, reward prediction errors can also be detected in humans using fMRI [48]. Within the basal ganglia, dopamine

is supplied to the striatum by the substantia nigra pars compacta, which neighbours the VTA. As in the VTA, dopamine-depended reward prediction error signals can also be detected in the striatum using fMRI [49].

1.3.3 Reward features beyond value

However, dopamine seems to also encode other types of surprises beyond value prediction errors. For example, when reward has an unexpected identity, while the value remains the same, activity can also be detected in the dopaminergic midbrain [50]. Apart from the dopaminergic midbrain, such identity prediction errors are also found in the OFC [51]. Another type of surprise that has been found to evoke dopaminergic activity is the unexpected onset of a stimulus, over and above the value associated with that stimulus [52, 53]. It has also been reported that dopaminergic activity responds to rare rewards [54]: Rothenhoefer and colleagues devised a task in which monkeys were rewarded with 3 different amounts of juice (0.2 ml, 0.4 ml, or 0.6 ml) with either equal probability or varying probability (Fig 1.5A). By recording extracellular dopamine, they discovered that dopaminergic neurons respond more strongly to 0.2 ml and 0.6 ml of juice if these rewards were rare (i.e. drawn from a Normal distribution in Fig 1.5A), rather than occurring with the same probability as 0.4 ml of juice (i.e. drawn from a Uniform distribution in Fig 1.5A). This is indicated by the steeper slope for the Normal distribution in Fig 1.5B compared to the Uniform distribution, and suggests that dopamine also encodes the rarity of rewards.

However, alternative explanations for this finding are also possible. One other known feature of dopaminergic reward prediction errors is that these are scaled up or down depending on the standard deviation of the distribution rewards are sampled from [55, 56]. While the absolute magnitudes of the juice amounts in Fig 1.5A are the same for both distributions, the Uniform distribution has a larger standard deviation than the Normal distribution. Thus, dividing the negative prediction error associated with 0.2 ml of juice and the positive prediction error associated with 0.6 ml of juice by the standard deviations of the underlying distributions can also

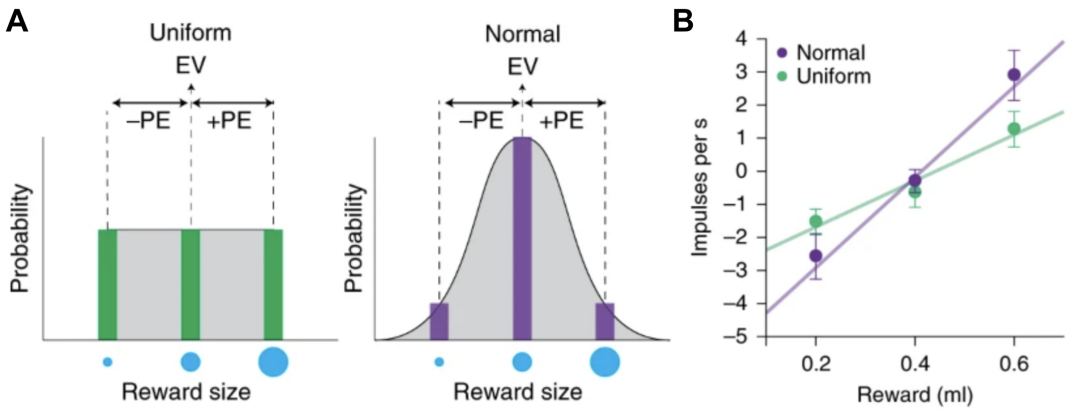


Figure 1.5: **Dopaminergic response to rare rewards.** In their task, Rothenhoefer and colleagues [54] manipulated the distributions rewards were drawn from. (**A**) Rewards were either drawn from a Uniform or a Normal distribution, which made small and large reward more likely to occur in the Uniform compared to the Normal distribution. However, because the amounts of reward were the same for both distributions, prediction errors in the Rescorla-Wagner sense were the same for both distributions. (**B**) Extracellular dopamine responds more strongly to the small and large rewards for the Normal compared to the Uniform distribution. Figure adapted from [54]

explain the pattern observed in Fig 1.5B: The slope associated with the Uniform distribution is shallower because it is divided by a larger number.

In Chapter 2 I further examine surprise signals in the macaque brain. Using fMRI, I show patterns of activity for classic reward prediction errors as discussed above, and contrast these with activity for other kinds of surprises. I show how activity is distinct for visuospatial surprise (a stimulus occurring in an unexpected location) and rare reward surprise. The latter is of particular interest as due to our task-design, these rare rewards do not coincide with prediction errors as described in the Rescorla-Wagner rule. We achieve this by contrasting reward amounts sampled from a Uniform and a U-shaped distribution, which gets around the issues discussed with regards to comparing a Uniform and Normal distribution in Fig 1.5: In our task, the rare reward is also the average expected reward and thus scaling is irrelevant as no reward prediction error occurs. Nevertheless, I demonstrate that macaques can detect these rare rewards, which indicates another way in which reward in the brain is represented beyond its average value.

1.3.4 Learning the value of choices

Just as it is beneficial for an agent to form a representation of the distribution of expected rewards, it is also beneficial for an agent to form a distribution of the opportunities within an environment. One way to learn such a distribution is by learning in parallel with different learning rates, which allows the agent to track changing value over different time scales. Using fMRI, it has been shown that within the dACC activity in more anterior voxels is better explained by assuming that they compute the value of choices using smaller learning rates, while more posterior voxels update with a comparatively higher learning rate [57]. Along with learning at different timescales, dACC also represents the value of alternatives options. For example, in a task where monkeys had to track the changing values of three options to make good choices, dACC was representing the value of the best alternative option, regardless whether this option was available but unchosen on a trial, or unavailable entirely [58]. Such a value can be computed in highly complex ways that include factors beyond its average value, such as the time horizon of a choice and the risk associated with it [59, 60]. Apart from representing the value of an opportunity, dACC also represents a course of action to achieve the opportunity [61]. This links back to the role of dACC in exploratory behaviour, which was discussed earlier.

While the dACC represents long-term value, a representation of the immediate current value of a choice is often found in vmPFC. To show this distinction, Boorman and colleagues [62] designed a task in which value could be determined both long and short term. While constant long-term value signals were found in dACC, vmPFC activity reflected short-term immediate value of choices. In particular, they found that vmPFC activity reflected the difference in value between the chosen and the unchosen option—i.e. it compares the value of an option to an alternative. Such a value comparison has been hypothesised to arise from mutual inhibition between two competing subsets of neurons, each representing the value of an option. It has been shown in a biophysically plausible attractor model that such mutual inhibition gives rise to value signals like the one observed in vmPFC [63–66]. Lesions to vmPFC in

macaques [67] and humans [68, 69] reduces the accuracy of choices, indicating that the value representation in vmPFC is necessary for an accurate choices process.

In line with also encoding the value of the unchosen option, vmPFC has also been implicated in showing regret signals when an opportunity was missed. Steiner and Redish [70] developed a task in which rats had to choose whether to wait for a reward or skip an opportunity in favour of another one. If the rats skipped an opportunity but then encountered a worse opportunity afterwards, cells in the OFC represented the missed action. In humans, vmPFC has also been linked to regret: Activity in the region was linked to regret signals using fMRI [71], and patients with lesions to the area do not express regret [72]. Such regret signals can also be thought of as predictions errors for unchosen options that can be used to update the value of the unchosen option that is computed in vmPFC. However, they might also signal to vmPFC whether that the kind of value-computation it is involved in should be used less in the future as it has resulted in a missed opportunity.

In Chapter 3 I demonstrate the existence of counterfactual prediction error signals in vmPFC in macaques. In our task, macaques were shown the hypothetical reward outcome of choices they did not make, which allowed me to compute prediction errors both for the chosen and the unchosen reward. I then link these counterfactual prediction errors to activity in vmPFC.

One feature of value signals in vmPFC, which I also observe, is that the sign of the BOLD signal in macaques is reversed compared to humans. While the sign is positive in humans, such as for example in the study discussed above [62] but also multiple other studies [32, 73–75], consistently negative signs of the BOLD signal have been found in macaques [58, 76–78]. It has been hypothesised [2] that the negative sign is due to a slightly differently implementation of the attractor network discussed above [64–66].

1.3.5 Structural knowledge in the brain

However, while the dACC tracks opportunities in an environment, and vmPFC tracks the value of the current choice, neither links these opportunities to the

states they occur in, i.e., assign the credit to larger contexts or specific properties of options. For example, let us say that when crossing the road, two outcomes are possible: we make it to the other side and reach our home, or we get hit by a car and end up in hospital. Estimating which of these outcomes is likely to occur requires an understanding of the causal structure of the world, e.g. linking a car that is approaching to potential danger. DACC is likely not involved in learning structural knowledge; however, it might be able to use such knowledge once it has been acquired. A region that has been implicated in learning such states is the orbitofrontal cortex (OFC) [79]. What I call OFC here is located more laterally on the orbital surface than vmPFC.

In a study [80] demonstrating the distinction between ACC and OFC, mice were trained on a task in which water reward could be delivered from two different locations. However, at any one time only one of these locations would probabilistically deliver reward, and the rewarded location would switch occasionally but unpredictably (Fig 1.6). When ACC was deactivated, it took mice longer to change their behaviour and they persisted longer in choosing a location that was no longer rewarded, presumably because ACC is needed to represent the opportunity that comes with choosing the other location. In the task, the researchers also manipulated how difficult it was to detect changes by changing the probabilities with which water was delivered and the location changed. They also included a condition in which it was harder to switch locations as they placed a physical barrier between them. Importantly, mice with ACC inactivation showed no difference between these conditions but were equally impaired throughout. By contrast, when OFC was deactivated, an impairment was seen only in more difficult conditions.

To further investigate this effect, the researchers had set up their task such that it could be solved using two different strategies: One strategy to solve the task is to track the probabilities of reward for each location with a learning rule such as the Rescorla-Wagner rule described above. The agent would then choose the location with a higher reward probability while occasionally exploring the other location. However, a better strategy for this task is to figure out the underlying

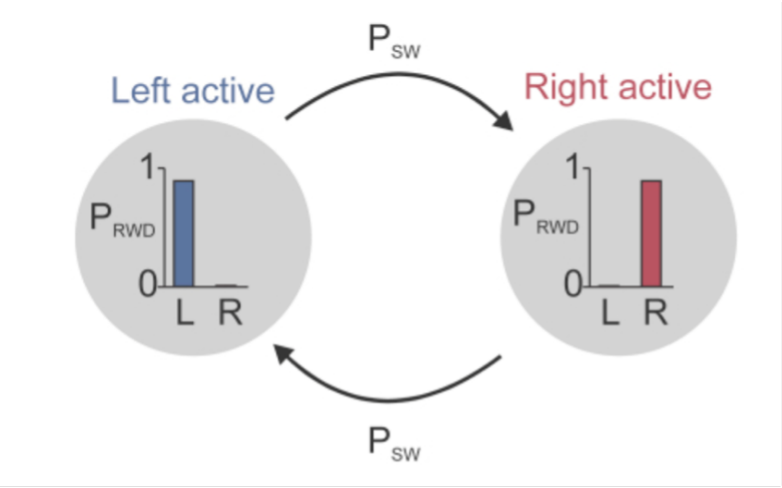


Figure 1.6: **Task structure used by Vertechi and colleagues.** On each trial, either the left or the right side probabilistically delivers reward with a probability < 1. Which side delivers reward can also probabilistically switch on every trial. Figure adapted from [80].

task structure and make use of the fact that experiencing a single reward from any location indicates that this is the currently active location (Fig 1.6)). This latter strategy was used by mice with both an intact ACC and OFC, and also by mice with an inactivated ACC. In stark contrast, mice with an inactivated OFC used the former strategy, and thus resorted to a simpler form of inference. This demonstrates that representing the task states—which of the two locations is currently rewarded—requires the OFC.

However, while the rodent OFC seems to encode such structural knowledge [81–83], the evidence for structural knowledge in the primate OFC is more mixed [2]. It has been suggested that in primates, anterior medial prefrontal cortex (amPFC; see Fig 1.1A for anatomical location) represents structural knowledge. Activity related to such structural knowledge has been found in amPFC across a range of studies [76, 84–86].

Structural knowledge of environmental states—correctly assigning credit to larger contexts or properties of options—is crucial for making complex decisions. Identifying a relevant context or state is often also necessary the kinds of decisions discussed previously, e.g. what values to track or compare with each other.

1.4 Task engagement

In the previous sections it was always implicitly assumed that it was better to act—e.g. to explore or exploit—than not to act at all. However, whether to act at all or not to engage with a task is also a decision that can be made. Neurally, perigenual anterior cingulate (pgACC; see Fig 1.1A for anatomical location) has been implicated in integrating the benefits and costs of whether to pursue an action, which makes it well situated to decide whether to initiate an action. In one study [87], rats were trained to run through a T-maze. One side of the maze contained a small reward, while the other contained a large reward but required the rat to climb a barrier. While healthy rats preferred the high effort high reward arm of the maze, rats with a lesion to the pgACC preferred the low effort low reward arm of the maze. Importantly, they were still capable of climbing the barrier if they had to, indicating that without a pgACC they were lacking the necessary motivation to pursue the higher value option.

In another study [88], microstimulation of the pgACC made macaques less likely to pay a cost—in the form of an air puff—to obtain a juice reward. Here, the air puff and the reward were delivered together but the macaques also had a choice to not engage with the cue at all and forgo both cost and reward. This suggests that pgACC is necessary to initiate a costly action. Cost can take on many forms, including cognitive costs. In a sequential decision-making task [60], human participants could either perform costly mental simulations to compute the complex value of their choices, or use a simple heuristic to compute value. It was found that pgACC is more active the more participants performed such costly computations.

All the studies discussed above examine engagement by manipulating the effort that choices require. However, the willingness to engage with a task also depends on factors intrinsic to the decision-maker: our motivation to engage with effortful tasks fluctuates throughout the day; sometimes we seem motivated to overcome great costs, while at other times even the slightest bit of effort makes us give up on a task. Such intrinsic fluctuations of motivation have not been studied much as they, by definition, do not depend on external factors that can be manipulated experimentally.

In Chapter 4, I use data from different reward-based decision-making tasks to study such intrinsic engagement, and present evidence for a representation of general task engagement in pgACC. I analysed pauses in the behaviour of macaques performing four distinct decision-making tasks, and used a computational model to quantify the degree to which macaques were engaged or disengaged with the tasks. By controlling for all the factors manipulated in the tasks, such as reward history, I was able to link pgACC activity to task-independent engagement. I also show that stimulating pgACC using ultrasound alters the pattern of how macaques engage with a task. Moreover, I contrast this task engagement with another measure of motivation, which is how fast or slow animals respond on engaged trials.

One benefit of the approach I take in Chapter 4 to study motivation is that I combine data from four distinct tasks, including the tasks described in Chapters 2 and 3. This allows for examining the generalisability of my results beyond the specifics of a singular task. Generalisability is often established by replicating effects across studies, where similar behavioural and neural patterns are shown. However, results can also sometimes surprisingly be less generalisable than expected. For example, it has been shown that even when performing learning tasks that are similar in design, fitted model parameters of the same participants do not generalise well between different tasks [89]. Neurally, it has been shown that individual neurons in ACC shift their firing state if a rat switches the rule it uses to navigate. Even if the rat returns to a rule that it employed earlier, the neurons do not revert back to their original state but instead form an altogether new state [90].

Beyond activity patterns, neurotransmitters have also been linked to motivational states. In particular, motivation has also been linked to the neurotransmitter dopamine, just like reward prediction errors. There have been attempts to theoretically link the reward prediction error function of dopamine with its role in motivating behaviour [91–93] but not all the empirical predictions of such theories could be found experimentally [94, 95]. Dopamine has been implicated in energising behaviour [96], and it has been proposed that its role is in overcoming the cost associated with obtaining a reward [97]. How exactly, and if at all, dopamine takes

on the dual role of signalling reward prediction errors and energises behaviour remains an area of active debate [98].

Structural knowledge of environmental states—correctly assigning credit to larger contexts or properties of options—is crucial for making complex decisions. Identifying a relevant context or state is often also necessary the kinds of decisions discussed previously, e.g. what values to track or compare with each other.

2

Systems in macaques for tracking prediction errors and other surprises

Contents

Abstract

Animals learn from the past to make predictions. These predictions are adjusted after prediction errors i.e. after surprising events. Generally, most reward prediction error models learn the average expected amount of reward. However, here we demonstrate the existence of distinct mechanisms for detecting other types of surprising events. Six macaques learned to respond to visual stimuli to receive varying amounts of juice rewards. Most trials ended with the delivery of either one

or three juice drops so that animals learned to expect two juice drops on average even though instances of precisely two drops were rare. To encourage learning, we also included sessions, during which the ratio between one and three drops changed. Additionally, in all sessions, the stimulus sometimes appeared in an unexpected location. Thus, three types of surprising events could occur: reward amount surprise (i.e. a scalar reward prediction error), rare reward surprise, and visuospatial surprise. Importantly, we can dissociate scalar reward prediction errors – rewards that deviated from the average reward amount expected – and rare reward events – rewards that accorded with the average reward expectation but which rarely occurred. We linked each type of surprise to a distinct pattern of neural activity using fMRI. Activity in the vicinity of the dopaminergic midbrain only reflected surprise about the amount of reward. Lateral prefrontal cortex had a more general role in detecting surprising events. Posterior lateral orbitofrontal cortex specifically detected rare reward events regardless of whether they followed average reward amount expectations, but only in learnable reward environments.

2.1 Introduction

Animals including humans learn from the past to predict the future. This enables them to adjust to their environment and is critical for survival. One type of prediction that animals make concerns the reward value of potential choices that might be taken [99, 100]. After the choice and the outcome is experienced, the correctness of the original expectation is evaluated. When the outcome is better than expected then there is a positive prediction error (PE) and the animal should revise its estimate of the choice's future value upwards. When the outcome is worse than expected – there is a negative PE – the animal should revise its future estimate of the choice's value downwards.

Neurophysiological investigations in animals have shown that areas such as the dopaminergic midbrain encode reward PEs [1, 56, 99, 101–103]. In addition, human functional magnetic resonance imaging (fMRI) studies have also examined whether activity throughout the brain reflects reward [104–110] PEs. While some

neuroimaging studies have found evidence for PE coding in the dopaminergic midbrain, they have also reported PE coding in other structures including striatum and prefrontal cortex. One aim of the current study was, therefore, to investigate the nature of PE coding in macaque prefrontal and cingulate cortex [111, 112] and compared it with that seen in other parts of the brain.

We also, however, intended to address a more general debate about the precise nature of the reward expectations that animals hold. While it is clear that they hold representations of average expected reward size, they also hold more specific representations about the nature and identities of the outcomes that they hope to receive after a choice [113–121]. If that is the case then it might be anticipated that they would also experience PEs not just about the amount of reward – scalar reward value – but about other features of the reward [122].

It is possible that PE-related activity may actually reflect either scalar reward PEs or PEs about other features of a surprising reward experience [123, 124]. In the current study we disentangled two distinct types of reward PEs reflecting the richness of the reward representations held by monkeys. Specifically, we wanted to look at rare reward event (RRE) signals that might indicate that an unusual reward event has occurred. If such signals exist then they would complement the scalar reward PE (sRPE) signals typically studied. Alongside mean reward level, the frequency at which a given reward occurs is also important information that should be keenly monitored for learning purposes. In other words, animals may have a representation of the most frequent or modal reward levels, in addition to the mean reward level, to be expected from an average trial to guide their behaviour [125]. Pierce and Hall [126] suggested reward novelty is an important driver of the associability of any stimulus.

In the current experiment, it was possible to motivate behaviour with primary reinforcers delivered directly to the animals' mouths. Because of the animals' training with relatively small but discriminable quantities we could vary reward magnitudes parametrically and have a sufficient number of trials for animals to learn and change their reward expectations. To produce RREs it was necessary

to dissociate the reward frequency from absolute PEs. Therefore, we inverted the usual experience of receiving rewards close to the mean reward level most frequently. We did this by making actual instances of reward at the mean level very unlikely. This was achieved by devising schedules delivering mostly either one or three drops but occasionally delivering two drops. In such a schedule the average reward expectation is close to two drops but actual instances of two drops occurred only rarely. In other words, the reward distribution was a bimodal distribution with two equal peaks either side of the mean. Thus, while the delivery of a two drop reward would entail no sRPE, because it matched the average reward amount expectation, it should constitute a large RREs because the delivery of two drops is such a rare and novel experience. Careful experimental design ensured that sRPE and RREs shared only 0.049% of variance so that their neural correlates could be dissociated from one another.

However, as the world is full of minor or currently irrelevant changes, it is beneficial to selectively monitor reward frequency only when actively learning. Thus, we also examined whether increasing reward volatility (known to boost learning) [41, 127], also increases attention to other reward features such as RREs. We employed two types of schedules that we refer to as "changing/learnable" and "stable/unlearnable" (Fig 2.1AB). Animals should learn to expect either a higher or lower average reward in different parts of the changing/learnable schedules, potentially rendering a RRE more surprising or noteworthy because there is less uncertainty about whether to expect one or three drops of juice. By contrast, in the stable/unlearnable sessions, the two primary outcomes remain equally likely throughout so no outcome is more frequent or expected and a RRE should be less surprising too.

In human neuroimaging studies it is often difficult to determine if all PE-related activity reflects encoding of the reward PE *per se*. This is because it is difficult to motivate human subjects to perform tasks in the MRI scanner with primary rewards. Instead task performance is typically motivated by visual cues that act as secondary, or other higher order cues that predict money given to the participant

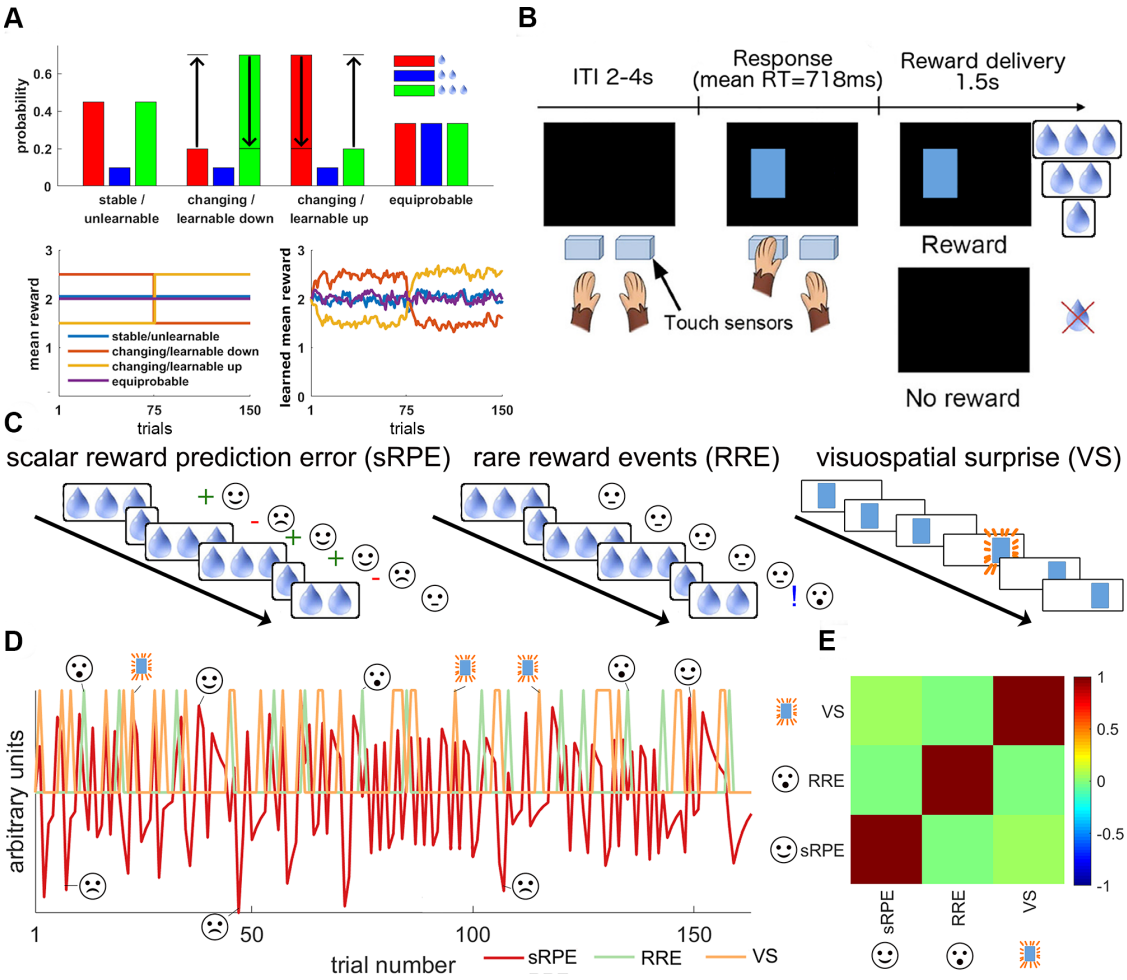


Figure 2.1: **Trial and task structure.** (**A**) Four reward schedules were employed. All had the same mean reward over the whole session but in the learnable sessions (changing/learnable up; changing/learnable down) the average reward reversed halfway. RREs, which had the same scalar value, two drops, as the mean reward, only had a low probability of occurrence, $P(RRE) = 0.1$, except in the equiprobable condition where they were just as likely to occur as one and three drop outcomes (top panel). In changing/learnable sessions monkeys have to learn that the average reward is higher/lower than expected at the beginning of a session and reverses halfway through the session (bottom panels). (**B**) On each trial a visual cue (blue rectangle) appeared on either the left or the right of the screen. This instructed the monkey to make a hand response towards a touch sensor next to the corresponding side of the screen. If they responded correctly then they received a juice reward of one, two, or three drops in size. (**C**) Illustrations of the three surprising effects of interest. sRPEs (left) occur when the obtained reward is better (i.e. three drops) or worse (i.e. one drop) than the expected average (two drops). RREs (middle) occur when an infrequent reward is obtained (i.e. two drops). VS (right) occurs when the stimulus switches sides. (**D**) An example (stable/unchanging) session illustrating how sRPEs, RREs, and VSs occur on different trials. Trials on which each of the three types of surprise occur are marked. (**E**) Correlation matrix between the three main effects of interest showing that our task design allowed for separately examining the effects of sRPEs, RREs, and VSs.

at the end of the experiment. It is therefore possible that apparent reward PE-related activity simply reflects the visual surprise associated with the appearance of a particular visual cue that is acting as a secondary reinforcer rather than the surprising level or nature of the reward it predicts [123, 128]. That visual surprise may be confounded with reward PE is an important consideration because it has been argued that dopaminergic neuron responses reflect the salience of cues and not just reward PEs [129]. While the salience of cues may be affected by many factors such as physical intensity, visuospatial surprise – appearance at a surprising location – may also contribute to salience. In addition to comparing sRPEs and RREs, we also considered whether reward PE and visuospatial surprise are encoded in the same manner by the same brain structures.

Specifically, we designed our experiment to concurrently measure the impact of visuospatial stimulus surprise (VS) and two distinct types of primary reward PEs. To infer levels of neural activity associated with these different error signals across the frontal cortex and striatum, we used fMRI. By studying macaques, it was possible to examine neural responses to PEs concerning primary reward that were of consequence and interest for the animals instead of visual tokens like those typically used in human neuroimaging. By carefully designing the order in which visual stimuli and rewards were delivered not only did we decorrelate sRPEs and RREs, we also decorrelated both sRPE and RREs from VS. It was therefore possible to identify neural activity associated with each type of surprising events.

2.2 Results

2.2.1 Behaviour

We wanted to investigate the behavioural and neural effect of spatial surprise and different types of reward surprise. We therefore designed an experiment that enabled us to look at three effects separately: sRPE, RRE, and VS (Fig 2.1C).

On each trial, animals were presented with a single blue rectangle appearing on either the left or the right side of the screen. VS could be examined because the common stimulus side reversed periodically with occasional rare stimuli appearing

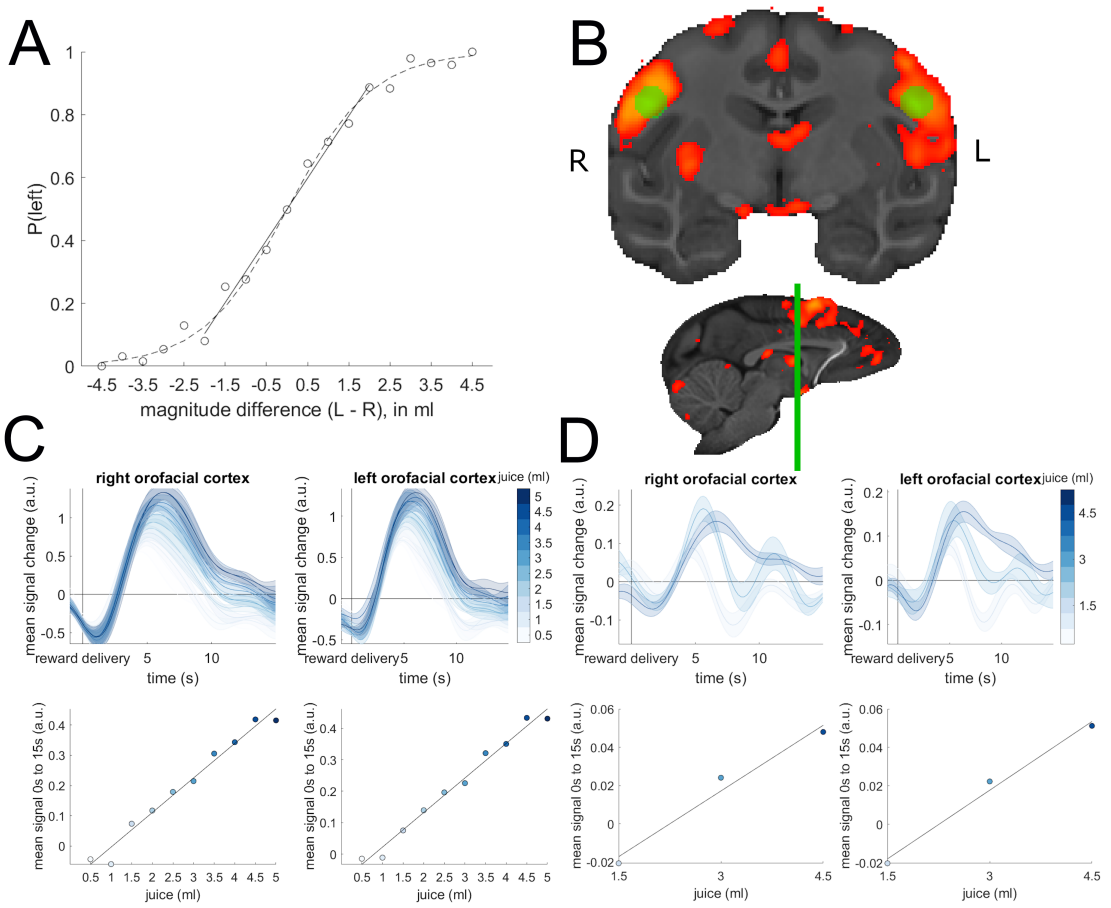


Figure 2.2: (Continued on the following page.)

on the opposite of the common location. Animals received a juice reward of one, two, or three drops after touching the response sensor on the side corresponding to the stimulus (Fig 2.1B). To first validate that monkeys can reliably distinguish between such juice amounts, we reanalysed data from a different task and showed that it was indeed the case when both behavioural and neural data were analysed (Fig 2.2). The schedules were designed so that the average reward expectation across the session was for two juice drops to be delivered. Delivery of three or one drops therefore constituted positive and negative sRPEs respectively (Fig 2.1CD).

The juice rewards followed four different schedules that also allowed us to study the effects of RRE: during stable/unlearnable sessions (Fig 2.1A) the mean reward was constantly kept at two drops, but actually two drops of juice were only delivered in 10% of trials. This made receiving two drops of juice a surprising event or

Figure 2.2: **Neural and behavioural evidence that macaques can distinguish between juice amounts.** To validate that monkeys can reliably distinguish between juice amounts of 1.5 ml, we reanalysed data from a visual discrimination task. In this task, monkeys (n = 4) had to select one of 2 displayed stimuli that were probabilistically rewarded with a juice amount between 1 and 10 drops (0.5 ml per drop). The colour of the stimuli varied in shade between blue and green. (**A**) The proportion of times the left stimulus was chosen as a function of the difference in reward magnitude (in ml) between the left and the right stimulus. For the slope of the curve in the central region (between magnitude differences of -2 ml and 2 ml), we can see that discrimination performance improves by 9.89% on average per drop (solid line; $t(7) = 23.08, p < 0.001$). (**B**) The thresholded and cluster-corrected map of activity covarying with reward from a whole-brain analysis of this task (shown with a threshold of z = 4.5). The cluster shows prominent bilateral activity in the orofacial sensorimotor cortex. We placed spherical ROIs at the most posterior local maxima in both the left and right orofacial activity (i.e., in the somatosensory cortex indicated by green ROIs; F99 coordinates 22.6, -4.02, 11.1 and -23.1, -3.51, 11.1) and extracted the BOLD time courses. (**C**) The extracted time courses from the ROIs indicated in (B) split by the amount of juice received (from 0.5 ml to 5 ml in 0.5-ml intervals) after reward delivery (top) and averaged over a window of 15 s (bottom). BOLD activity becomes stronger the more juice monkeys received in both the right $\chi^2(1) = 20.38, p < 0.001$) and left $\chi^2(1) = 17.687, p < 0.001$) orofacial somatosensory cortex. (**D**) The extracted time course from the same (now a priori) ROIs for the present study, again split up by the juice amount the monkeys received (1.5 ml, 3 ml, and 4.5 ml). As can be seen, BOLD signals are larger the more juice the monkey receives in both the right $\chi^2(1) = 8.892, p = 0.003$) and left $\chi^2(1) = 10.984, p < 0.001$) orofacial somatosensory cortex. We thus conclude that our monkeys can reliably distinguish between 1, 2, and 3 drops of juice and that distinct activity patterns associated with each outcome are available as inputs to any neural learning mechanism in the present study. The averaged BOLD time course for both the left and right orofacial area is also shown in Fig 2.7E.

RRE, even though it corresponded to the average expected reward for every trial. Five animals completed six of these sessions and one animal completed five of these sessions. However, a weakness of this schedule is that reward monitoring and learning may be minimal or ineffective because the schedule is static and the reward environment is unchanging [41, 127]. We therefore created two further schedules that were more changing in nature – changing/learnable sessions – in which two drops were again delivered on only 10% of trials but the average reward changed up or down halfway through a session (Fig 2.1A). Once again, the two drops of juice reward correspond to the average reward across the session (although

the actual average was below or above depending on the part of the schedule; see Fig 2.1A lower panel). Once again because of their rarity, actual two drop reward occurrences are RREs. Each of our monkeys completed four of these sessions. We refer to these sessions as 'changing/learnable' in contrast to the 'stable/unlearnable' sessions because of two features they have: first, we expect monkeys to have formed strong priors about two drops of juice as the average reward amount because of the high number of stable/unlearnable sessions. Thus, when they encounter a changing/learnable session where the mean reward amount starts at either 1.5 or 2.5 drops, they must re-learn average reward expectations. Moreover, once the mean reward changes halfway through a session, they yet again must re-learn average reward expectations (Fig 2.1A bottom right). The second reason we call these sessions 'changing/learnable' is that the uncertainty about juice amount (one, two, or three drops) that can be reduced in these sessions via learning is greater than in the stable/unlearnable sessions. This is because the inherent irreducible uncertainty about juice amount that is built into the schedules is greater for the stable/unlearnable than for the changing/learnable sessions: in the stable/unlearnable sessions the monkeys can at best figure out that one and three drops of juice are equally likely (45% and 45%; Fig 2.1A top) and thus expect them with equal probability. In contrast, in the stable/unlearnable sessions the monkey can figure out that on any trial the probability of one specific outcome is 70% and the other two outcomes are unlikely (10% and 20%; Fig2.1A top) and thus form less uncertain expectations about what juice amount to expect.

Finally, we included a small number of sessions of an additional control condition during which receiving one, two, and three drops of juice was equally likely (Fig reffig:21A); now two drop rewards are no longer RREs. Each of the six monkeys completed two of these sessions.

We wanted to assess whether factors related to VS, sRPE, and RRE affected behaviour. To do so we ran a generalised linear mixed-effect model (GLME1, Fig 2.3A; see Materials and Methods), predicting whether a performance lapse occurred (either an error response not directed to the correct side of the screen or an outlier

response [any trial that had an unusually short (< 50ms) or long (> 4000ms) response time (RT) indicative of task disengagement] on any given trial. Our six monkeys lapsed 15.95%, 15.17%, 11.80%, 10.24%, 9.89%, and 6.89% on average over all sessions they completed. It is intuitive that VS or negative/low sRPE events will be associated with performance lapses, but it is less clear that this will be true of RRE. It is, however, likely that performance lapses will diminish in frequency when scalar reward levels are high. We therefore combined the previously received rewards into a single regressor of scalar reward expectation (sRE) for the current trial through a Rescorla-Wagner learning model. Our results show that a positive sRE on the previous trial decreases the likelihood of a performance lapse on the current trial; in other words, greater reward expectation on the previous trial decreases the likelihood of a performance lapse on the current trial ($\chi^2(2) = 6.918, p = 0.031$; two degrees of freedom for fitting the learning rate and the inverse temperature; Fig 2A first column). To estimate the learning rate we used to calculate the sRE, we used another GLME (GLME2, Fig 2.3B) that included the reward on the previous five trials as individual regressors, and then estimated a Rescorla-Wagner type reinforcement learning model from these beta weights by finding the Rescorla-Wagner model that best describes the observed weights on previous rewards. We thus obtained a learning rate ($\alpha = 0.257$). This procedure is broadly equivalent to fitting a truncated reinforcement learning model which also controls for other confounds (see Materials and Methods for details). To separately test for an effect of reward that decayed with the distance in the past (reward history), we also used GLME2 to fit a line through the beta weights of the previous five rewards for each monkey separately and then tested whether these slopes differ from zero. We indeed found a consistent effect of reward history ($t(5) = 10.264, p < 0.001$).

We did not find any significant effect of having experienced a RRE on the previous trial when all data were considered in aggregate ($\chi^2(1) = 0.004, p = 0.950$ in GLME1; third column in Fig 2.3A, seventh column in Fig 2.3B). However, we also performed these analyses for different session types (Fig 2.4). While it is difficult to anticipate the impact, if any, that RREs might have on performance

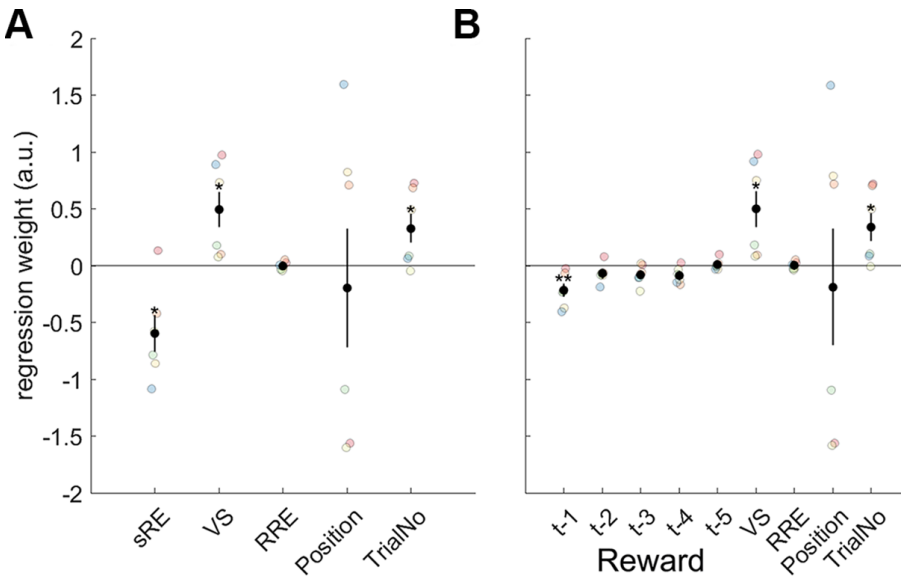


Figure 2.3: **Behavioural effect of surprising events.** GLMEs predicting errors or outliers as a function of three types of surprising event: sRE, VS, and RRE. The two GLMEs only differ in that, in (**A**), sRE is indexed by a single regressor while in (**B**) sRE is not a single regressor but instead its component parts are made explicit in terms of the reward outcome experienced on the last five trials (t-1 to t-5). Dots represent the beta weight associated with each regressor determined by a GLME applied to each monkey separately. Black dots indicate group mean effects according to the full GLMEs and vertical bars indicate the standard errors of the means (SEMs). The Position effect indicates that animals were more likely to make errors or outlier responses depending on the side of the screen they had to respond to but that different animals had different side biases. Errors and outliers became more likely as each session progressed and trial number (TrialNo) increased.

lapses, we noticed that RREs were more likely to be followed by performance lapses in the changing/learnable compared to the stable/unlearnable sessions in five of the six animals although the change occurred in the opposite direction in the sixth animal. A further set of GLMEs were used to predict RT as opposed to performance lapses (Fig 2.5) and revealed VS slowed RTs ($\chi^2(1) = 9.409, p = 0.002$; GLME3.) Finally, both GLMEs (GLME1 and GLME2) also revealed significant effects of VS: lapses of performance were more likely to occur if the stimulus appeared in an unexpected location ($\chi^2(1) = 5.837, p = 0.016$ in GLME1; second column in Fig 2.3A, sixth column in Fig 2.3B).

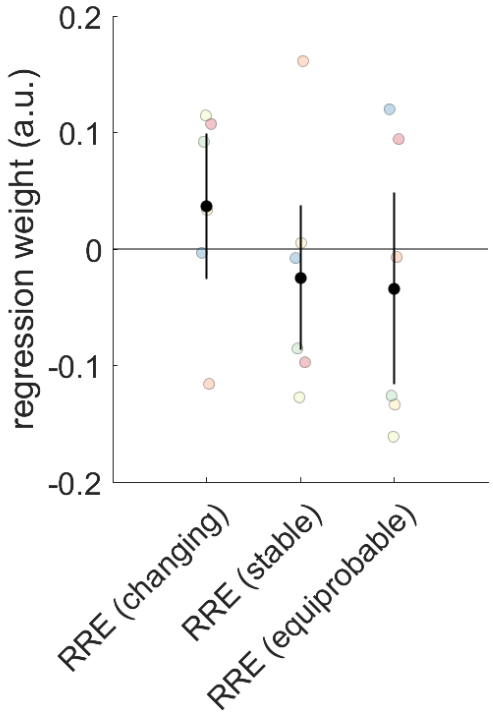


Figure 2.4: **Behavioural differences of RRE between session types.** To assess whether behavioural RRE effects were modulated by session type, we ran GLME1 separately for changing/learnable, stable/unlearnable, and equiprobable sessions. The effect was not significant in changing/learnable sessions ($\chi^2(1) = 0.318, p = 0.573$; left column), stable/unlearnable sessions ($\chi^2(1) = 0.150, p = 0.699$; middle column), or equiprobable sessions ($\chi^2(1) = 0.134, p = 0.714$; right column).

2.2.2 FMRI

In our experiment, expectations could be violated in three main ways: VS, sRPE, RRE. To identify brain areas associated with these three types of surprising event, we ran a three-level multiple regression analysis by employing a general linear model (GLM). For each monkey we used a fixed-effects model between session and applied the FLAME 1+2 procedure from the FMRIB Software Library (FSL) on the highest hierarchical level (level combining animals). We focus on effects on frontal cortex, striatum, and in the vicinity of the dopaminergic ventral tegmental area and substantia nigra (VTA/SN) in the midbrain because these areas are the ones that have been most frequently related to reward value expectation and prediction error coding in both human and non-human primates [53, 100, 110,

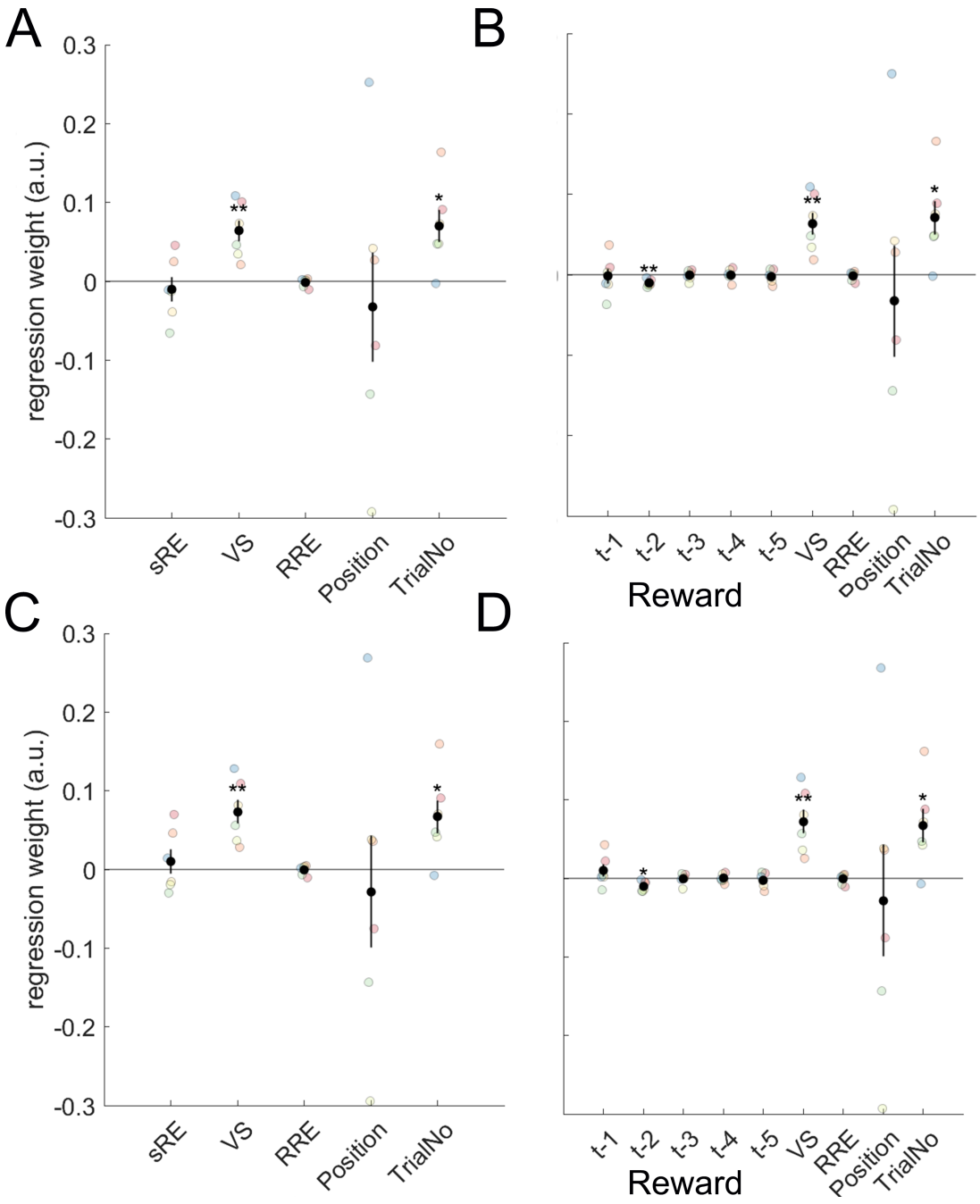


Figure 2.5: **Behavioural effects on RTs.** We also ran GLMEs with RTs as the dependent variable that contained the same independent variables as GLME1 and GLME2. For GLME3 (**A**) and GLME4 (**B**), outlier trials were excluded. GLME3 and GLME4 only differed in that in GLME3, the rewards from the previous 5 trials are combined into the sRE using a learning rate estimated from GLME2 as described in the main text. In GLME3, VS was significant ($\chi^2(1) = 9.409, p = 0.002$), whereas sRE ($\chi^2(1) = 0.433, p = 0.511$) and RRE ($\chi^2(1) = 0.185, p = 0.667$) were not. In GLME5 (**C**) and GLME6 (**D**), we excluded all outlier, error, and repeat trials (trial after an error). Again, in GLME5, VS was significant ($\chi^2(1) = 9.631, p = 0.002$), whereas sRE ($\chi^2(1) = 0.444, p = 0.505$) and RRE ($\chi^2(1) = 0.030, p = 0.862$) were not.

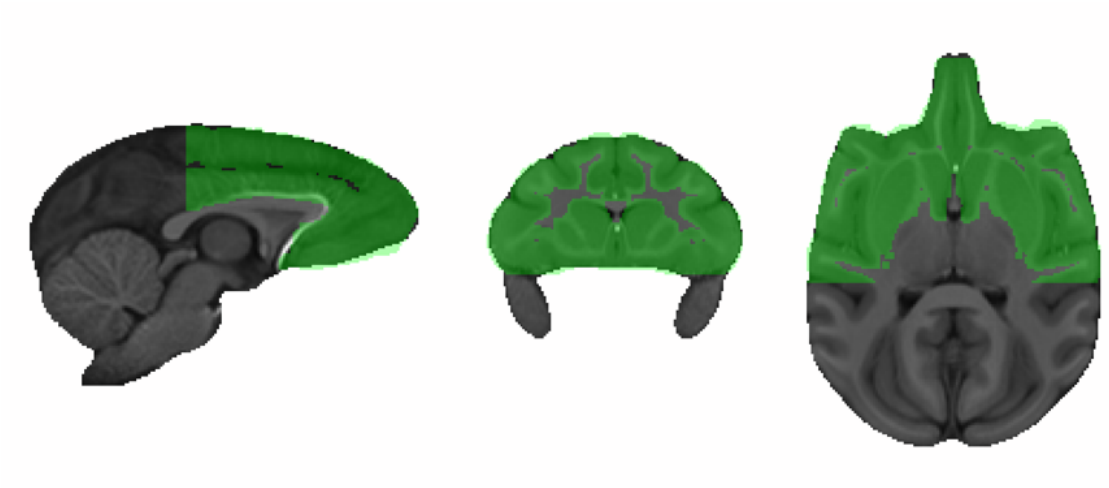


Figure 2.6: **VOI covering the prefrontal cortex and striatum.** As we were primarily interested in prefrontal cortex and striatum and the nature of their PEs, we specified a VOI covering both. All whole-brain analyses in the chapter were carried out using this VOI.

130–134]. For this reason we analysed data in a volume of interest (VOI) covering frontal cortex and striatum (Fig 2.6), and a precisely localised region of interest (ROI) in the much smaller dopaminergic midbrain region VTA/SN. To further examine the co-occurrence of effects within the VOI, we place functional ROIs at peaks of the sRPE, RRE and VS effects.

In this way we attempted to ensure that we were both able to detect surprise and prediction error responses wherever they occurred in the striatum or prefrontal cortex; such responses have previously been reported in several sub-regions in these structures, moreover within the striatum it is not clear that the strongest prediction error/surprise signals can be mapped onto just the ventral striatum, caudate, or putamen. At the same time, the approach increased the statistical power of our analyses to examine neural activity in brain regions that were of *a priori* interest.

We anticipated that the VTA/SN's smaller size would preclude other analysis approaches commonly used in fMRI such as spatial cluster-based statistics that are most beneficial when there are large areas of activation, however, the activity in VTA/SN, as in the striatum and frontal cortex was so prominent that it could also be identified using a standard cluster-based correction procedure for multiple comparisons corrected across the whole brain. Thus in addition, whole brain

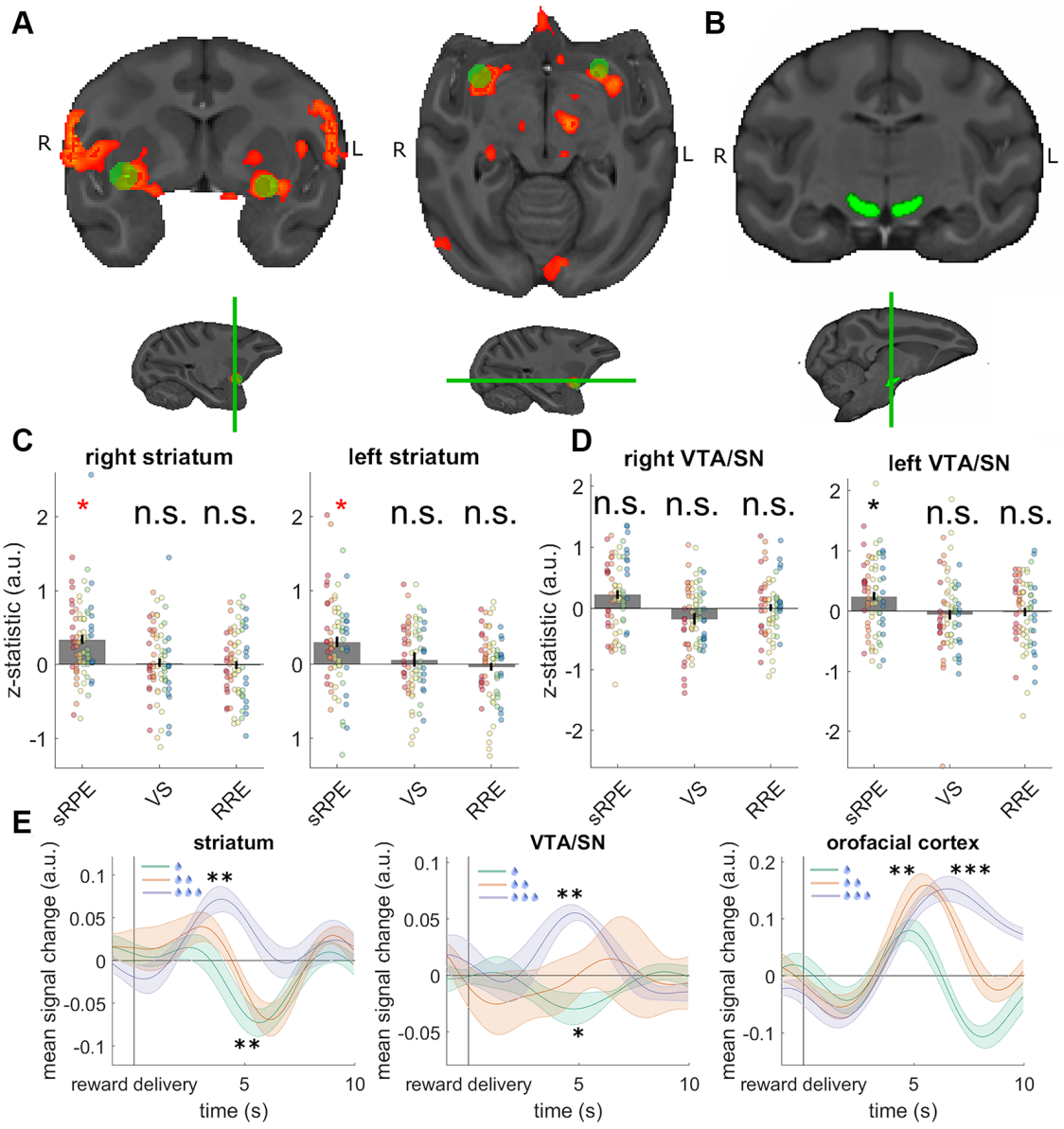


Figure 2.7: (Continued on the following page.)

cluster-corrected results are reported in Tables A.5,A.6, and A.7 and are discussed briefly in the Discussion.

We used data from 71 of the 76 sessions we had acquired (fMRI data from one session were corrupted and unrecoverable). For sRPE, VS, and RRE our hypotheses necessarily focused on activity that was positively related to each type of surprising event, but we also note other patterns of activity where they existed.

To look for evidence of sRPE signals, we first performed an analysis in our VOI in the frontal cortex and striatum and in our ROI of interest in the VTA/SN. The

Figure 2.7: (sRPEs in the striatum and VTA/SN. (**A**) Prominent sRPE effects were observed in the right and left orofacial somatosensory and motor cortex and ventrolateral striatum extending into the nucleus basalis of Meynert. (**B**) We used a priori ROIs in the right and left VTA/SN to extract z-statistics of the regressors from the whole-brain analysis for each session. (**C**) The z-statistics of each session from a spherical ROI placed at the peak activity of the cluster encompassing the right and left ventrolateral striatum. Different colours indicate data from different animals, with the grey bar showing the grand mean. A red asterisk indicates significance according to the whole-brain analysis, and a black asterisk indicates significance according to a test on the extracted z-statistics. The ROI-based analyses illustrated the presence of the sRPE effects in the left and right striatum but revealed no evidence for VS or RRE signals at the same locations. (**D**) The z-statistics for the ROI in the right and left VTA/SN revealed an overall significant effect of sRPE but no effects of VS or RRE. When testing the right and left VTA/SN separately, the z-statistics for sRPE in the left VTA/SN were significant, with no effects of VS or RRE. The right VTA/SN revealed a similarly signed sRPE effect although it was, on average, smaller in size and there was more variation across individuals and sessions. Once again VS and RRE effects were not significant in right VTA/SN. (**E**) BOLD time courses extracted from 3 ROIs in the ventrolateral striatum, the SN/VTA, and the orofacial cortex. The time courses are averaged over both hemispheres. For the location of the orofacial cortex ROI see Fig 2.2. Time courses are illustrated for each level of juice reward the monkey received (1, 2, or 3 drops). The orofacial area activity reflects reward amount, and thus all three time courses exhibit an initial positive peak. In contrast, ventrolateral striatum and SN/VTA process sRPEs, and thus, the time course for receiving 1 drop of juice—which is associated with a negative sRPE—results in suppressed activity (a negative activity change).

analysis revealed four main clusters of activity (cluster $p < 0.05$, cluster forming threshold of $z > 2.3$; Fig 2.7A; see Table A.1 for cluster locations). While this analysis approach, focused on a priori areas of interest, was our primary one, we note that the same results were evident in a whole-brain cluster-corrected analysis (Table A.5). Two of these clusters were located in the left and right ventrolateral striatum respectively, and two in the left and right ventral sensorimotor cortex near the region occupied by the orofacial representation. Extracting the z-statistics of each session from the individual regressors of our whole-brain analysis from spherical ROIs with a 7.5mm diameter in the left and right striatum respectively confirmed a clear effect of sRPE (Fig 2.7C). Next, we examined whether the same ROI carried information about VS or RRE. However, further analysis of the extracted z-statistics revealed no effect of VS in the left ($\chi^2(1) = 0.378, p = 0.539$; second column right panel Fig 2.7C)

or right ($\chi^2(1) = 0.133, p = 0.715$; second column left panel Fig 2.7C) ventrolateral striatum. We also found no effects of RRE in the left ($\chi^2(1) = 0.485, p = 0.486$; third column right panel Fig 2.7C) or right ($\chi^2(1) = 0.023, p = 0.88$; third column left Fig 2.7C) ventrolateral striatum. Finally, we examined whether VS or RRE might exert a significant influence on activity but only in a specific session-type (stronger learning effects are predicted in the changing/learnable sessions [41, 127]). This approach, however, failed to find any evidence for VS or RRE coding in striatum even in the changing/learnable sessions (Fig 2.8).

To test whether a sRPE signal was present in the BOLD activity in the ROI placed over the dopaminergic VTA/SN region in the midbrain, we warped an *a priori* defined mask of the left and right VTA/SN into session space and extracted the z-statistics from the regressors of our whole-brain analysis for this ROI for each session (Fig 2.7B; although note, as already mentioned, the effects were sufficiently strong to survive whole brain cluster correction; Table A.5; Fig 2.9). We found an overall effect of sRPE in both the right and left VTA/SN, while controlling for the different hemispheres ($\chi^2(1) = 5.940, p = 0.015$; left columns in both panels of Fig 2.7D) but found no effects of VS ($\chi^2(1) = 2.187, p = 0.325$; central columns in both panels of Fig 2.7D) or RRE $\chi^2(1) = 0.021, p = 0.885$; right columns in both panels of Fig 2.7D). When testing the right and left VTA/SN separately, we found an effect of sRPE in the left VTA/SN $\chi^2(1) = 6.298, p = 0.012$; first column right panel of Fig 2.7D) although it did not reach significance in the right VTA/SN ($\chi^2(1) = 2.146, p = 0.143$; first column left panel in Fig 2.7D).

Finally, we considered the possibility that the sRPE effect might simply be artefact of the learning rate used when estimating monkeys' reward value expectations. In line with observations previously made by Wilson and Niv [135], a control analysis provided evidence for sRPEs in the VTA/SN regardless of the precise learning rate used (Fig 2.10).

We found no effect of VS in the left ($\chi^2(1) = 0.635, p = 0.425$; second column right panel of Fig 2.7D) or right $\chi^2(1) = 2.434, p = 0.119$; second column left panel Fig 2.7D) dopaminergic midbrain. We also found no effects of RRE in

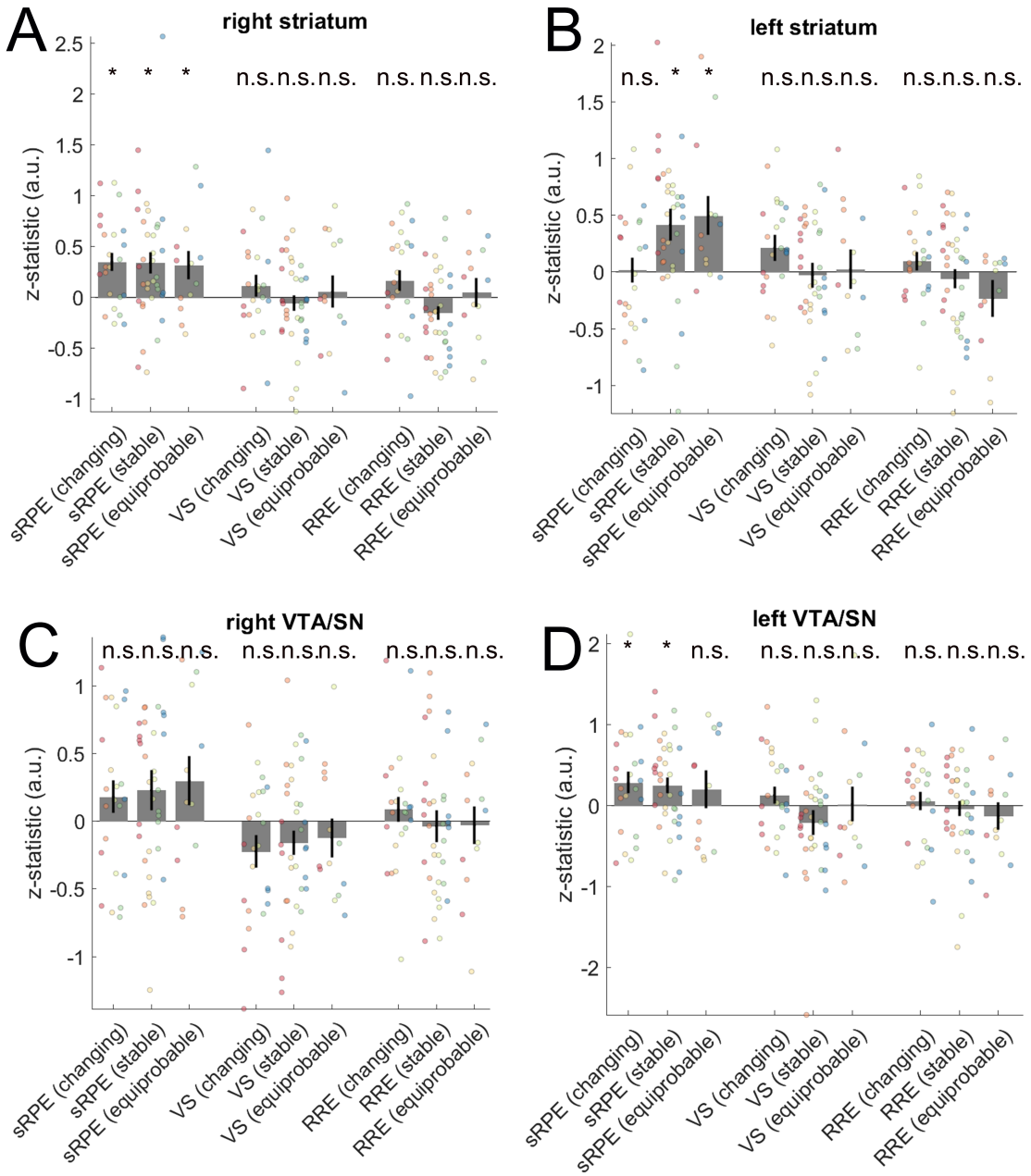


Figure 2.8: **Neural effects of sRPE by session type.** We examined the extracted z-statistics shown in Fig 2.7, here split up by session type. Labelling conventions are the same as in Fig 2.7.

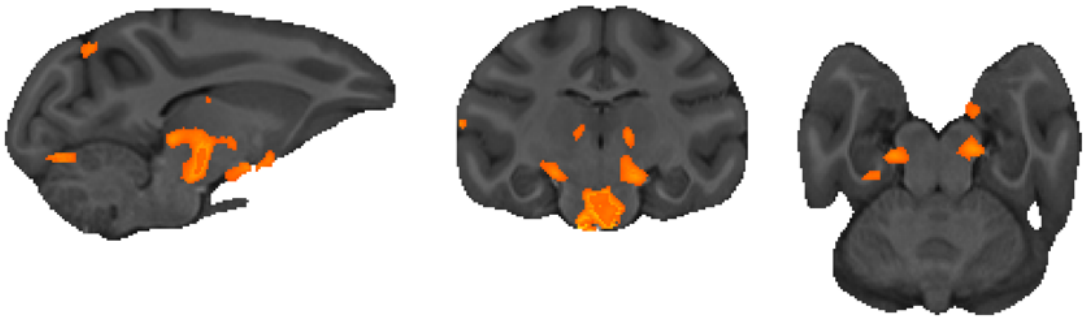


Figure 2.9: **Whole-brain results for sRPE in the dopaminergic midbrain.** When running a cluster correction on the whole brain (without the VOI) for our sRPE regressor we also found activity in the dopaminergic midbrain. The precise location of these clusters can be found in Table A.6.

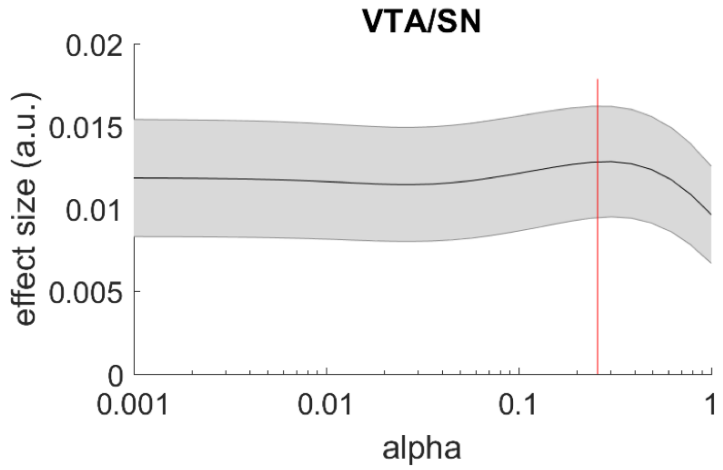


Figure 2.10: **VTA/SN sRPE signals do not depend on the learning rate.** To consider the possibility that sRPE signals might be an artefact of the learning rate we used when estimating the monkeys' reward value expectations, we examined the BOLD response in the VTA/SN ROI further. By running a linear regression on the extracted BOLD time course time-locked to reward delivery (with terms coding for a constant, sRPE, VS, RRE, and trial number), we can determine the effect size of the sRPE regressor when constructing it using different learning rates. Effect size is here defined as the averaged beta weights for regressions run on the 10 s after reward delivery. The red line indicates the learning rate we used in the main text (0.257). As can be seen, the empirical learning rate is close to the peak effect strength in VTA/SN, but the effect strength is positive throughout regardless of learning rate.

the left $\chi^2(1) = 0.076, p = 0.783$; third column right panel of Fig 2.7D) or right $\chi^2(1) = 0.025, p = 0.874$; third column right panel of Fig 2.7D) dopaminergic midbrain. Finally, we examined the effects of sRPE, VS, and RRE in specific session-types and (Fig 2.8) in order to test whether VS or RRE effects might be present in the changing/learnable sessions in which learning effects were expected to be stronger; no evidence of their presence was found.

To illustrate these results, we extracted BOLD timecourses from the ROIs in the ventrolateral striatum, the SN/VTA and a control region in the orofacial cortex (Fig 2.7E; see Fig 2.2 for the location of the control region). When splitting up the timecourses by the amount of juice received (1, 2, or 3 drops) we observe suppressed activity after receiving one drop in the ventrolateral striatum ($\chi^2(1) = 10.219, p = 0.001$) and the SN/VTA ($\chi^2(1) = 6.250, p = 0.012$), which is due to the negative sRPE the monkeys experience. After receiving three drops of juice, which is associated with a positive sRPE, we observe enhanced activity in the ventrolateral striatum ($\chi^2(1) = 8.954, p = 0.003$, Fig 2.7E left) and the SN/VTA ($\chi^2(1) = 10.191, p = 0.001$, Fig 2.7E centre). In contrast, an area that processes reward amount such as the orofacial sensorimotor cortex shows a different activity profile: we observe no effect different from baseline for one drop of juice because of there is initially a positive BOLD response but this is quickly followed by a negative change ($\chi^2(1) = 2.1.03, p = 0.147$) but we observe positive activity after receiving two ($\chi^2(1) = 8.931, p = 0.003$) or three ($\chi^2(1) = 13.463, < 0.001$) drops of juice (Fig 2.7E right).

Next, we examined the effects of RRE in an analysis conducted in our VOI in the frontal cortex and striatum and in our ROI in the VTA/SN. Note, as already mentioned, careful experimental design ensured that sRPE and RREs shared only 0.049% of variance so that their neural correlates could be dissociated from one another. Moreover, by including terms relating to both sRPE and RRE in the same GLMs we ensured that activity actually related to sRPE could not be misinterpreted as activity related to a RRE. No effects survived cluster corrections when we combined all session types at the contrast level. However, when we

focused only on changing/learnable sessions, in which there was a possibility for animals to learn the changing statistics of the environment, we found a cluster of activation in the striatum and extending to the posterior lateral orbitofrontal cortex (plOFC). The activity was situated lateral to the lateral orbital sulcus just anterior to the anterior insula in or near area 47/12o and extended dorsally towards the ventral tip of the arcuate sulcus in or near areas ProM and 44 (Fig 2.11AB; Table A.2). The RRE-related activity in the striatum was situated in a more lateral and anterior area than was the case for the sRPE-related activity (Fig 2.7A). Again, this result was also apparent in a whole brain cluster-corrected analysis (Table A.6). Extracting the z-statistics of the regressors of our whole-brain analysis from spherical ROIs with a 7.5mm diameter placed at the peak activity of the cluster in striatum (Fig 2.11A) and a subpeak in plOFC (Fig 2.11B) illustrates the presence of the RRE effect in both areas (Fig 2.11CD).

Further analysis revealed no such RRE signals in the striatum ($\chi^2(1) = 0.435, p = 0.510$,last column Fig 4C) or in the plOFC ($\chi^2(1) = 1.026, p = 0.311$,last column Fig 2.11D) during stable/unlearnable sessions. We then examined whether the same ROIs carried information about sRPE or VS (Fig 2.11CD). We found evidence of a significant VS effect in the anterior lateral striatum ($\chi^2(1) = 7.729, p = 0.005$; Fig 2.11C) and a negative effect of sRPE in plOFC ($\chi^2(1) = 5.258, p = 0.022$; Fig 2.11D). We are cautious about over interpreting the latter effect given that it would not survive correction for multiple comparisons.

Another way to test for effects of RRE is to examine if activity during changeable/learnable sessions is significantly different from activity during equiprobable sessions. Running such an analysis on the whole brain level revealed similar patterns of activity in the anterior lateral striatum and plOFC (Fig 2.12; Fig 2.13; Table A.4). In particular, our plOFC cluster during changing/learnable sessions remains broadly similar when contrasting RRE effects with either of the other two conditions separately (changeable/learnable sessions versus equiprobable sessions and also changeable/learnable versus stable/unlearnable sessions (Fig 2.13)).

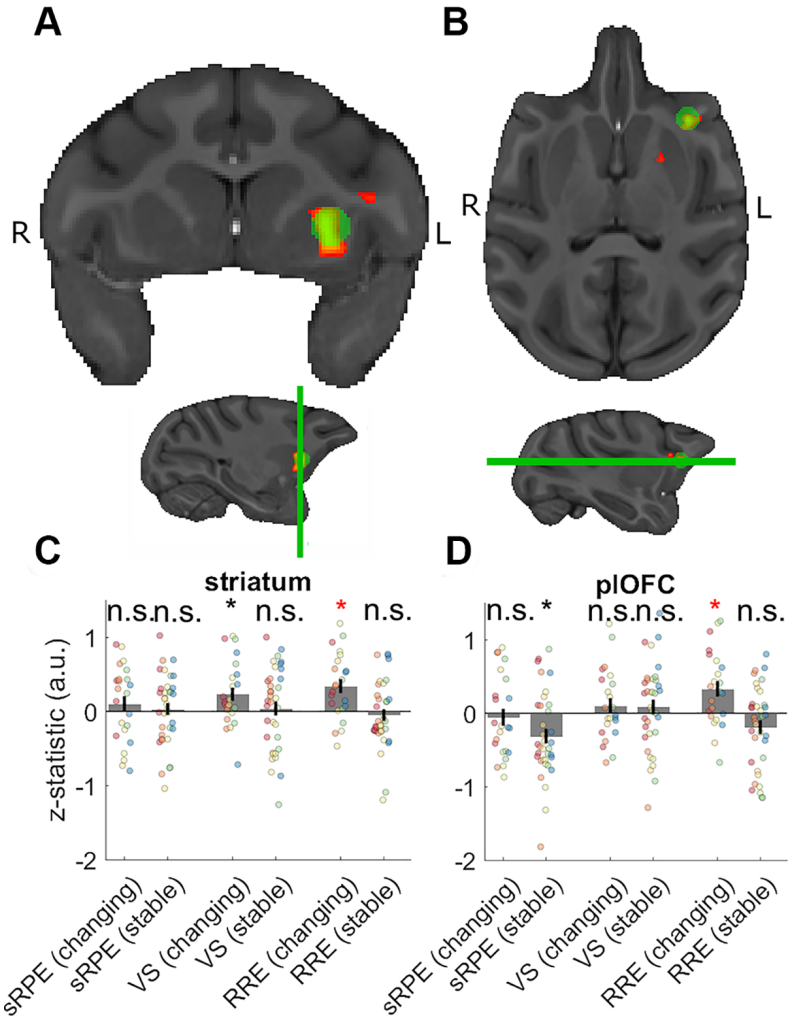


Figure 2.11: **Neural activity related to rare reward events.** (**A**) RRE effects in anterior lateral striatum during the changing/learnable sessions. (**B**) RRE effects in plOFC during changing/learnable sessions. The z-statistics of each session from a spherical ROI placed in (**C**) the striatum and (**D**) the plOFC. The ROI based analysis illustrates the effects of RRE during changing/learnable sessions. Additionally, we found a significant positive effect of VS during changing/learnable sessions in the anterior lateral striatum. There was also a significant negative effect of sRPE during stable/unlearnable sessions in the plOFC although we are cautious about over-interpreting this result as it would not survive correction for multiple comparisons. A negative sRPE indicates a stronger response when reward outcomes are worse than expected

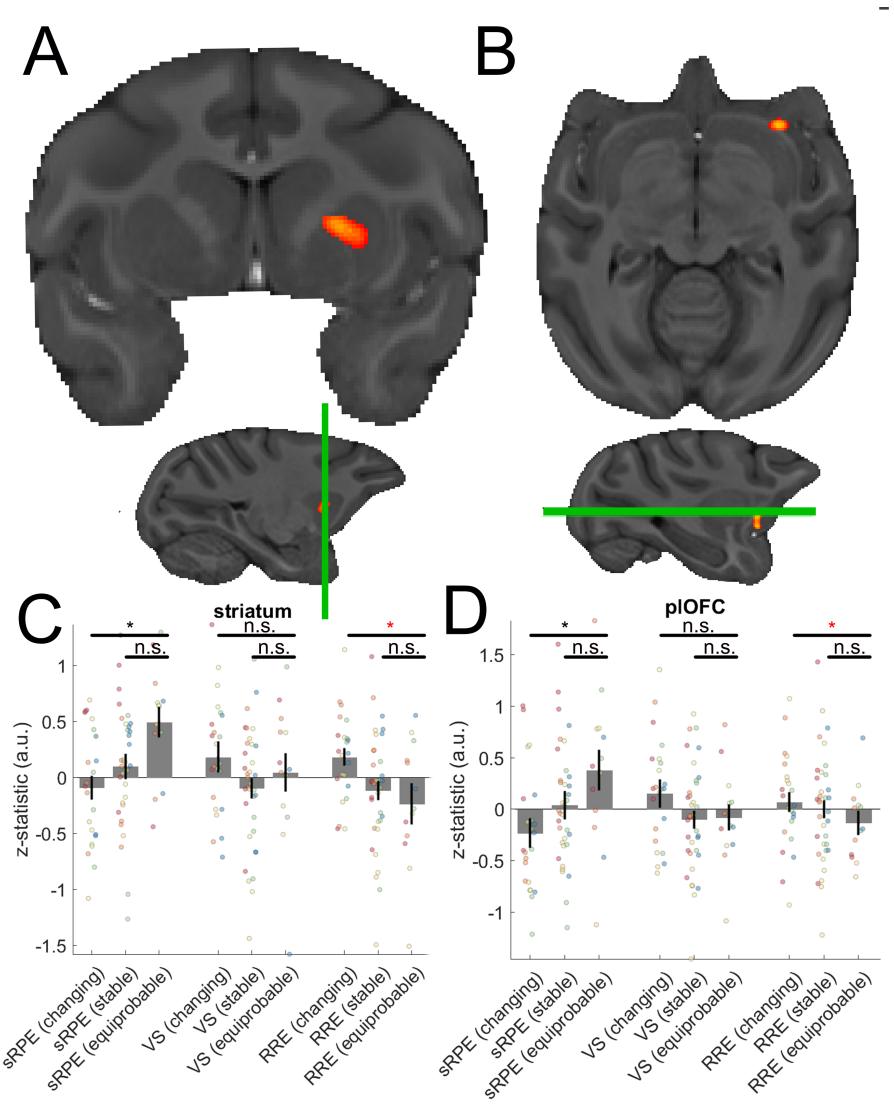


Figure 2.12: (Continued on the following page.)

Finally, we examined whether effects of VS were found at any other location in our VOI in the frontal cortex and striatum and in our ROI of interest in the VTA/SN. Four clusters were found. One cluster was found in lateral prefrontal cortex (lPFC; Fig 2.14A). Three other clusters were found just outside prefrontal cortex [the VOI was generously sized to include most of frontal cortex and adjacent tissue (Fig 2.6) so that no prefrontal signals of potential interest were overlooked.] These were situated in premotor cortex, secondary somatosensory cortex, and posterior cingulate area 23 (Table A.3). Again, the effect was strong enough to survive a whole brain cluster-corrected analysis (Table A.7). Extracting the z-statistics of

Figure 2.12: **Comparing neural RRE effects between session types.** As an alternative to the analysis shown in Fig 2.11, we also examined RRE by comparing changing/learnable and stable/unlearnable sessions against the equiprobable control sessions. (**A**) On the whole-brain level, we found significant activity in the striatum for a contrast comparing changing/learnable and equiprobable sessions. (**B**) We also found activity in the plOFC for the contrast comparing changing/learnable and equiprobable sessions although now the activity was even more centred on the boundary with the anterior insula. All local maxima of the cluster are shown in Table A.4. (**C**) Extracting the z-statistics from an ROI placed at the peak activity illustrates the difference for RRE between changing/learnable and stable/unlearnable sessions in the striatum (third column from the right and second column from the right). We did not find a significant difference between stable/unlearnable and equiprobable sessions for RRE ($\chi^2(1) = 0.516, p = 0.474$). Our analysis revealed a significant difference between changing/learnable and equiprobable sessions for sRPE in the striatum ($\chi^2(1) = 7.071, p = 0.008$) but with a different sign. For sRPE, the difference between stable/unlearnable and equiprobable was not significant ($\chi^2(1) = 3.160, p = 0.076$). For VS, neither the difference between changing/learnable and equiprobable ($\chi^2(1) = 0.110, p = 0.415$) nor between stable/unlearnable and equiprobable sessions ($\chi^2(1) = 0.516, p = 0.472$) was significant. (**D**) In the lOFC, we again did not find a significant difference between stable/unlearnable and equiprobable sessions for RRE ($\chi^2(1) = 0.665, p = 0.415$). For sRPE, there again was a significant difference between changing/learnable and equiprobable sessions in the opposite direction ($\chi^2(1) = 4.007, p = 0.045$). The difference between changing/learnable and equiprobable sessions was not significant for sRPE ($\chi^2(1) = 2.289, p = 0.130$). For VS, again neither the difference between changing/learnable and equiprobable ($\chi^2(1) = 2.552, p = 0.110$) nor between stable/unlearnable and equiprobable sessions ($\chi^2(1) = 0.002, p = 0.965$) was significant.

the regressors of our whole-brain analysis of each session from spherical ROIs with a 7.5mm diameter illustrates the presence of the VS effect in lPFC (Fig 2.14B).

Next, we examined whether the same lPFC ROI carried information about sRPE or RRE. Extracting the z-statistics from the lPFC revealed no effect of sRPE $\chi^2(1) = 0.120, p = 0.729$; first column Fig 2.14B) and no positive effect of RRE. There was, however, a significantly negative effect of RRE ($\chi^2(1) = 4.304, p = 0.038$; third column Fig 2.14B). A negative RRE effect is likely to indicate a response to a surprisingly large or small reward amount (sometimes referred to as an unsigned scalar reward prediction error). Additional analyses of specific session types are reported in Fig **??**. It is perhaps worth noting here that while the aspect of VS we consider here might contribute to the salience of a stimulus, so might other

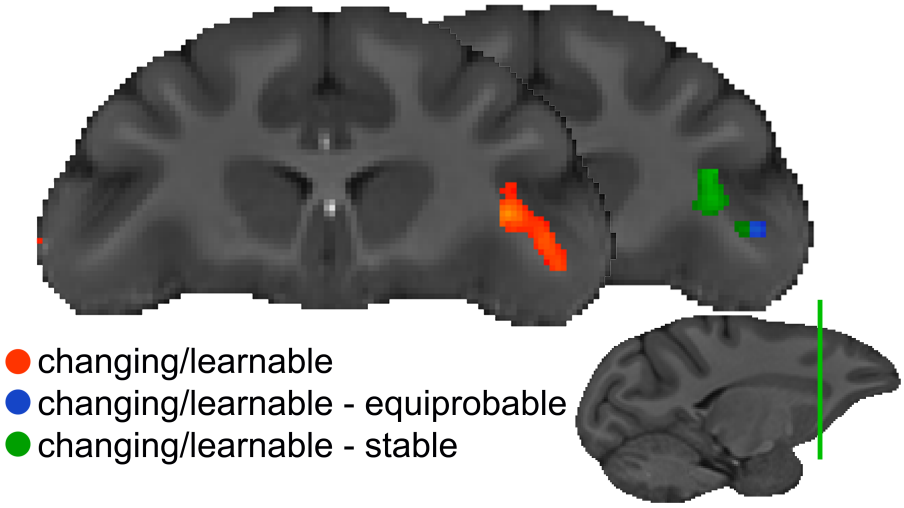


Figure 2.13: **Locations of neural RRE activity.** RRE activity in plOFC in changing/learnable sessions (red) is located in the same area as activity when we contrast changing/learnable and equiprobable sessions (blue) or changing/learnable and stable sessions (green). All effects shown here are at a threshold of 2 and are shown without applying any cluster correction.)

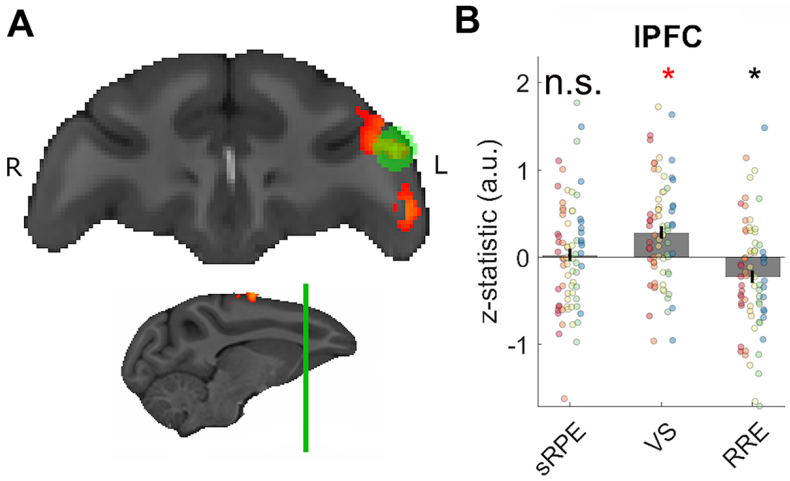


Figure 2.14: **Neural activity related to visual surprise.** (**A**) VS effects found in lPFC. (**B**) The z-statistics of each session from a spherical ROI placed at the peak activity of the cluster in lPFC. Labelling conventions are the same as in Fig 3. The ROI-based analyses confirmed the presence of the VS effects. There was no effect of sRPE. There was a significant negative effect of RRE which is indicative of encoding of surprise about the scalar reward value (it corresponds to an unsigned reward prediction error).

factors. While no VS effect was found in the vicinity of the dopaminergic midbrain, it is quite possible that other aspects of visual stimuli contribute to the transient response that has been noted to occur at the onset of a visual stimulus associated with reward. Schultz, Lak, and Stauffer [52, 53] refer to this response to the onset of a reward-predicting stimulus as being due to the "physical impact" of the stimulus. By contrast, the VS effect that we examined compared the neural response to identical stimuli that were equally unexpected in time, but which differed because they were either in an unexpected or expected spatial location.

2.3 Discussion

Animals, including humans, learn from the past to predict the future. New predictions are formed using prediction errors (PEs) that occur during surprising experiences. There has been considerable interest in the role that PEs about reward amount play on behaviour and the manner in which they are encoded by activity in the dopaminergic midbrain [99, 100]. Here, however, we found behavioural and neural evidence for three distinct types of PEs in a group of six macaques.

We ran a relatively simple experiment in macaques in which they touched a rectangle appearing either on the left or right side of the screen with correct responses delivering one, two, or three drops of juice. However, our key manipulations lay in the location of the target as well as the frequency of the juice amounts. Our schedules allowed us to examine the impact of visual PEs when the rectangle appeared on an unexpected side of the screen. Reward amount PEs (also known as scalar reward PEs) were triggered when the amount of juice deviated from the average of recent experiences (as this regressor is signed, lower rewards led to suppression of activity, while more reward let to increased activity). The delivery of two reward drops conformed with the average reward amount but at the same time it was the rarest outcome to occur, leading to a third form of surprising effect, i.e. rare reward events (RREs). The experimental design, therefore, allowed us to dissociate three types of surprising events, visuospatial surprise (VS), scalar

reward prediction errors (sRPE), and rare reward events (RREs) and uniquely link them to changes in behaviour and brain activity.

The three types of events were associated with different effects on behaviour. Macaques were more likely to have a lapse in performance following VS (they either made an error response to the wrong side of the screen or an outlier response with a speed outside the usual range). By contrast, they were less likely to lapse after a high scalar reward value (positive effect of sRE). The novel reward experience on RRE trials (receiving two drops), however, while rare and thus surprising, did not significantly affect behaviour in this very simple visual response task.

We did find consistently distinct neural responses for all three types of surprising events (Fig 2.15). sRPEs significantly modulated activity in the vicinity of the dopaminergic midbrain consistent with previous neurophysiological studies [1, 56, 99, 101–103] and human neuroimaging studies [104–110]. There was no evidence that VS or RRE significantly altered activity in the same region. The findings can be linked to those reported by Lak and colleagues [136] who examined macaque dopaminergic midbrain neuron responses to rewards that varied in type, amount, and riskiness. They showed that activity reflected PEs for the integrated subjective reward value of the outcome. In other words, when multi-attribute reward outcomes were experienced, the contributions of the attributes on the scalar, subjective reward value also determined the PE response.

Several other studies have examined whether the activity of dopaminergic neurons or the BOLD signal in the dopaminergic midbrain responds to changes in the type of reward experienced even when its scalar value remains the same [50, 137–139]. These studies have highlighted important similarities in the way in which the dopaminergic midbrain responds to both value and reward identity prediction errors. For example, across participants, there are correlations in the sizes of responses to value and identity prediction errors [50]. The RREs that we studied here are similar in some respects to identity prediction errors. They are both changes in the nature or type of reward rather than its value. However, the identity prediction errors investigated in previous studies reflect changes in features

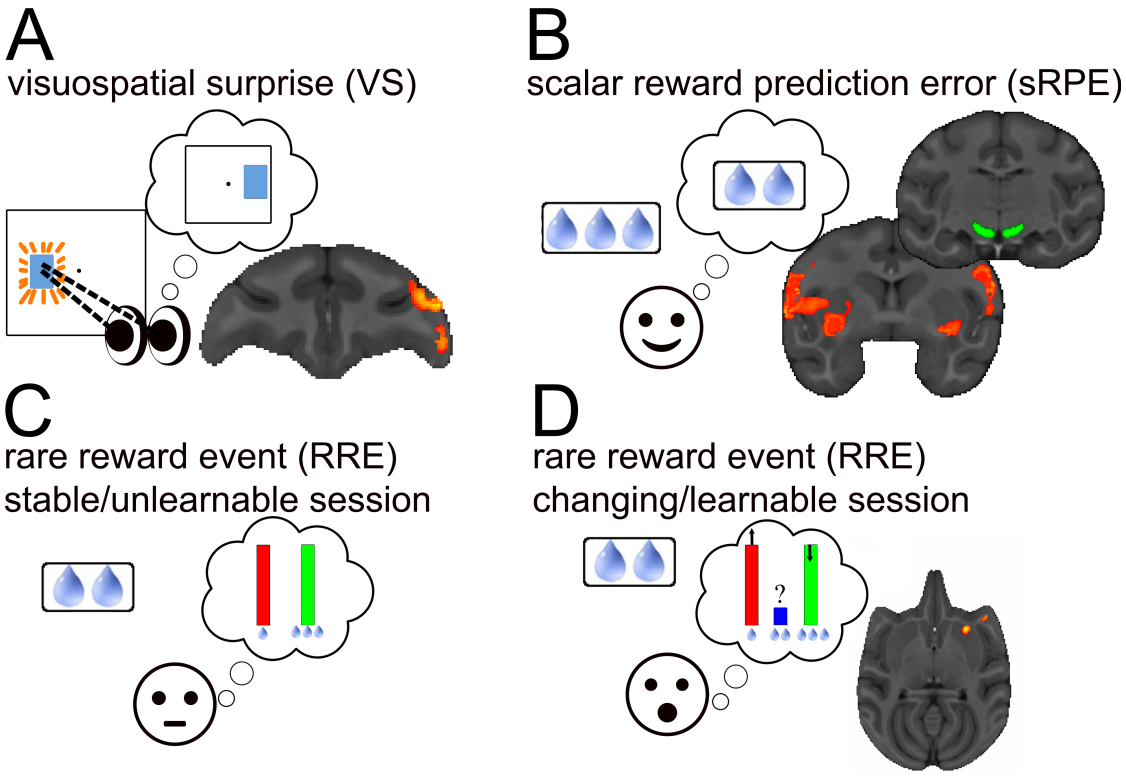


Figure 2.15: **Summary of the neural results.** (**A**) VS occurred when the stimulus is expected to occur on one side of the screen (thought bubble) but surprisingly appears on the other side of the screen. We found activity in the posterior lPFC in response to VS. (**B**) sRPEs occurred when the macaque experienced a reward level that was higher or lower than its reward expectation (thought bubble). We found activity in the dopaminergic midbrain and ventrolateral striatum in response to sRPE. (**C**) RRE occurred when an infrequent reward (2 drops) was sampled. We found no neural activity for RRE results in in stable/unlearnable sessions. This result might be because in such sessions less monitoring of reward occurrences takes place because occurrences of 1 or 3 drops cannot be predicted (thought bubble). (**D**) In contrast, we found activity for RRE in lateral striatum and plOFC in changing/learnable sessions. This might be because in such sessions the frequency of 1 and 2 drops of juice is actively tracked to estimate the current average reward rate, rendering obtaining 2 drops of juice a potential event of interest that is encoded as a new task relevant state (thought bubble).

of the reward outcomes that are intrinsic to the outcomes themselves. By contrast, the RREs that we study here are neither changes in reward value nor changes in a taste or odour identity feature intrinsic to the reward itself. Instead, RREs are events that are only surprising in the context of knowledge of the composition of the current reward environment. It is only if the monkeys represent this aspect of the task space they are navigating, that they will detect the occurrence of RREs.

The absence of a VS response in the current study is also interesting. In theory it is possible that if the dopaminergic midbrain responds to surprising sensory events such as surprising visual stimuli then this could underlie the dopaminergic response to identity prediction errors. It has been reported that the human dopaminergic midbrain is responsive to surprising visual identities in a study employing face and house stimuli [140]. On the other hand, it has been recently reported that reward identity prediction error responses in the dopaminergic midbrain are not proportional to the perceptual dissimilarity between the different rewards [137], suggesting that reward identity prediction errors are computed in some more abstract reward space.

There are three ways in which we might interpret the current failure to find VS responses in the dopaminergic midbrain. One possibility is that the absence of VS responses reflects the difficulty of imaging a small brain region such as the dopaminergic midbrain. From this perspective it is important to remember that absence of evidence for an effect is not tantamount to evidence of absence of the effect. Against such considerations, however, it would also be important to note that VS effects were found in the current study in several cortical regions that correspond with those previously seen to be responsive, in humans, to surprising visual stimulus identities and that our imaging techniques were sensitive enough to identify sRPEs in the dopaminergic midbrain [140]. A second possibility is that there is a fundamental difference between the visuospatial surprise we have studied here and the visual identity surprise studied in previous investigations [140]; maybe the dopaminergic midbrain is responsive to the latter but not the former. Finally, it is possible that human participants in the previous study obtained some pleasure or satisfaction in correctly predicting the visual stimuli they were about to see

even though, as in the current study, these were carefully decorrelated from reward events. It has certainly been argued that human participants are often motivated by a desire to perform a task well and are engaged by some of the task's design features as much or more than by the monetary rewards they will receive at the end of the experiment [141]. From this perspective, it would seem that when this level of task engagement is lacking, as in the monkeys in the current study, no dopaminergic response to VS is observable.

Dopaminergic neurons in the midbrain have been reported to response to salient stimulus events in addition to responding to their reward value and reward value prediction error associated with a stimulus [52]. This has sometimes been described as a response to the "physical impact" of the stimulus or to the stimulus' capacity for "behavioural activation" as opposed to its precise scalar reward value [53]. In theory the type of VS we investigated here might contribute to a stimulus' impact and capacity for behavioural activation. However, it also important to note that our experimental design and VS analysis controlled for other basic features of the visual stimuli we studied. VS-related activity was identified by comparing neural responses to visual stimuli of the same brightness, in the same locations, and with the same temporal predictability; only the predictability of the stimulus location differed between surprising and unsurprising VS events. Secondly, we also controlled for the "physical impact" of our visual stimuli in our GLMs by including a constant term at stimulus presentation but did not examine this effect further. As in the study conducted by Lak and colleagues [52], however, this analysis approach in the current study means that sRPE in the VTA/SN cannot be explained away as being the consequence of visual surprise.

Similarly, we found sRPE, but not RRE or VS, was associated with activity in the ventrolateral striatum, suggesting ventrolateral striatum also mostly coded for the current aggregate average value expectation. Thus, although animals may detect surprising reward events that are not sRPEs, they appear to employ neural mechanisms beyond the dopaminergic midbrain or ventral striatum to do so.

By contrast, rare reward events (RRE) were associated with a very different pattern of activity in more lateral parts of striatum and plOFC. The activity was situated lateral to the lateral orbital sulcus in or near area 47/12o and extended dorsally towards the ventral tip of the arcuate sulcus in or near areas ProM and 44. Interestingly, RREs were only tracked in sessions with changing average reward, i.e. when active monitoring and learning about the reward environment was more likely to occur [41, 127]. In this learning context, surprising events such as RREs might provide important information about a new state [142, 143]. On the other hand, in the stable/unlearnable session animals cannot learn to expect either high or low reward levels because one drop and three drop outcomes are equally likely. Just as neither one drop nor three drop outcomes provide important information about reward state and so will be disregarded as uninformative, so too will RRE outcomes also be disregarded.

It has been argued that OFC is particularly important for reward learning and credit assignment [79]. OFC has also been proposed to represent task structure and to locate one's current latent task state within that structure. However, whether and how OFC decides when to encode a new task relevant state has been unclear [79]. Our data suggests that plOFC is specifically sensitive to novel reward experiences, or new state creation, when an agent has to consider what state they are currently in.

Although there has, as yet, been little investigation of the neural basis of RREs, the presence of an RRE signal in plOFC is consistent with the fact that it is known that OFC does not just hold representations of reward size but rather representations that specify other features of the reward outcome [114–119]. The current results, therefore, suggest that while it is important for animals to detect when rewards are unexpected in terms of a one dimensional scale or common currency such as subjective value, equally animals track additional information about whether the nature and type of reward is surprising. Other authors looking for evidence of activity when surprising types of reward occur have also noted activity in orbitofrontal cortex [50, 137]. It is possible that evidence that the dopaminergic midbrain encodes a broader range of surprising events will emerge in future studies

employing other types of surprise, interactions between different types of surprise, other behavioural paradigms, or other recording techniques. Nevertheless, the current results suggest the possibility that such forms of surprise coding may be present, or even more prominent, beyond the dopaminergic midbrain in regions such as the plOFC and perhaps elsewhere in the frontal cortex.

The plOFC not only tracked RREs but it did so in a relatively selective manner. The same plOFC region did not respond to VS. In addition, there was no positive response to sRPEs. However, it is not the case that VS had no impact on brain activity. Within prefrontal cortex VS was most prominently associated with activity in posterior lPFC. This may be consistent with claims that lPFC is part of a network responding to unexpected state transitions even when they do not have immediate implications for reward [131, 144]. In addition, outside the frontal lobe mask in which we conducted our analysis, exploratory analysis also revealed VS effects in the intraparietal sulcus and inferior parietal lobule. A related pattern of activity in the parietal cortex has been reported in fMRI studies of visuospatial attentional shifting in macaques [145, 146].

In addition, lPFC exhibited a pattern of negative activity change in relation to RREs. This means that it is most active when either one or three drop rewards were received; in other words lPFC activity is decreased when RREs occur. Instead lPFC responds to any reward magnitude deviating from the average expectation regardless of whether it is larger or smaller than the average [147]. Thus even though lPFC may not detect RREs it appears to encode both surprising reward amounts and surprising visual events as might be expected from a domain general learning mechanism [128]

We note that we were able to demonstrate specific prediction error and surprise effects because each region that was identified as showing a significant effect of sRPE, RRE, or VS was then examined to see if it also held either of the two other types of activity patterns. Conducting a test in this way provides a strong demonstration of specificity if the secondary tests fail to find other prediction error/surprise effects even when not correcting for multiple comparisons. We refrained, however, from comparing activity patterns in different areas statistically in case it might be argued

that such comparisons are circular given that the areas usually had been first identified by their response in relation to a particular statistical contrast.

In summary, different types of surprising reward and visual events are associated with distinct effects on behaviour and distinct neural circuits. Even within the domain of reward learning there are distinct mechanisms linked with learning, on the one hand, about reward amount and its subjective value [136] and, on the other hand, mechanisms for learning about the precise features of potential rewards, such as their frequency, and for detecting rare reward events. While learning about reward value is associated with the dopaminergic midbrain and some divisions of the striatum, learning about other reward features and detecting rare reward events is associated with the OFC as well as with other divisions of the striatum. .

2.4 Materials and Methods

Six rhesus monkeys (*macaca mulatta*, one female) participated in the experiment. The animals weighed between 8.6 and 13.5 kg and were 7-8 years of age. They were kept on a 12-h light dark cycle, with *ad-lib* access to water for 12-16h after testing and throughout the day on non-testing days. All procedures were conducted under licences from the United Kingdom (UK) Home Office in accordance with the UK Animals (Scientific Procedures) Act 1986 and by the University of Oxford Animal Care and Ethical Review committee.

In the behavioural task, monkeys sat in the sphinx position in a purpose-built MRI-safe chair (Rogue Research, CA). In order to prevent head movements during fMRI data acquisition, an MRI-compatible cranial implant (Rogue Research, CA) was surgically implanted under anaesthesia.

Each trial began with a blank screen (2-4 seconds, mean 3 seconds) followed by a presentation of the rectangle. The monkeys responded by touching either of two custom-built infrared sensors placed in front of them. Each manual response was classified as either correct (the monkey touched the response sensor adjacent to the stimulus) or incorrect (the monkey touched the other sensor). Each correct response yielded a juice reward of one, two, or three drops ($\sim$1.5ml each drop)

after a delay of 200ms. The juice delivery took 1.5 seconds and after an inter-trial interval of 2-4 seconds (mean 3 seconds) the next trial began (Fig 2.1A). If the response was incorrect, the trial was repeated until the monkey made the correct response. Importantly, the spatial cue position (left or right) varied independently of reward magnitude (one, two, or three drops) that was given for each correct trial. Reward was thus decorrelated from spatial position of the stimulus.

The task design enabled us to examine VS because the side on which the stimulus was shown reversed after 11-19 trials on the same side (mean 15 trials) although occasional stimuli appeared on the opposite side throughout. 11-13 sessions of 150 rewarded (i.e. correct) trials were collected for each of six animals while they were in the MRI scanner. Thus, we could examine VS by comparing trials on which the stimulus had appeared on the same or the opposite side compared to the previous trial. Three types of sessions were performed by the monkeys:

Stable/unlearnable sessions: In six sessions, one and three drops were delivered randomly in 90% of the reward trials (45% for each reward size) and two drops were delivered in 10% of rewarded trials (Fig 2.1B, shown in blue). Mean reward over a session was kept at two drops. However, even if two drop rewards accorded with the average reward expectation, they were only rarely delivered and so they were in this sense the most surprising outcomes. Therefore, it was possible to identify activity related to RRE by comparing 2 drop reward outcomes with all other outcomes and it was possible to identify sRPE-related activity by calculating the parametrically varying scalar reward prediction error associated with each outcome. The sRPE and RRE regressors shared only 0.049% of variance. One weakness of the schedule, however, is that it is static and does not change over the course of the session. Because no learning is possible in such situations, learning mechanisms may not be deployed. We attempted to remedy this deficiency by using additional schedules.

Changing/learnable sessions: In two different reward schedules (each comprising two sessions), the mean reward changed either from an average of 1.5 drops to an average of 2.5 drops (‘changing up’ sessions; Fig 2.1B, yellow line) or in the

opposite direction, from 2.5 drops down to 1.5 drops ('changing down' sessions; Fig 1B, red line) halfway through the session. In these changing/learnable sessions, either one or three drops were delivered on 90% of trials on average, across the whole session. Therefore, once again it was possible to identify activity related to RRE by comparing 2 drop reward outcomes with all other outcomes and it was possible to identify sRPE-related activity by calculating the parametrically varying scalar reward prediction error associated with each outcome. However, it is possible that effects may be stronger in the changing/learnable sessions than the stable/unlearnable sessions because the reward environment is genuinely getting either better or worse. In order to estimate which is the case animals must pay attention to outcomes. In these sessions, the sRPE and RRE regressors shared 0.1551% of the variance.

Equiprobable sessions: In a third, control condition (comprising two sessions), we kept the average reward stable but each reward magnitude had an equal probability of 1/3, thereby eradicating any reward frequency effects (Fig 2.1B, shown in purple). Therefore, once again it was possible to identify activity related to sRPE-related activity by calculating the parametrically varying scalar reward prediction error associated with each outcome. Now, however, there may be less RRE effect because two drop reward outcomes are no less frequent, and therefore no more surprising, than one or three drop outcomes.

2.4.1 Behavioural analysis

We fitted generalised linear mixed effects models (GLMEs) to assess the impact of task manipulations on behaviour. As explained, we were interested in the effects of VS, sRPEs, and RRE on behaviour. Our binary regressor for VS encoded whether the stimulus appeared on the same side as on the previous trial or not. For this regressor, we zeroed out trials after the monkey had made an error, as these trials were repeated in our task design and would thus never lead to a visuospatial surprise. Our binary regressor coding for whether the current trial was a (surprising) two drop event (RRE) was also zeroed out for trials following an error. To calculate sRPEs, we first needed to compute the expected reward on each trial. To do so, we

first had to establish the degree to which outcomes from previous trials contributed to the reward expectation that animals would hold on the current trial. To do this we ran a GLME that included the rewards of the previous five trials as regressors. We then calculated the learning rate that best described the fitted beta weights of these five regressors, and used it to calculate the expected reward (sRE) on each trial. Other regressors we included as confounds in our GLMEs were the position of the target on the screen (left or right) to account for a potential bias toward one side, and the trial number to account for time-on-task effects.

We used this GLME to predict lapses of performance (whether any trial was an error or an outlier with a very long RT). Both errors and outliers were coded in the same way in this analysis. We first marked all trials with an RT over 4000ms (indicating disengagement from the task) or under 50ms (indicating impulsive responding also consistent with a failure to engage with the task) as outliers. Additionally, trials with an RT of more than 2.5 standard deviations away from the mean were also marked as outliers.

We used this GLME to predict lapses of performance (whether any trial was an error or an outlier with a very long RT). Both errors and outliers were coded in the same way in this analysis. We first marked all trials with an RT over 4000ms (indicating disengagement from the task) or under 50ms (indicating impulsive responding also consistent with a failure to engage with the task) as outliers. Additionally, trials with an RT of more than 2.5 standard deviations away from the mean were also marked as outliers.

GLME1: ErrorOrOutlier $\sim$ 1 + VS + sRE + RRE + Position + TrialNumber + (1 + VS + sRE + RRE+ Position + TrialNumber | Monkey) + (1 | Monkey:Session)

GLME2: ErrorOrOutlier $\sim$ 1 + VS + Rewardt-1 + Rewardt-2 + Rewardt-3 + Rewardt-4 + Rewardt-5 + RRE+ Position + TrialNumber + (1 + VS + Rewardt-1 + Rewardt-2 + Rewardt-3 + Rewardt-4 + Rewardt-5 + RRE+ Position + TrialNumber | Monkey) + (1 | Monkey:Session)

The learning rate to compute sRE in GLME1 was obtained from GLME2. This was done by minimizing the Euclidean norm between the weights of the previous five

rewards of a truncated Rescorla-Wagner reinforcement learning model (described by two free parameters: a learning rate and an inverse temperature) and the regression coefficients for Rewardt-1 to Rewardt-5 in GLME2. This procedure allowed us to both fit a learning rate while simultaneously accounting for other experimental factors in the linear model.

Additionally, we also ran each of these GLMEs separately for session types (i.e. stable/unlearnable, changing/learnable, and equiprobable), and for each animal. Moreover, as a further control, we ran the same GLMEs with RTs instead of performance lapses as the dependent variable, assuming a gamma distribution and using the log link function. For these RT GLMEs, we either excluded all outlier trials (**GLME3** and **GLME4**; see Fig 2.5AB), or all outlier, error, and repeat trials (trials after an error were repeated and thus did not result in a VS) (**GLME5** and **GLME6**; see Fig 2.5CD).

To test for the significance of individual regressors in our GLMEs, we fitted the models once with and once without the regressor in question and performed a likelihood ratio test between the two models.

2.4.2 MRI Data acquisition and preprocessing

Imaging data were collected using a 3T clinical MRI scanner and a four-channel phased-array receive coil in conjunction with a radial transmission coil (Windmiller Kolster Scientific, Fresno, CA). For each monkey, structural images were acquired under general anaesthesia, using a T1-weighted MP-RAGE sequence with a resolution of 0.5 x 0.5 x 0.5 mm, repetition time (TR) = 2.05 s, echo time (TE) = 4.04 ms, inversion pulse time (TI) = 1.1 s, and flip angle of 8°. Two or three structural images per subject were averaged. Anaesthesia was induced by intramuscular injection of 10 mg/kg ketamine, 0.125 - 0.25 mg/kg xylazine, and 0.1 mg/kg midazolam. fMRI data were collected during task performance with a gradient-echo T2* echo planar imaging (EPI) sequence with a resolution of 1.5 x 1.5 x 1.5 mm, interleaved slice acquisition, TR = 2.28 s, TE = 30 ms and flip angle of 90°. At the end of each session, to aid image reconstruction, a proton-density-weighted image was acquired

using a gradient-refocused-echo (GRE) sequence with a resolution of 1.5 x 1.5 x 1.5 mm, TR = 10 ms, echo time TE = 2.52 ms and flip angle 25°.

EPI data were prepared for analysis following a dedicated non-human primate fMRI processing pipeline [148] using tools from FMRIB Software Library (FSL) [149], Advanced Normalization Tools (ANTs) [150], and the Magnetic Resonance Comparative Anatomy toolbox (MrCat; https://github.com/neuroecology/MrCat). In short, after EPI data were reconstructed offline using a SENSE algorithm [151] time-varying spatial distortions were corrected using restricted non-linear registration, first to a session specific high-fidelity EPI, then to each animal's T1w structural image, and finally to a group-specific template in CARET macaque F99 space [152]. Functional images were temporally filtered (high-pass cutoff at 100s) and spatially smoothed (using a 3mm full width at half maximum Gaussian kernel).

2.4.3 FMRI analyses

Whole-brain analysis was conducted using a hierarchical General Linear Model (GLM) approach. Specifically, we first fitted every session individually before combining them for each monkey on a second hierarchical level using fixed effects in F99 standard space. Finally, we combined the data from all monkeys on a third hierarchical level using the FLAME 1+2 procedure from FSL [149] and using standard cluster-based thresholding criteria of $z > 2.3$ and $p < 0.05$ cluster-corrected [153]. Sixteen regressors of interest were designed for each session. Additional confound regressors were used to index head motion and volumes with excessive noise. All regressors were z-score normalised and the data were analysed using FSL's FEAT (FMRI Expert Analysis Tool). To model the hemodynamic response function, each regressor was convolved with a single gamma function (mean lag = 4.5s, standard deviation = 2s, therefore peaking 3.5s after the event, which is consistent with a faster hemodynamic response function (HRF) in macaques than humans). The analysis and cluster correction were run only in a predefined volume of interest (VOI) covering the frontal cortex and striatum, as shown in Fig 2.5.

We were interested in the main effects of VS, sRPE, and RRE. We included two regressors, coding for VS occurring when stimuli were presented on the left and right side of the monitor to account for nonlinear differences between the two sides. These regressors were then combined on the contrast level to allow identification of neural activity related to VS independent of precise location of the surprising event. Finally, we note that, as in the behavioural analysis, we zeroed out the VS regressor whenever an error occurred on the previous trial. This is because after an error the trial is repeated until the monkey answers correctly, making it not spatially surprising any longer.

To examine the neural correlates of sRPE, we included six regressors at the outcome phase of each trial: one regressor encoding the magnitude of the current reward (zero, one, two, or three drops), and five regressors encoding the reward magnitudes on the five previous trials. These six regressors were combined on the contrast level by subtracting the previous five rewards (which determine the animal's expectation about reward on the current trial) from the current reward. The regressors encoding the previous five outcomes were weighted according to the learning rate we fitted to our behavioural GLME (thus the reward expectation on the current trial is more influenced by recent previous rewards and less influenced by more distant past rewards). Additionally, because we observed some activity in the orofacial sensorimotor and gustatory cortex that could reflect the cortical activity correlates of swallowing and tasting the juice from the most recent outcome, we also contrasted the current reward outcome against a reward expectation estimate based on the weighted outcomes of four previous trials but leaving out the most recent trial. RRE events (receiving two drops of juice) were encoded with separate regressors both at outcome (whether the current reward is two drops) and at decision (whether the last reward was two drops). Finally, we also included two constants, one at decision and one at outcome, as well as a regressor that controlled for hand movements registered by the infra-red touch sensor into the GLM. The decision constant accounts for the neural impact of visual stimulus presentation or motor activation [52, 53]. The outcome constant accounts for the neural response to

receiving reward, regardless of magnitude [52, 53]. Note that by including terms relating to both sRPE and RRE in the same GLMs we can ensure that activity actually related to sRPE cannot be misinterpreted as activity related to a RRE.

After having identified clusters of activity for our effects of interest, we were interested in whether the same or different regions process VS, sRPE, and RRE. To this end we transformed the locations of the peaks of activity of our whole brain analysis into session space, following the non-linear deformation field. There, we extracted the average z-statistics for spheres with a diameter of 7.5mm centred at the warped peaks for all regressors of interest. For the ROI covering SN/VTA, we used a mask from a recently published atlas [154], which we warped into session space. When effects were illustrated with extracted BOLD timecourses, these timecourses were extracted from the same ROIs and were upsampled by a factor of 10 using spline interpolation. To test for statistical significance, we average the timecourses of each session over time (from 0s to 10s in Fig 2.7 and from 0s to 15s in Fig 2.2) and compare the effect over sessions and monkeys against baseline while controlling for subject-by-subject differences by modelling monkeys as random effects.

3

Strategic exploration in the macaque's prefrontal cortex

Contents

Abstract

Humans have been shown to strategically explore. They can identify situations in which gathering information about distant and uncertain options is beneficial for the future. Because primates rely on scare resources when they forage, they are also thought to strategically explore, but whether they use the same strategies as humans and the neural bases of strategic exploration in macaques are largely unknown. We designed a sequential choice task to investigate whether macaques mobilise strategic exploration based on whether that information can improve subsequent choice, but also to ask the novel question about whether macaques adjust their exploratory choices based on the contingency between choice and information, by sometimes providing the counterfactual feedback, about the unchosen option. We show that macaques decreased their reliance on expected value when exploration could be beneficial, but this was not mediated by changes in the effect of uncertainty on choices. We found strategic exploratory signals in anterior and mid-cingulate cortex (ACC/MCC) and dorsolateral prefrontal cortex (dlPFC). This network was most active when a low value option was chosen which suggests a role in counteracting expected value signals, when exploration away from value should to be considered. Such strategic exploration was abolished when the counterfactual feedback was available. Learning from counterfactual outcome was associated with the recruitment of a different circuit centred on the medial orbitofrontal cortex (OFC), where we showed that macaques represent chosen and unchosen reward prediction errors. Overall, our study shows how ACC/MCC-dlPFC and OFC circuits together could support exploitation of available information to the fullest and drive behaviour towards finding more information through exploration when it is beneficial.

3.1 Introduction

In the general theoretical framework of optimal foraging [14], foraging is an optimisation problem that can be solved by a cost-benefit analysis. In many species, foraging can be accounted for by simple behaviours – approach/avoidance of an observed and immediately available source of food – that require no mental representations. In those models, exploration is often defined as a random process, where noise in behaviour can lead animals to change behaviour by chance [22, 25, 155, 156]. However, in species relying upon spatially and temporally scattered resources, such as fruits, the computation of costs and benefits of foraging should extend in space and time [157]. Strategic exploration implies a specific representation of potential future action and outcomes, which allows animals to select specific options that would be better over longer time or spatial scale, rather than using a general heuristic such as win/stay lose shift. It implies foregoing immediate rewards to collect information about delayed, distant or risky rewards. But it is essential to maximise rewards over a longer time and spatial scale. For frugivorous animals such as primates, it might be critical for survival.

Humans have been shown to strategically explore [27, 34, 158, 159], but there is little evidence in other species. Inspired by Wilson and colleagues' "horizon" exploration task [27], we developed a task to investigate whether macaques mobilise strategic exploration based on whether that information can improve subsequent choice. Importantly non-human primate models provide insights into the evolutionary history of cognitive abilities, and of the neuronal architecture supporting them [160]. Given the rhesus macaques' ecology (including feeding), they should also be able to use strategic exploration, but the extent to which they can mobilize strategic exploration might be different from that of humans. Based on the similarities in circuits supporting cognitive control and decision-making processes in humans and macaques [161, 162], one could further hypothesise that the same neurocognitive processes (the same computational model) might be recruited but not to the same extent (different weights).

As in Wilson and colleagues' original study, we manipulated whether the information could be used for future choices by changing the choice horizon [27]. On *short horizon trials*, the information provided by the outcome of the choice could only be used for the current choice and was then worthless going forward. On *long horizon trials*, it could be used to guide a sequence of four choices. By comparing exploration in both conditions, we could test whether the animals reduced their reliance on value estimates (*random exploration*) and increased their preference for more uncertain options (*directed exploration*) when gathering information was useful for future choices in the long horizon [27]. In addition, we manipulated the contingency between the choice and the information by varying the type of feedback that macaques received. In one experimental condition, the animals could only receive information about an option by choosing it (*partial feedback* condition). In the other experimental condition, there was no contingency between the choice and the information as they received information about the outcome of both the option they chose and the alternative option (*complete feedback condition*). In the latter case, the information about the options could be learned from the counterfactual outcomes – the outcome that would have been obtained had a different choice been made. This type of feedback is sometimes referred to as "hypothetical" [163] or "fictive" feedback [164]. With this complete feedback condition, we could probe whether macaques decreased their exploration and relied more on value estimates when information was freely available. A strategic explorer would only actively explore – and forgo immediate rewards – when it is useful for the future (long horizon) and that it is the only way to obtain information (partial feedback). In addition to behavioural data, neural data were collected using fMRI to probe the neural substrates of strategic exploration. Our analysis was focused on regions previously identified in fMRI studies on reward valuation and cognitive control in macaques [58, 165–169]. Finally, we took advantage of the different feedback conditions to explore how macaques update their expectations based on new information. Specifically, we investigated the behavioural and neural consequences of feedback

about the outcome of their choice and – in the complete feedback condition – on counterfactual feedback from the alternative.

We found that rhesus macaques engaged in strategic exploration by decreasing their reliance on expected values when it was useful for the future (long horizon) and that active sampling was the only way to obtain information (partial feedback). In other words, macaques strategically adjusted the degree to which they use random exploration depending on both horizon length and feedback type. Neurally, we found prefrontal strategic exploration signals in the ACC/MCC and dlPFC. However, we did not find a significant effect of uncertainty (directed exploration) on choices.

When making choices in a sequence (long horizon), we found evidence that macaques used counterfactual feedback to guide their choices. Complementing this activity at the time of decision, we found overlapping chosen and unchosen outcome prediction error signals in the OFC, at the time of receiving the outcome. The counterfactual prediction errors in the OFC are particularly interesting as they point to the neural system that allowed the macaques to forgo having to make exploratory choices in the complete condition, which could also change how the MCC-dlPFC network represented the value of the chosen option.

3.2 Results

3.2.1 Probing strategic exploration in macaques

Three macaques performed a sequential choice task inspired by Wilson and colleagues [27]. In this paradigm called the horizon task, macaques were presented with one choice (short horizon) or a sequence of four choices (long horizon) between two options (Fig 3.1A). Each option belonged to one side of the screen and had a corresponding touch pad located under the screen (see Material and Methods for details). Both types of choice sequence (long and short horizon) started with an 'observation phase' during which animals saw four pieces of information randomly drawn from both options and reflecting outcome distribution of each option. They received at least one piece of information per option (Fig 3.1B). Each piece of information was presented exactly like subsequent choice outcomes as a bar length

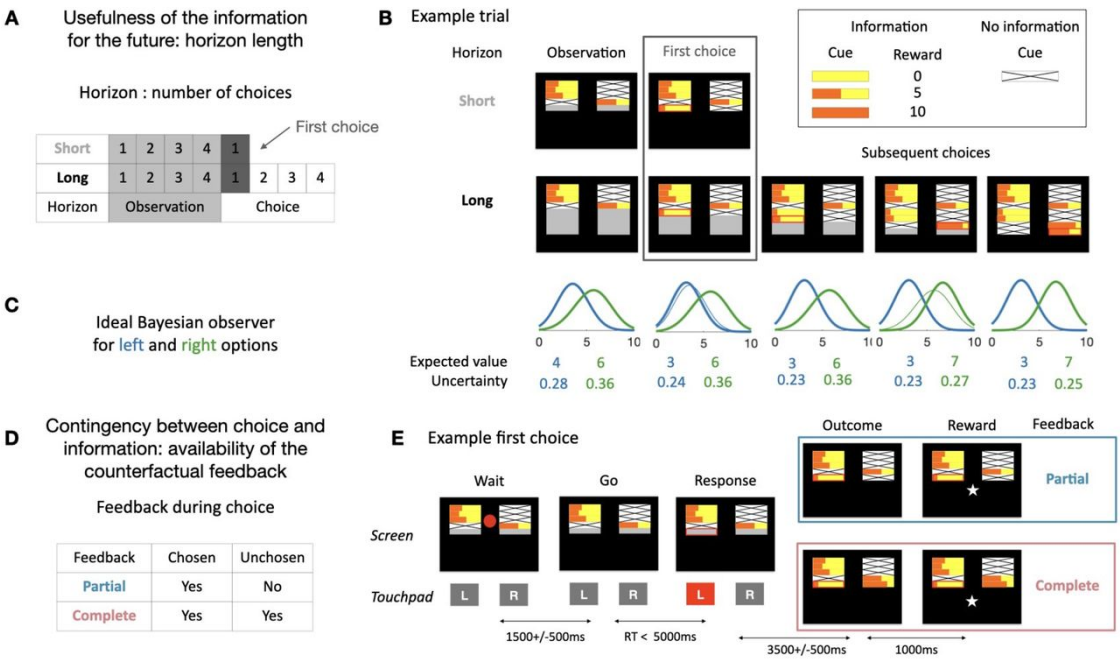


Figure 3.1: **Task and model.** (Continued on the following page)

(equivalent of 0 to 10 drops of juice) drawn from each option's outcome distribution. The animals had been trained that the length of the orange bar on a yellow background indicated the number of drops of juice associated with that specific option on a given trial (Fig 3.1B). One option was associated with a larger reward (more drops of juice) on average than the other. The means of the distributions were fixed within a sequence but unknown to the macaque. The animals only received the reward associated with the option they chose at the end of each choice.

First, we manipulated whether the information gathered during the first choice could be useful in the future. During a session, we varied the number of times the animals could choose between the options (horizon length). The horizon length was visually cued (Fig 3.1AB). Second, we manipulated the contingency between choice and information by varying the type of feedback macaques received after their choices. In the partial feedback condition, they only saw the outcome for the option they chose. In the complete feedback condition, they saw the outcome of both the option the chose and the alternative option (Fig 3.1DE). The feedback condition was not cued but was fixed during a session.

Figure 3.1: **Task and model.** (**A**) During the task, we manipulated whether the information could be used in the future by including both long and short horizon sequences. In both trial types the animals initially received four samples ('observations') from the unknown underlying reward distributions. In short horizon trials they then made a one-off decision between the two options presented on screen ('choice'). In long horizon trials they could make four consecutive choices between the two options (fixed reward distributions). On the first choice (highlighted) the information content was equivalent between short and long horizon trials (same number of observations), whereas the information context was different (learning and updating is only beneficial in the long horizon trials). (**B**) Example short and long horizon trials. The macaques first received some information about the reward distributions associated with choosing the left and right option. The length of the orange bar indicates the number of drops of juice they could have received (0-10 drops). The horizon length of the trial is indicated by the size of the grey area below the four initial samples. The macaques then make one (short horizon) or four (long horizon) subsequent choices. As the animals progressed through the four choices, more information about the distributions was revealed. Displayed here is a partial information trial where only information about the chosen option is revealed. (**C**) Ideal model observer for the options of the example trial shown in B (color code corresponds to the side of the option). The distributions correspond to the probabilities to observe the next outcome for each option. The expected value corresponds to the peak of the distribution and the uncertainty to the variance. Thick lines correspond to post outcome estimate and thin lines to pre-outcome estimates (from the previous trial). (**D**) We also modulated the contingency between choice and information by including different feedback conditions. In the partial feedback condition the animals only receive feedback for the chosen option. In contrast, in the complete feedback condition they receive feedback about both options. (**E**) Example partial and complete feedback trials (both short horizon). Here, the observation phase shown in (B) is broken up into the components the macaques see on screen during the experiment. Initially, the samples were displayed on screen but a red circle in the centre indicates that the animals could not yet respond. After a delay, the circle disappears, and the macaques could choose an option. After they responded, the chosen side was highlighted (red outline). After another delay, the outcome was revealed. In the partial feedback condition (top) only the outcome for the chosen option was revealed. In contrast, in the complete feedback condition (bottom) both outcomes were revealed. After another delay the reward for the chosen option was delivered in both conditions.

To assess macaques' sensitivity to the expected value and the uncertainty about the options, we set up an ideal observer Bayesian model (see Material and Methods for model details), which estimates the probability of observing the next outcome given the current information (Fig 3.1C). This model uses only the visual information available on the screen to infer the true underlying mean value of each options but does not use the horizon nor the feedback type as those were irrelevant for this inference. We extracted the expected value (peak of the probability distribution of the next observation i.e., most likely next outcome) and the uncertainty (variance) of the options from the model to evaluate the animals' sensitivity to these variables. If macaques did not engage in strategic exploration, the effect of expected value should be unaffected by the manipulations of horizon and feedback as was the case for the model.

3.2.2 The horizon length and the type of feedback modulate macaques' exploration

We first focused our analysis on the first choice of the trial, as the information about the reward probability of two options was identical across horizons and feedback conditions, such that choices should only be affected by the contextual manipulations (horizon and feedback type). If macaques were sensitive to whether the information could be used in the future, they would explore more in the long compared to the short horizon. This is because information obtained early in a trial can only beneficial for subsequent choices in long horizon trials. Moreover, exploration should only occur when obtaining information is instrumentally dependent upon it, i.e., in the partial feedback condition (Fig 3.2A).

We first ensured that the animals' choices were influenced by the expected value computed by the Bayesian model. We looked at the accuracy (defined as choosing the option with the highest expected value according to the model) during the first choice. For the two horizon lengths and in both feedback conditions, accuracy was above chance level (t-test compared to a distribution with a mean at 0.5; partial feedback short horizon: $t(40) = 10.029, p < 0.001$, partial feedback long horizon:

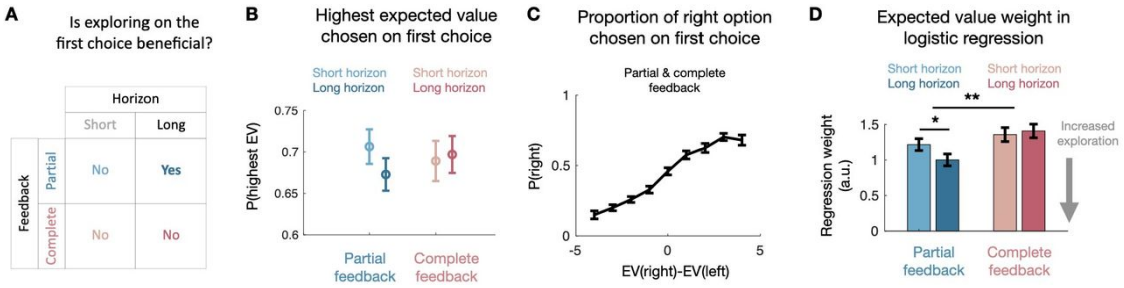


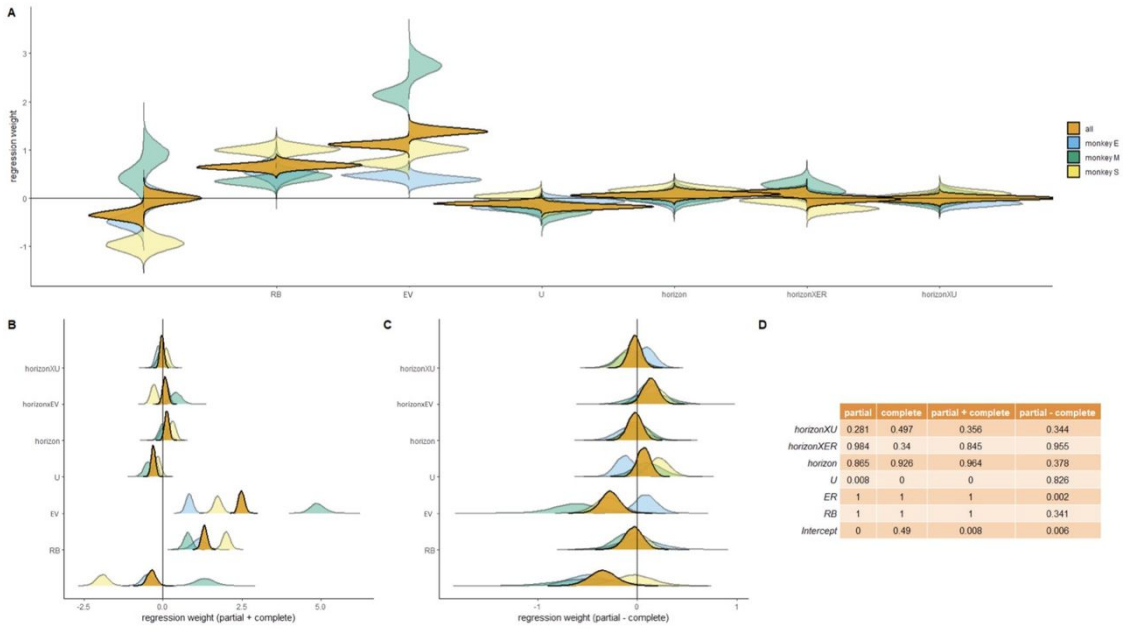
Figure 3.2: **First choice.** (**A**) In our experimental design, on the first choice of a horizon, directed exploration is only sensible in long horizon trials in the partial feedback condition. This is because in short horizon trials the information gained by exploring is of no use for subsequent choices, so a rational decision-maker would only choose based on the expected value of the options. Moreover, in the complete feedback condition all information is obtained regardless of which option is chosen, so an ideal observer would again always choose the option with the highest expected value. (**B**) The proportion of trials in which the macaques chose the option with the higher expected value is above chance level (0.5) across both feedback conditions and horizons. Mean across sessions (partial feedback: 41 sessions, complete feedback: 40 sessions). (**C**) The animals' choices are sensitive to nuanced differences in expected value. Mean across all sessions (81 sessions). (**D**) According to the logistic regression model predicting macaques' first choices in a horizon (see main text and methods for details), macaques' first choices are less driven by expected value in the partial than in the complete feedback condition. Within the partial feedback condition, they are less driven by expected value in long then in short horizon trials. No such difference was found in the complete feedback condition. This is evidence that macaques deliberately modulate their exploration behavior to explore more on partial feedback long horizon trials, where exploration is sensible (see (A)). Error bars indicate standard error to the mean in B-C and standard deviation in D.

$t(40) = 8.930, p < 0.001$, complete feedback short horizon: $t(39) = 7.906, p < 0.001$, complete feedback long horizon: $t(39) = 8.9634, p < 0.001$, Fig 3.2B). Therefore, macaques used the information provided by the informative observations on each trial to guide their choices. However, using this raw behavioural measure did not uncover the effects of feedback type (2-way ANOVA with interaction on the raw accuracy: $F(1, 158) = 0.02, p = 0.87$), horizon length ($F(1, 158) = 0.36, p = 0.55$) nor their interaction ($F(1, 158) = 0.93, p = 0.34$) as it does not isolate the effect of expected value and uncertainty on choices. The animals also adjusted their choices to variations in expected value, as can be seen when pooling together both feedback conditions and horizon lengths (Fig 3.2C, see statistical significance in Fig 3.2D).

Although choices were guided by the expected value of the options above chance-

level, the animals still sometimes chose the less valuable option in both conditions and horizons (Fig 3.2BC). We examined whether macaques were less driven by expected value on partial feedback long horizon trials, as exploration is only a relevant strategy on these trials (Fig 3.2A). To test this hypothesis, we ran a logistic regression predicting responses during first choices in the partial and complete conditions. As regressors, we added the expected value according to our Bayesian model, the uncertainty according to our Bayesian model, the horizon (short/long), and the interactions of expected value and uncertainty with horizon. Moreover, we added two potential biases, a side bias and tendency to repeat the same action. We fitted regressors to vary by condition (partial or complete feedback) and by animal, and modelled sessions as random effects for each macaque, with all regressors included as random slopes. We confirmed that in both feedback conditions macaques tended to choose the option with the highest expected value ($p < 0.001$ in the partial condition and $p < 0.001$ in the complete; one-sided test, based on sample drawn from Bayesian posterior, see Material and Methods). We identified that macaques relied more on the difference in expected value in the complete than in the partial feedback condition ($p = 0.0024$; one-sided test), and in short horizon than in the long horizon in the partial condition only ($p = 0.0163$ in the partial condition and $p = 0.6598$ in the complete; one-sided test). Thus, animals engaged in strategic exploration by reducing their reliance on expected value. In other words, animals strategically modulated the degree to which they use random exploration both depending on the horizon length and feedback type.

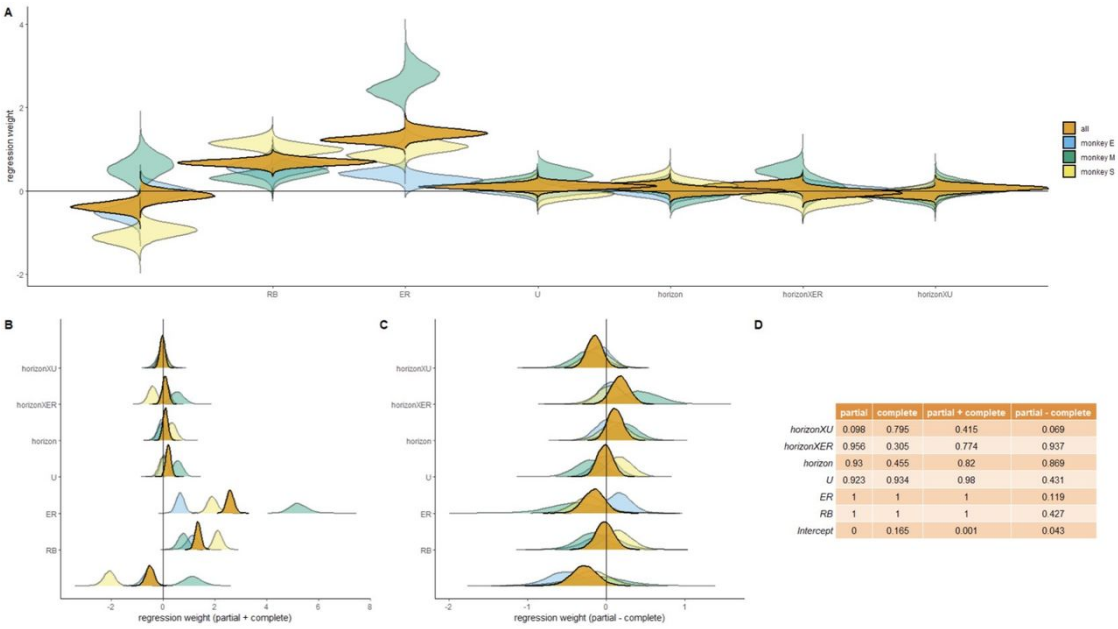
We next looked at the effect of uncertainty. Exploratory behaviours should be sensitive to how much they can reduce uncertain i.e., the animals should optimally pick the most uncertain option when they explore [27]. We found that macaques were sensitive to uncertainty overall, avoiding options that were more uncertain in the partial and the complete feedback conditions ($p = 0.0081$ in the partial condition and $p = 0.00025$ in the complete; one-sided test) (see Fig 3.3 for full model fit). This risk aversion was driven by the difference in number of information presented as when we restricted our analysis to the trials where they received 2



	partial	complete	partial + complete	partial - complete
horizonXU	0.281	0.497	0.356	0.344
horizonXER	0.984	0.34	0.845	0.955
horizon	0.865	0.926	0.964	0.378
U	0.008	0	0	0.826
ER	1	1	1	0.002
RB	1	1	1	0.341
Intercept	0	0.49	0.008	0.006

Figure 3.3: **Model fit predicting choosing the right option on screen on first choices** (shown in Fig 3.2D and described in detail in the Methods section). (**A**) Predictors are from left to right: Intercept (i.e., a side bias), repetition bias (RB), expected value of difference between right and left according to our Bayesian model (EV), uncertainty difference between right and left according to our Bayesian model (U), horizon length (short horizon is positive, long horizon is negative), the interaction between horizon and expected value (horizonXER), and the interaction between horizon and uncertainty (horizonXU). The distributions are the posteriors of the parameter estimates, shown both for each animal individually and averaged over animals. Fits from the partial feedback sessions are shown on the left, and from the complete feedback sessions on the right. (**B**) Data from the same fit as in (A) but now summed up over both partial and complete feedback sessions. (**C**) Data from the same fit as in (A) but now we computed the difference between partial and complete feedback sessions. (**D**) One-sided p-values for all parameters are computed as the number of samples of the posterior greater than 0. To compute the p-value for effects smaller than 0, the p-values in the table can be subtracted from 1.

information about each option, macaques showed a small preference for the more uncertain option ($p = 0.077$ in the partial condition and $p = 0.066$ in the complete; $p = 0.02$ when combined; one-sided test) (see Fig 3.4 for full model fit). However, we found no statistically reliable difference in the sensitivity to the uncertainty across the experimental conditions. Therefore, uncertainty did not play a key role in strategic exploration in our task. This indicates that our macaques did not use directed exploration to strategically guide their choices.



	partial	complete	partial + complete	partial - complete
horizonXU	0.098	0.795	0.415	0.069
horizonXER	0.956	0.305	0.774	0.937
horizon	0.93	0.455	0.82	0.869
U	0.923	0.934	0.98	0.431
ER	1	1	1	0.119
RB	1	1	1	0.427
Intercept	0	0.165	0.001	0.043

Figure 3.4: **Full model fit predicting choosing the right option on screen on first choices with equal information.** The same model as in Fig 3.3 but only fit to trials during which the available choices on screen were the same on each side (2 and 2). All conventions are the same as in Fig 3.3.

3.2.3 Macaques learn from chosen and counterfactual feedbacks

We next assessed whether macaques used the information they collected during their previous choices to update their choice, and how the nature of the feedback affected this process. To this end, we focused our analysis on choices from long horizon trials. On such trials, macaques' accuracy (defined as choosing the option with the highest expected value according to the model) was always above chance level (t-test compared to a distribution with a mean at 0.5; all $p < 0.001$) and increased as they progressed through the sequence (t-test compared to a distribution with a mean at 0 of the distribution regression coefficients of the trial number onto the accuracy (both z-scored) for each session; partial feedback condition: $t(40) = 11.3653, p < 0.001$, complete feedback condition: $t(39) = 5.6590, p < 0.001$) (Fig 3.5A). We inferred that this improvement was due to the use of the information collected during the choices. To examine this, we isolated the change in expected value compared to the initial 'observation phase' (see Material and Methods). We found that macaques

were sensitive to the change in expected value both for the chosen option (in the partial and complete feedback conditions) and the unchosen option (counterfactual feedback in the complete feedback condition only) (Fig 3.5B-C, see statistical significance in Fig 3.5E). Macaque displayed a significant tendency to choose the same option (t-test compared to a distribution with a mean at 0.5; all $p < 0.001$), which sharply increased after the first trial (paired t-test between the first choice and the subsequent choices; all $p < 0.001$) and kept increasing after the first choice (t-test compared to a distribution with a mean at 0 of the distribution regression coefficients of the trial number onto the probability to choose the same option (both z-scored) for each session; partial feedback condition: $t(40) = 5.3026, p < 0.001$, complete feedback condition: $t(39) = 3.1265, p = 0.003$) (Fig 3.5D).

We investigated the determinants of these effects by performing a logistic regression for all non-first choices. We added regressors for the expected value and uncertainty during the observation phase (which served as a baseline for subsequent choices), and regressors for the change in these baselines as new information was revealed as they progressed through the horizon. We also added three potential biases in choices: a side bias, the tendency to repeat the same action, and a bias for choosing the option most often chosen (see Fig 3.6 for full model fit). Just as with the previous regression model for first choices, we again allowed regressors to vary by condition and animal and modelled sessions as random effects. We confirmed that macaques remained sensitive to the difference in expected value during the observation phase, and that guided the first choice ($p < 0.001$ in the partial condition and $p < 0.001$ in the complete; one-sided test). Consistent with the choice behaviour on the first choice, macaques relied more on this difference in the complete than in the partial feedback condition in subsequent choices ($p = 0.0192$, one-sided test; Fig 3.5E). The animals were biased towards repeating the same choice ($p < 0.001$ in the partial condition and $p < 0.001$ in the complete; one-sided test), but this bias was also more pronounced in the partial feedback condition ($p = 0.002$, one-sided test; Fig 3.5E) as can already be seen in Fig 3.5B. The animals also preferred to choose the option most chosen ($p < 0.001$ in the partial condition

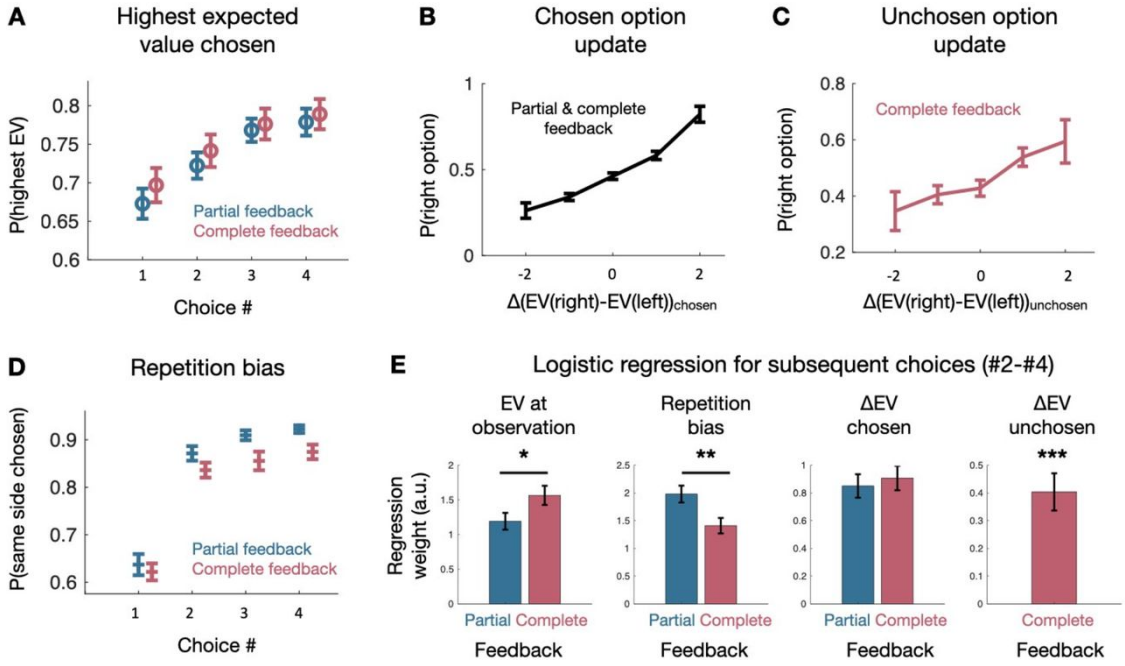


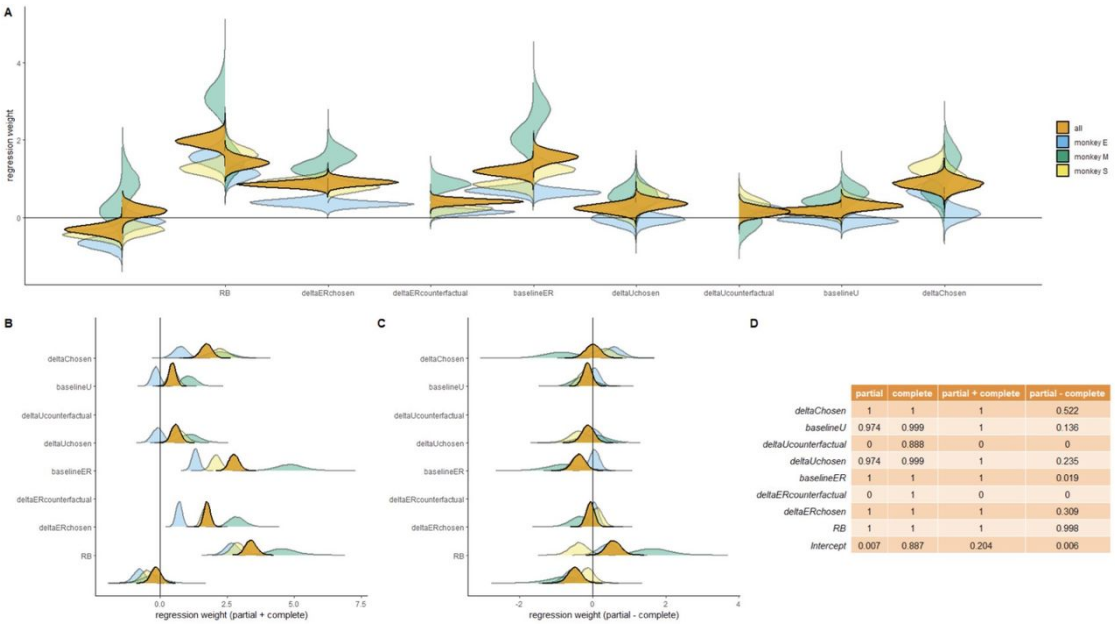
Figure 3.5: **Behavioural update.** (**A**) As macaques progressed through the long horizon, they were more likely to choose the option with the higher expect reward in both the partial and complete feedback condition. Mean across sessions (partial feedback: 41 sessions, complete feedback: 40 sessions). (**B**) The animals were sensitive to changes in the expected value compared to the baseline expected value they experienced during the observation phase both for the chosen option (mean across all sessions (81 sessions)) and (**C**) the unchosen option (mean across all complete feedback sessions (40 sessions)). (**D**) The animals were also more likely to repeat their choice as they progressed through the long horizon. Mean across sessions (partial feedback: 41 sessions, complete feedback: 40 sessions). (**E**) Results of the single logistic regression model predicting 2, 3, and 4th choices in the long horizon. In both the partial and complete feedback macaques were sensitive to the expected value at observation but more so in the complete than the partial feedback condition (left). The animals tended to repeat previous choices in both conditions but more so in the partial than in the complete feedback condition (centre left). In both conditions, macaques were sensitive to the change in expected value compared to the observation phase with no significant difference between conditions (centre right). In the complete feedback condition macaques were also sensitive to the change compared to baseline of the additional information they received. Error bars represent standard error of the mean in A-D and standard deviation in E.

and $p < 0.001$ in the complete; one-sided test), which explained the increase in repetition bias over time, but this was not affected by the feedback type (partial > complete: $p = 0.309$) (Fig 3.6). The animals were sensitive to the change in expected value when the information was related to the chosen option ($p < 0.001$ in the partial condition and $p < 0.001$ in the complete; one-sided test), with no statistical difference between the partial and complete feedback conditions (partial > complete: $p = 0.6913$). Finally, in the complete feedback condition, macaques were sensitive to the change in expected values obtained from the counterfactual feedback ($p < 0.001$; one-sided test; Fig 3.5E).

Overall, we found that on top of being more sensitive to the expected value difference during the initial evaluation, macaques were less likely to be biased towards repeating the same action when they had counterfactual feedback to further guide their choices in the complete feedback condition. They were able to learn about the options, using both the chosen and the counterfactual feedback when it was available.

3.2.4 Strategic exploration signals in ACC/MCC and dlPFC

To identify brain areas associated with strategic exploration, we ran a two-level multiple regression analysis using a general linear model (GLM). For each individual session, we used a fixed-effects model. To combine sessions and animals, we used random effects as implemented in the FMRIB's Local Analysis of Mixed Effects (FLAME) 1 + 2 procedure from the FMRIB Software Library (FSL). We focused our analysis on regions previously identified in fMRI studies on reward valuation and cognitive control in macaques [58, 165–169]. Thus, to only look at the regions we were interested in and to increase the statistical power of our analysis, we only analysed data in a volume of interest (VOI) covering frontal cortex and striatum (previously used by [169]; see Fig 2.6). We used data from 75 (41 partial feedback and 34 complete feedback) of the 81 (41 partial feedback and 40 complete feedback) sessions we had acquired (fMRI data from 6 sessions were corrupted and unrecoverable). Details of all regressors included in the model can be found in the Methods section. In addition to the analysis in the VOI, we examined the



	partial	complete	partial + complete	partial - complete
deltaChosen	1	1	1	0.522
baselineU	0.974	0.999	1	0.136
deltaUcounterfactual	0	0.888	0	0
deltaUchosen	0.974	0.999	1	0.235
baselineER	1	1	1	0.019
deltaERcounterfactual	0	1	0	0
deltaERchosen	1	1	1	0.309
RB	1	1	1	0.998
Intercept	0.007	0.887	0.204	0.006

Figure 3.6: **Model fit predicting choosing the right option during subsequent choices in the long horizon** (choices 2-4; shown in Fig 3.5E and described in detail in the Methods section). (**A**) Predictors are from left to right: Intercept (i.e. a side bias), repetition bias (RB), the change in expected value between the right and left option revealed by choices made during this horizon, compared to the initial expected value for this horizon, i.e. the baseline (deltaERchosen), the change in expected value between the right and left option revealed by feedback about the unchosen option, compared to the initial expected value for this horizon (deltaERcounterfactual), the difference in initial expected value between the right and left option available, i.e. the expected value difference at first choice (baselineU), the change in uncertainty between the right and left option revealed by choices made during this horizon, compared to the initial uncertainty for this horizon (deltaUchosen), the change in uncertainty between the right and left option revealed by feedback about the unchosen option, compared to the initial uncertainty for this horizon (deltaUcounterfactual), the difference in initial uncertainty between the right and left option available, i.e. the uncertainty difference at first choice (baselineU), the difference between how often the right option has been chosen over the left option during this horizon (deltaChosen). All other conventions are the same as in Fig 3.3, also for panels **B**-**D**.

activity in the functionally and anatomically defined regions of interest (ROIs). These ROIs were not chosen a priori but were selected based on the activity in the in the VOI. The goal of these analyses was either: i) to examine the effect of a different variable than the one used to define the ROI in our VOI, which is an independent test so we could look for statistical significance of this different variable on the activity in the ROI, ii) to illustrate an effect revealed in the VOI, which is not an independent test, so we did not do any statistical analysis.

To examine how macaques use initial information displayed during the observation phase of the task differently depending on the horizon and the feedback condition, we examined the brain activity when the stimuli were presented on the first choice ('wait' period; Fig 3.1D). Crucially, there was no difference in the visual inputs between the partial and the complete feedback condition, as the nature of the feedback was not cued and fixed for blocks of sessions. We first investigated the main effects of our two manipulations: the overall effect of the horizon and feedback type on brain activity.

We combined all sessions and looked for evidence of different activations in the long and short horizon. We found a significantly greater activity for the long horizon in 3 clusters (cluster $p < 0.05$, cluster forming threshold of $z > 2.3$; Fig 3.7A, see Table B.1 for coordinates of cluster peaks). One cluster was centred on the pregenual anterior cingulate cortex (pgACC) and the striatum and two clusters of activities were centred on the dlPFC and extended in the lateral orbitofrontal cortex (lOFC, area 47/12o; see methods for more details about OFC subdivisions) with one on each hemisphere. In an independent test, we placed ROIs by calculating the functional and anatomical overlap for each Brodmann area 24, 46 and 47/12o and extracted the t-statistics of the regressor to examine the effect the contingency between choice and information (feedback condition). We observed no effect of the feedback type in ACC ($p = 0.19$) and lOFC ($p = 0.53$), but we found a main effect of feedback type in the dlPFC (2-way ANOVA, $F(144, 147) = 4.86, p = 0.029$) and no interaction anywhere (ACC: $p = 0.29$, dlPFC: $p = 0.9$ and lOFC: $p = 0.78$). This revealed that a subpart of the pgACC and the lOFC were sensitive to the

horizon length, while the dlPFC showed an additive sensitivity to the horizon length and the feedback type, such that it was most activated in the long horizon and partial feedback, when exploration is beneficial.

We next examined the effect of the feedback in our VOI. We found one cluster around the MCC that was significantly modulated by the difference between the activity during the complete and partial feedback conditions during stimuli presentation on the first choice (Fig 3.7C, yellow contrast; see Table B.1 for coordinates of cluster peaks). To examine this effect further, and although it is not an unbiased test, we defined an ROI by taking the overlap between our functionally defined cluster and Brodmann area 24'. Extracted the t-statistics of each session from the regressor from this ROI revealed that the MCC is more active at the time of choice in the complete feedback condition but not in the partial feedback condition (Fig 3.7D). We found no interaction between the horizon length and the feedback type in our VOI. Thus, a different subpart of the MCC that was sensitive to the horizon length, was sensitive to the type of feedback.

Behaviourally, we observed that strategic exploration was implemented by decreasing the influence of expected value on the choice. We therefore next looked for evidence of stronger expected value signals in complete feedback condition compare to the partial feedback condition. We tested the expected value of the chosen option, the unchosen option and the difference in expected values between the chosen and unchosen options. We only found activity related to the expected value of the chosen option. We found two clusters of activities bilaterally in the MCC (area 24') and the left dlPFC (area 46) that were modulated by the contingency between choice and information (Fig 3.7C; see Table B.1 for coordinates of cluster peaks). We again placed two ROIs by calculating the functional and anatomical overlap for Brodmann areas 24' and 46 and extracted the t-statistics of the regressor. Although this is not an unbiased test, we can see that the MCC and dlPFC seemed to be active when an option with a low expected value was chosen, whereas in the complete feedback condition, they were more active when choosing high expected value options (Fig 3.7E for illustration). We found however no difference of the

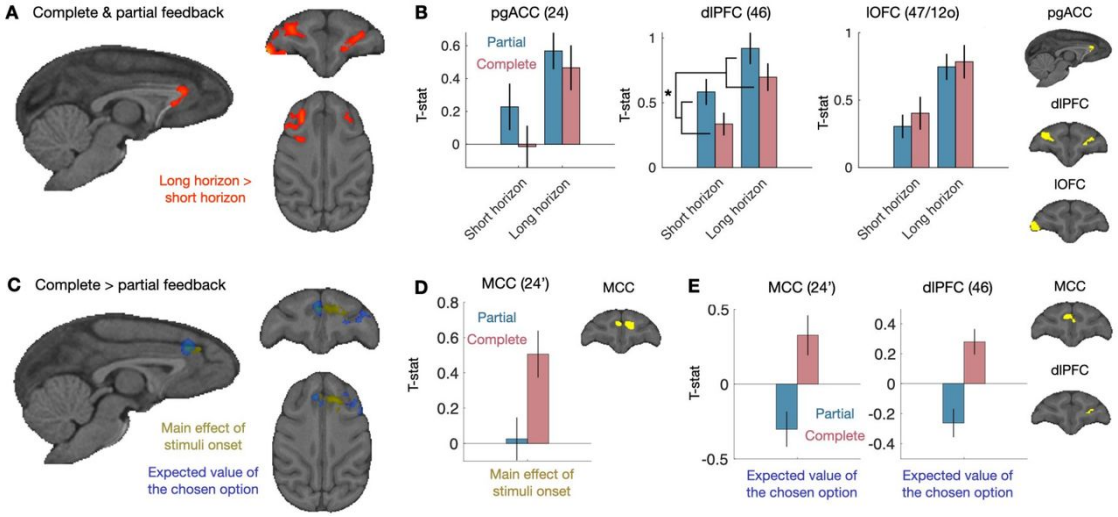


Figure 3.7: **First choice neural results.** (**A**) When combining partial and complete feedback sessions, we found clusters for a differential in activity in long horizon than short horizon in the pgACC, the dlPFC and the lateral OFC. Cluster $p < 0.05$, cluster forming threshold of $z > 2.3$. (**B**) We placed ROIs (in yellow) in the overlap of the functional cluster and anatomical region and extracted t-statistics for the difference between long horizon and short horizon. Mean across sessions (partial feedback: 40 sessions, complete feedback 34 sessions). (**C**) We looked for differences in how the contingency between choice and information (complete vs. partial feedback) modulates the initial information that was presented before first choices. Within our VOI, we found clusters of activity in MCC both for the main effect of feedback type and a greater sensitivity to expected value in the complete feedback condition. We also found a cluster of activity in dlPFC for a greater sensitivity to expected value in the complete feedback condition. (**D**) We placed an ROI (in yellow) in the part of MCC that is activated by the main effect of feedback type and extracted the t-statistics of the regressor for every session. We found that the effect we observe in the VOI is driven by increased activity in the complete feedback condition, whereas there is no activity in the partial feedback condition. Mean across sessions (partial feedback: 40 sessions, complete feedback 34 sessions). (**E**) We also placed ROIs (in yellow) in the parts of MCC and dlPFC where we found significant clusters in the VOI for the interaction of feedback type and expected value and extracted the t-statistics for the expected value regressor of every session. Plotting these regressors separately for feedback type reveals that both MCC and dlPFC were more active when an option with high expected value was chosen in the complete feedback condition, whereas they were more active when an option with low expected value was chosen in the partial feedback condition. Mean across sessions (partial feedback: 40 sessions, complete feedback 34 sessions). Error bars represent standard error of the mean.

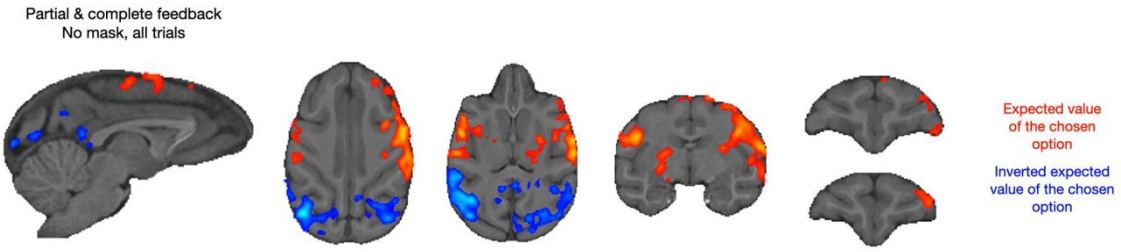


Figure 3.8: **Expected value of the chosen option.** Without mask and when taking the activity before the choice in all trials (not just first choice trials), we observed large activations related to the expected value of the chosen option (which is the same as the chosen action in our task) spanning from the motor cortex/somatosensory cortex, the dorsolateral prefrontal cortex, the OFC and striatum, as well as an inverted signal in the visual areas (Cluster $p < 0.05$, cluster forming threshold of $z > 2.3$).

strength of this sensitivity between short and long horizons. Thus, we found that the availability of the counterfactual feedback in the complete feedback condition decreased – and potentially even inverted – the sensitivity of the MCC and dlPFC to the expected value of the chosen option.

Finally, we looked for signals that were related to the expected outcome of the chosen option and that were common to both feedback conditions. Consistent with previous studies [170–173], when we combined the partial and complete feedback conditions session and took all trials in the 'wait' period, we found a large activation related to the expected value of the chosen option (which is the same as the chosen action in our task) spanning from the motor cortex/somatosensory cortex, the dlPFC, the OFC and striatum, as well as an inverted signal in the visual areas in the whole brain (without mask, Fig 3.8).

Overall, we found that pgACC and MCC reflected the horizon length and the type of feedback respectively. The dlPFC was linearly modulated both, with the strongest activation in the long horizon and partial feedback, when exploration is beneficial. Additionally, the feedback type modulated the effect of the chosen expected value on the activity of the MCC and the dlPFC, such that it was more active for low value choices only when obtaining information was contingent on choosing an option.

3.2.5 Chosen and counterfactual outcome prediction error signals in the OFC

We next examined the brain activity when the outcome of the choice is revealed ('outcome' period in Fig 3.1D) and macaques are updating their beliefs about the options. As the sequences of events played out differently in the partial and complete feedback conditions, we analysed each dataset separately in regard to feedback. At outcome, the partial feedback condition closely resembles previously reported results from fMRI studies in macaques [168, 169]. We looked for brain regions with an activity that was modulated by magnitude of the outcome prediction error signals, i.e., the difference between the outcome and the expectation (Fig 3.9A). Consistent with these studies, we found the expected clusters of activity in the medial prefrontal cortex and bilaterally in the motor cortex in our VOI (see Fig 3.9B for outcome-only related activity). When we time-locked our search to the onset of the reward (1 s after the display of the outcome, on a different GLM), we also found the classic prediction error related activity in the ventral striatum at the whole brain level (Fig 3.9C).

We then turned to the complete feedback condition, in which we simultaneous presented the outcome of the chosen and the unchosen, in order to examine the neural substrates involved in learning about counterfactual feedback and the extent to which they overlap with learning about chosen feedback. We looked in our VOI for brain regions with an activity that was modulated by the prediction error for the chosen option, and the unchosen option. We found a cluster of activity around the lateral OFC (lOFC, area 47/12o) that was negatively modulated by the prediction error for the chosen option and a cluster of activity around the medial orbitofrontal cortex (mOFC, area 14) that was negatively modulated by the prediction error of the unchosen option (Fig 3.10A; see Table **??** for coordinates of cluster peaks). These clusters intersected in the central part of the OFC (cOFC, area 13). Prediction error activity should show both an effect of outcome and expectation, with opposite signs. To independently test whether observed effects were prediction errors, rather than being driven by the outcome or the expectation

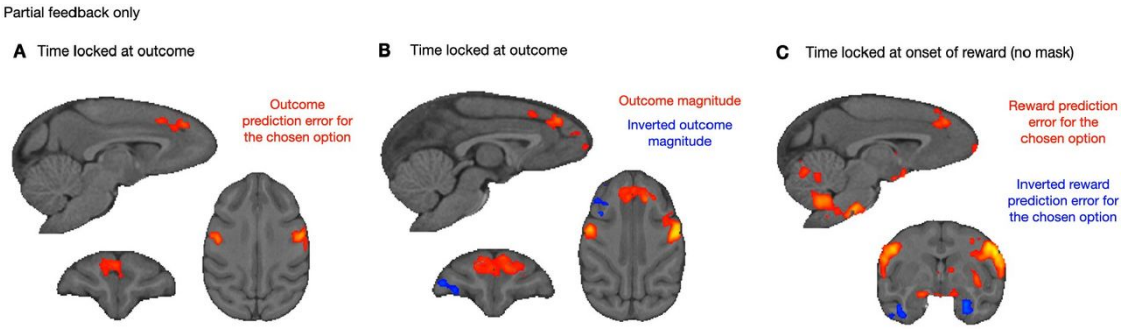


Figure 3.9: **Outcome prediction error and magnitude in the partial feedback condition.** (**A**) In the partial feedback condition and at the time of outcome, we found 3 clusters of activity that were positively modulated by the chosen option prediction error in the medial prefrontal cortex and bilaterally in the somatosensory and motor cortex in our VOI (Cluster $p < 0.05$, cluster forming threshold of $z > 2.3$). (**B**) We found the same 3 clusters when we looked for a positive modulation by the magnitude of the chosen outcome. We additionally found 1 cluster of activity in the right lateral OFC that was negatively modulated by the magnitude of the chosen outcome. (**C**) When we time-locked our search to the onset of the reward (1 s after the display of the outcome, with a different GLM), we found the same clusters as in A, as well as the classic prediction error related activity in the ventral striatum and a negative prediction error in visual areas (not shown) at the whole brain level.

alone, we extracted the t-statistics for both outcome and expectation in ROI defined by their outcome related activity only and looked for a modulation by the expectations (Fig 3.11). Again, we defined ROIs based the functional modulation by the magnitude of the chosen outcome and anatomical overlap. For the chosen outcome, we found that lOFC did not show a significant positive expectation for the chosen outcome ($p = 0.1083$) (Fig 3.10C). We found that the somatosensory cortex (area 3) showed a strong positive chosen outcome signal and as well as a positive modulation by the chosen expectation ($t(33) = 2.5246, p = 0.017$) and the ventrolateral prefrontal cortex (vlPFC) (area 45) had no sensitivity for the chosen expectation ($p = 0.95$) (Fig 3.11B). Using the same procedure with the unchosen outcome, we found that the cOFC showed a positive expectation about the unchosen outcome ($t(33) = 2.2617, p = 0.0304$), as well as a negative modulation by the chosen outcome ($t(33) = -2.8761, p = 0.007$) and a positive modulation by the expectation about the chosen outcome ($t(33) = 2.5560, p = 0154$). We found a similar pattern in the mOFC (unchosen expectation: $t(33) = 2.5130, p = 0.017$; chosen outcome:

$t(33) = -2.2455, p = 0.0316$; chosen expectation: $t(33) = 2.8729, p = 0.0071$). The ventral-medial prefrontal cortex (area 10m according to the atlas we used [154] but has been called 14m [174]) showed a negative modulation of its activity by the unchosen and the chosen ($t(33) = -3.079, p = 0.004$) outcomes but no sensitivity to the expectations. Overall, we found that the cOFC and mOFC both showed prediction error related activity for both the chosen and the unchosen outcomes, and with the same sign.

To test the OFC prediction error effects even further, we ran an exploratory correlational analysis, between the ROIs based the prediction error signal (t-statistic) and session specific t-statistic of the behavioural effect of the change in expected value on choices (estimated with a separate GLM for each session with the same regressors as in Fig 3.10B). We wanted to see whether the strength of the counterfactual outcome prediction error in the brain is predictive of how much an animal uses it in a particular session. Only in mOFC (and not cOFC) did we see the expected – albeit modest – correlation between increased negative counterfactual prediction error signals and increased behavioural impact of the counterfactual information (Fig 3.10D, $\beta = -0.1701 \pm 0.0971, t(32) = -1.7507, p = 0.0448$, one-sided test).

3.3 Discussion

Weighing up exploration to gather new information with exploitation of your current knowledge is a key consideration for humans and animals alike. Inspired by recent work carefully dissociating value driven exploration from simple lack of exploitation [27], we designed the horizon task to look at the behaviours and neural correlates of goal-directed evaluation of strategic exploration in rhesus macaques. While strategic value driven exploration is important to optimise the behaviour in time, it is equally important to be able to learn from observations related to choices not taken. In particular, being able to process counterfactual information during learning is key to optimise exploration for only the kind of situations when active sampling is necessary.

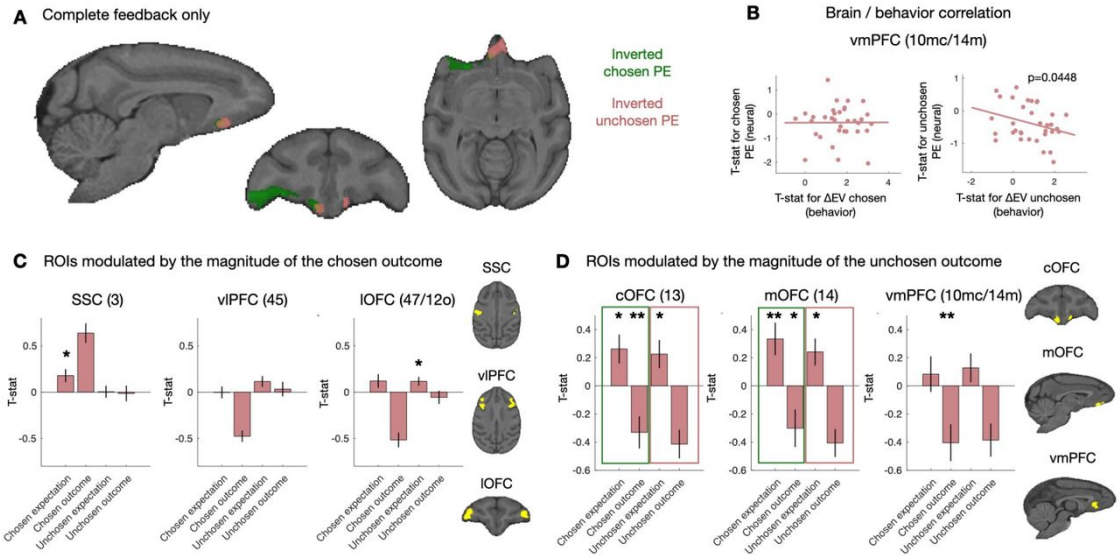


Figure 3.10: **Prediction error neural results.** (**A**) In complete feedback sessions only, we found clusters for inverted prediction error activity in the central part of OFC (area 13), extending into lateral OFC (area 47/12o). We also found inverted prediction error activity in the central OFC (area 13) and medial OFC (area 14) for the unchosen, counterfactual reward. (**B**) Brain-behaviour correlational analysis between the prediction error signal in the medial OFC (t-statistic) and session specific t-statistic of the behavioural effect of the change in expected value on choices (estimated with a separate GLM for each session). (**C**) We placed ROIs (in yellow) in the overlap of the functional cluster modulated by the magnitude of the chosen outcome and anatomical region. We extracted t-statistics for reward and expectation, both of the chosen and unchosen option. Prediction error activity should evoke both a reward and an expectation response with opposite signs. We did not found evidence for outcome expectation of the chosen option. Mean across complete feedback sessions (34 sessions). (**D**) When defining the ROIs (in yellow) according to the response to the magnitude of unchosen outcome, we find evidence for a classic reward prediction error and a counterfactual prediction error about the unchosen option both in central OFC and medial OFC: we observe activity related both to the obtained and the unobtained reward, and also activity related both the chosen and unchosen outcome expectation. Mean across complete feedback sessions (34 sessions). Error bars represent standard error of the mean.

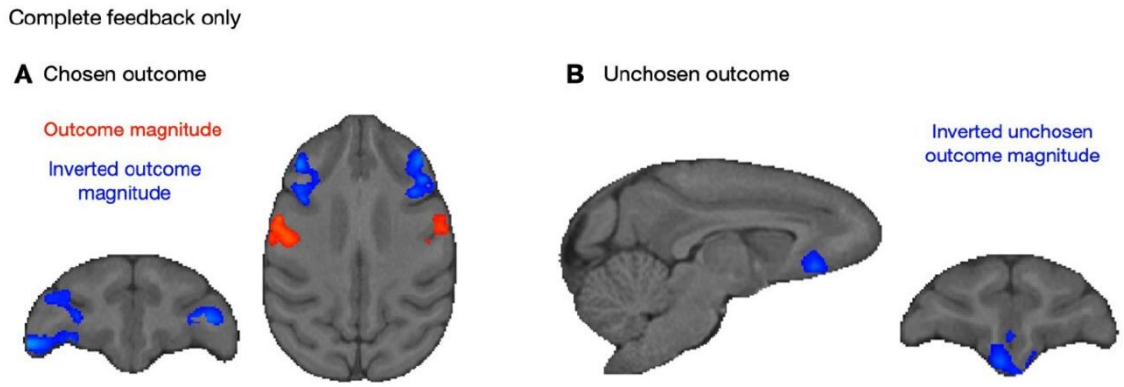


Figure 3.11: **Chosen and unchosen outcome magnitude in the complete feedback condition.** **(A)** In complete feedback sessions only, we found clusters for inverted chosen outcome magnitude activity in the right lateral OFC (47/12o) and bilaterally in the ventrolateral prefrontal cortex and 2 clusters in the somatosensory/motor cortex (3). **(B)** We found a cluster of activity for the inverted unchosen outcome magnitude in the central and medial OFC and the ventromedial prefrontal cortex.

3.3.1 Strategic exploration as a reduction of the effect of expected value on choices

We know that macaques can seek information before committing to a choice or to increase confidence about their decision [175–177]. However, here we showed that macaques could identify situations in which a strategic exploratory choice would lead to gaining information that would be beneficial for future decisions. Indeed, their choices were least influenced by expected value in the long horizon partial feedback condition, which is when there should be a drive to explore. This suggests that the animals had a representation of the significance of the information and used it to plan future actions. Our results demonstrate that they could discern both whether information will be useful in the future (greater exploration in long horizon) and that choosing an option is instrumental to get information about it (greater exploration in the partial feedback condition).

Exploration during value-based decision making has been conceived in different ways in the past. A simple way to account for exploration is the "epsilon-greedy strategy", in which a small fraction of choices is made towards the non-most rewarded option [22, 25]. Along the same line, another way to formalise exploratory choices is

through the noise or (inverse) temperature in the softmax choice-rule, which predicts that there are more exploratory choices when the expected values of the options are close [22, 25, 155, 156]. This process is also called *random exploration* because the relaxation of the effect of expected value on choices could allow for stumbling upon better options by chance [27, 159]. This form of exploration is negatively correlated with accuracy. Therefore, without varying the other features such as the usefulness of information for the future and the contingency between choice and information, it is impossible to know whether the animals made a mistake or were exploring the non-most rewarded option to obtain information about it.

Here we show that macaques, like humans, can perform sophisticated choices that take into account the prospective value of discovering new information about the options [60]. Our results revealed that foraging behaviours in macaques do not only rely on simple heuristics (e.g., win-stay/lose shift), but is also based on strategic exploration. To some extent it mirrors anticipatory switch to exploitative behaviour once enough information has been learned about the information, even when the expected outcome has not yet been obtained [178]. It also adds to recent works showing that complex socio-cognitive processes thought to be uniquely human such as mentalizing or recursive reasoning could be identified in rhesus macaques [179, 180]. However, contrary to human behaviour, our macaques only adapted by reducing their reliance on expected value (i.e., exploitative value) on choices, which corresponds to the degree to which they use random exploration. Humans also increase their preference for the most uncertain option when exploration is useful for the future, which has been referred to as directed exploration [27]. This suggests species-specificity in exploratory strategy.

3.3.2 Use of counterfactual feedback in subsequent choices

We next investigated whether and how the availability of the counterfactual feedback impacted their subsequent choices in the long horizon. First, we found that having more information about the options in the complete feedback condition improved accuracy. In general, macaques were more sensitive to the initial expected values of

the options when there was no contingency between choice and information, in the complete feedback condition, but then utilised the feedback about the chosen option to the same degree in both conditions. However, in the complete feedback condition macaques additionally also used the counterfactual feedback about the unchosen option to update their preference. Our results confirm that rhesus macaques are sensitive not only to direct reinforcers (i.e., the reward they obtain) but also counterfactual information [58, 163, 164, 181]. Here we clearly demonstrate that macaques can learn directly from counterfactual feedback, via prediction error. We also demonstrated this learning is associated with OFC activity. Identifying that an alternative action could have led to a better outcome and acting upon it has been shown to modulate OFC activity in rodents [70], suggesting that this ability was present in the last common ancestor to primates and rodents 100M years ago.

The availability of counterfactual feedback also helped compensate for the repetition bias that macaques displayed during the performance of the sequence. This form of engagement can also be considered in terms of the exploration/exploitation trade-off, where exploiting corresponds to staying with the current or default option: macaques committed to an option at the beginning of the trial and only changed option if there were sufficient evidence that it was worth it. Consistently, humans and animals show a tendency to over-exploit compared to the optimal policy in various tasks [182–184]. In general, there seems to be a cost associated with switching from the on-going representation or strategy to a new one [185, 186]. In our task, switching options also requires minimal physical effort as the macaques are positioned in the sphinx position in the scanner. Additional information about the options, particularly about the alternative option, seems to encourage the re-evaluation of the default strategy of persevering with the current option, enhance behavioural flexibility and increases the willingness to bear the cost associated with the physical resetting required by switching target.

3.3.3 Strategic exploration signals in ACC/MCC and dlPFC

Using fMRI, we investigated the neural correlates of the assessment of the possibility to use the information collected during the choice in the future, manipulated through horizon length, as well as the assessment of the contingency between choice and information, manipulated through the availability of the counterfactual outcome. Modulation of activity associated with exploratory behaviour in an uncertain environment has been recorded in humans and macaques in both the ACC and the MCC [25, 178, 187–189] but here we found interesting anatomical distinctions. We found that the pgACC was more active when the information could be used in the future in the long horizon. In humans, the pgACC activity has been shown to scale with the use prospective value more to guide choices [60]. Thus, the pgACC might be critical to organise the behaviour in the long run, beyond the immediate choice. The activity of a separate anatomical region of the MCC was modulated by the feedback type. The MCC has been shown to encode the decision to obtain information about the state of the world [190] and to integrate information about the feedback to adapt the behaviour [191]. Here we show that activity in the MCC was modulated prospectively by the feedback type. This activation was greater when more information was going to be provided i.e., in the complete feedback condition. Thus, the MCC could be involved in anticipating more learning or regulating exploration prospectively based on the feedback that will be received. Critically, the dlPFC displayed an additive effect of the usefulness of exploration for the future and the contingency between choice and information. It was most active when both were true, and exploration was sensible. Moreover, in the complete feedback condition, the MCC and the dlPFC were more active when the expected value of the chosen option was high. Such modulation is in line with studies in macaques and humans showing that neuronal activity in the MCC and the dlPFC does correlate with actions' values [192–195]. When the unchosen outcome was not available, MCC and dlPFC were more active when the expected value of the chosen option was low, which is consistent with the pursuit of an exploration strategy. Overall, the

coordinated roles of ACC and MCC participate to the regulation of exploratory /exploitative behaviours, not only in rhesus macaque but also in humans [196].

Computational modelling of the activity in ACC/MCC and dlPFC suggest that ACC/MCC could regulate decision variables in the dlPFC based on the strategic assessment for exploration [197]. Noradrenaline has been shown to modulate the noise in the decision process which could fuel random exploration and potentially give a mechanism for modulation of exploratory activity [35, 158, 184, 198]. Importantly, ACC/MCC also more generally interacts with other frontal lobe regions as well as monoamine systems and in particular the noradrenergic system making it a feasible mechanism for changing exploratory behaviours [199]. Specifically, a network consisting of the MCC, the dlPFC and potentially the locus coeruleus could support the relaxation of the effect of expected value on choices based on the context. Altogether, these results illustrate how ACC/MCC and dlPFC might dynamically switch modes to pursue different goals depending on the task demands [178, 187, 189]. Future studies will aim at testing whether switching mode is dependent on noradrenergic inputs and which causal role both regions play in changing into and out of strategic exploration.

3.3.4 Update signals for chosen and counterfactual outcomes in OFC

Being able to process counterfactual information during learning is key to reducing costly exploration to only the kind of situations when active sampling is necessary. Doing so requires an ability to process abstract information and learn from it similarly to experienced outcomes, without confusion between the two, which our macaques achieved. Neurally, we found classic activations in the partial feedback condition in response to the magnitude of the outcome and the prediction error of the chosen option. At reward delivery, we observed a prediction error signal in ventral striatum, which has previously been reported in neurophysiological studies [1]. We also observed prediction error activity at outcome in MCC, which had been shown previously in neurophysiological recordings [111]. We also found that

the prediction error for the chosen outcome modulated the activity of the OFC, but further examination showed that the lOFC was mostly sensitive to the chosen outcome. Previous studies have shown the lOFC involvement in learning and using choice-outcome associations to guide behaviour [200–202], and causal studies demonstrated its role in credit assignment [67, 203–205]. Here, in the presence of two outcomes – two stimuli – OFC could be crucial to integrate the information specifically related the chosen option. We also revealed that the central and medial OFC carried clear chosen prediction error signals.

However, our results go beyond chosen prediction error signals and add two additional dimensions to our understanding the neural processing of counterfactual information during exploration and learning. Firstly, we were able to map out for the first-time counterfactual prediction error signals in macaques in the cOFC and mOFC. Importantly, by using fMRI we could establish its specificity within the prefrontal cortex. In particular, we found signals for that counterfactual and the chosen outcomes but not the expectations in the vmPFC (10/mc14m). This adds to our knowledge of modulation of activity by counterfactual outcomes in gambling tasks had been reported in macaque lateral prefrontal cortex, MCC and OFC [163, 164]. Secondly, we found in the mOFC a relationship between the strength of the counterfactual prediction error signal and the extent to which the counterfactual outcomes influenced future choices (Fig 3.10). Encoding of the counterfactual outcome has also been observed in humans mOFC [71, 206, 207], and lesion of the mOFC in patients had been associated with an inability to use counterfactual information to guide future decisions [72]. Those results are compatible with the proposed broader role of the mOFC in representing abstract values [62, 208]. Here we show that it represents the comparison of the obtained counterfactual information with the expected counterfactual information. We found that the representation of the prediction error for the chosen and unchosen outcomes had the same sign at the time of outcome, which leads us to postulate that this update mechanism is independent of the frame of the decision [62, 207, 209].

Having identified the orbitofrontal source of counterfactual prediction errors in macaques opens up further possibilities to directly interfere with the neural processes in each system to see the effect it has on this complex adaptation of the animals' exploratory strategy. Furthermore, knowing how the brains of non-human primates might solve this complex sequential exploration task also sheds light on the building behavioural and neural blocks of reward exploration, learning and credit assignment.

3.3.5 Conclusion

Here we showed that macaques are able to assess the contingency between choice and information and the utility of information for the future when making strategic exploratory decisions. Different subparts of the ACC and MCC related to the assessment of these variables for strategic exploration, and the dlPFC represented them both additively, such that it was most active when exploration was beneficial. Only when the only way to obtain information was to explore did MCC and dlPFC show increased activity with less exploitative choices. This suggests a role in suppressing expected value signals when value guided exploration should to be considered. Importantly, to limit costly exploration to when it is necessary being able to process counterfactual information is key. We showed macaques could do this potentially by representing chosen and unchosen reward prediction errors in central and medial OFC. Furthermore, the strength of this signal in the mOFC was shown to be correlated with future decisions taken. Overall, our study shows how ACC/MCC-dlPFC and OFC circuits together might support exploitation of available information to the fullest and drive behaviour towards finding more information when it is beneficial.

3.4 Materials and Methods

3.4.1 Macaques

Three male rhesus macaques were involved in the experiment (Macaque M: 14kg, 7 years old, macaque S: 12kg, 7 years old and macaque E: 11 kg, 7 years old). They were kept on a 12-hour light dark cycle, with access to water 12–16 hours on testing

days and with free water access on non-testing days. All procedures were conducted under licenses from the United Kingdom (UK) Home Office in accordance with the UK The Animals (Scientific Procedures) Act 1986.

3.4.2 Task

During the task, macaques sat in the sphinx position in a primate chair (Rogue Research, Petaluma, CA) in a 3T clinical horizontal bore MRI scanner. They faced an MRI-compatible screen (MRC, Cambridge) placed 30cm in front of the animal. Visual stimuli were projected on the screen by a LX400 projector (Christie Digital Systems). The animals were surgically implanted under anaesthesia with an MRI-compatible cranial implant (Rogue Research) in order to prevent head movements during data acquisition. Two custom-built infrared sensors were placed in front of their left and right hands that corresponded to the stimuli on the screen. Blackcurrant juice rewards were delivered from a tube positioned between the macaque's lips. The behavioural paradigm was controlled using Presentation software (Neurobehavioral systems, Inc, CA, USA).

The task consisted of making choices between two options by responding on either the left or right touch sensor to select the left or right stimulus respectively. A trial consisted of a given number of choices (determined by the horizon length) between these two options (Fig 3.1A). Each option corresponded to one side for the entire trial (Fig 3.1B). After each choice, the animals received a reward associated with the chosen option (Fig 3.1E). The reward was between 0 and 10 drops (0.5 mL of juice per drop) and was sampled from a Gaussian distribution with a standard deviation of 1.5 and mean between 3 and 7. The means of the underlying distribution were different for the two options and remained the same during a trial, such that one option was always better than the other. After each choice, macaques also received a visual feedback on the reward (Fig 3.1E). This feedback was in the form of was an orange rectangle displayed in a yellow rectangular window, such that the wider the orange rectangle, the greater the amount of juice (Fig 3.1B). It remained on the screen for the remainder of the trial.

At the beginning of each trial, prior to making their first choice, macaques received 4 informative observations in total, which consisted of information about the reward they would have received if they had chosen the option (Fig 3.1B). This was displayed in the same manner as reward feedback and also remained on screen during the duration of the trial. For each informative observation, a non-informative observation was presented for the other option (Fig 3.1B). The non-informative observation was a white rectangle crossed by black diagonals. Half of the trials started with an equal amount of information about the two options (2 informative and 2 non-informative observations for each option), and the other half with an unequal (3 informative and 1 non-informative observations). The order and side were randomly determined.

A critical parameter was the number of choices in each trial (horizon length). In **short horizon** trials, macaques were only allowed 1 choice before a new trial with new stimuli started, whereas in **long horizon** trials, they were allowed to make 4 choices between the options. Horizon conditions were blocked (5 consecutive trials of equal horizons) and alternated in the session. A second key manipulation was whether feedback was received only for the option they chose (**partial feedback** condition) or whether they received information about both the reward they received for the chosen option and the reward they would have received for selecting the alternative option (**complete feedback** condition) (Fig 3.1E).

A trial would proceed as follows (Fig 3.1E, timings in Table 3.1): After an inter-trial interval during which the screen was black, the stimuli were displayed, consisting of a large grey rectangle and the 4 horizontal bars of feedback information (Fig 3.1BE). The length of the grey rectangle corresponded to the length of the horizon, which each line corresponding to a choice, simulated or actual. Informative or non-informative stimuli were displayed on the first four lines. After the display of the stimuli, a red dot at the centre of the screen disappeared and macaques were then allowed to choose between the two options by touching the corresponding sensor (in less than 5000 ms or the trial restarted). A red rectangular frame appeared around the line on the side of the chosen option. After a delay, the outcome – the reward

Choice	ITI	Go delay	Outcome delay	Reward delay	ICI
1st	4000 ± 1000 ms	500 ± 500 ms (1500 ± 200 ms for macaque E)	3500 ± 500 ms	1000 ms	2500 ± 500 ms
2nd to 4th	na		1500 ± 500 ms		1500 ± 500 ms

Table 3.1: **Timings**

feedback – was displayed inside the rectangle. In complete feedback condition only, the reward that would have been gained on the other side (informative stimulus) was also displayed at the same moment. After an additional delay, a white star appeared on the screen and the reward was delivered. After the end of the reward delivery, the star disappeared. In short horizon blocks, a new trial started after the inter-trial interval delay. In long horizon trials, the red dot appeared and then macaques could choose among the options. The events leading to the reward were similar than for the first choice, but the delays were shorter. At the end of the fourth choice, a new trial started. The feedback condition macaques were in was not explicitly cued but instead fixed both within and across several sessions (6 to 10 consecutive sessions). Sessions after a switch from one feedback condition to the other were included in the analysis since it only took one choice for macaques to know the feedback condition.

Macaque M performed 14 sessions in the partial feedback condition and 13 (2 were corrupted and unrecoverable for fMRI analysis) in the complete feedback condition, macaque S performed 13 and 12 (3 corrupted sessions) sessions in each condition respectively and macaque E performed 14 and 15 (1 corrupted session) sessions in each condition. Sessions with less than 50 trials completed for the horizon task or with more than 80% bias for one side were removed from the analyses.

3.4.3 Training

All macaques followed the following training procedure, which lasted several months in a testing room mimicking the actual scanner room: First, they learned the meaning of the informative observation stimuli by choosing between a rewarded (1 to 10 drops) and a non-rewarded (0 drop) observation stimulus, and then between

different rewarded (0 to 10 drops) observation stimuli. They next learnt to associate an option with an expected value by choosing between a non-rewarded option (0 drop) and a rewarded option, and then between rewarded options in the long horizon and partial feedback condition. We then introduced blocks of small and long horizon trials. The animals were then tested in the scanner room. They all had previous experience of awake behaving testing in the scanner. We discarded the first scanning session and then analysed the following ones if they corresponded to our inclusion criterions in terms of number of trials and spatial bias. Macaque M and S were introduced to the complete feedback condition during the training procedure; macaque E experienced it for the first time during testing.

3.4.4 Bayesian expectation model

We analysed the behaviour using an ideal Bayesian model which estimated the most likely next outcome given the previous observations about the options. Outcomes were randomly drawn form a distribution of mean μ and fixed standard deviation. $P(x|\mu)$ is the probability that an outcome x would be observed given that it came from a distribution of mean μ. Since outcomes were independently dram from a distribution of mean μ, the probability of observing a set of outcomes $\{x_1 \dots x_n\}$ was

$$P(\{x_1 \dots x_n\}|\mu) = \prod_{i=1}^{n} P(x_i|\mu).$$

Using Bayes' rule, we computed the probability that this observation was generated by a distribution of mean μ as

$$P(\mu|\{x_1 \dots x_n\}) = \frac{P(\{x_1 \dots x_n\}|\mu)P(\mu)}{P(\{x_1 \dots x_n\})} = \frac{\prod_{i=1}^{n} P(x_i|\mu)}{\sum_{j}^{N} \prod_{i=1}^{n} P(x_i|\mu_j)}.$$

For each observation, we then computed the probability of a new observation as

$$P(x_{n+1}|\{x_1 \dots x_n\}) = \sum_{j=1}^{N} P(x_{n+1}|\mu_j)P(\mu_j|\{x_1 \dots x_n\}).$$

Thus, we can compute the probability distribution of the future outcomes given a set of observations (Fig 3.1C shows how the distributions change with new observations).

In our model, the expected value (EV) of an option is the mean of the probability distribution of the set of observations. The uncertainty (U) about what the next outcome was represented by the variance of the distribution. In general, the more informative observations the subject has access to for an option, the closer the expected value to the actual mean of the underlying distribution and the smaller the uncertainty about this quantity. The weight of the expected value controls a specific form of exploration: the reward-based exploration. Reducing this parameter allows exploring options by relaxing the tendency to choose the most rewarded option.

3.4.5 Choice model fit

We first focused our analysis on the first choice of the trial because it was similar in terms of information content (4 informative observations) across horizon lengths and feedback conditions. Contrary to subsequent choices in the long horizon, the expected value and the uncertainty about the expected value (which decreases with the number of informative observations, from 1 to 3) associated with each option were uncorrelated on the first choice. Indeed, in the partial feedback condition, if the option with the higher expected value is chosen more often, the uncertainty about its expected value decreases specifically, inducing a correlation between expected value and uncertainty about it. For these first trials t, we model the probability of picking the option that is presented on the right side of the screen as

$$P(right_t) = \sigma(b_{SB} + b_{RB}RB_t + b_{horizon}horizon_t + b_{EV}EV_t + b_U U_t + b_{ERhorizon}ER_t horizon_t + b_{Uhorizon}U_t horizon_t) + e_t$$

using logistic regression. Here, σ is the sigmoid function, RB_t is a categorical predictor that control for a repetition bias, EV_t and U_t denote the difference between the expected value / uncertainty of the options on the right and left side of the screen, $horizon_t$ is a categorical predictor for whether trial t is a short or long horizon trial, and e_t is the residual.

For the remaining trials (second, third, and fourth choice in the long horizon), we are interested in whether the animals change their behaviour as new information becomes available. We model these trials as

$$\begin{aligned} P(right_t) = \sigma(&b_{SB} + b_{RB}RB_t + b_{\Delta chosen}\Delta chosen_t + b_{baselineEV}baselineEV_t \\ &+ b_{\Delta ERchosen}\Delta EVchosen_t + b_{baselineU}baselineU_t + b_{\Delta Uchosen}\Delta Uchosen_t \\ &+ b_{\Delta EVunchosen}\Delta EVunchosen_t + b_{\Delta Uunchosen}\Delta Uunchosen_t) + e_t. \end{aligned}$$

For this logistic regression, we used an additional bias: $\Delta chosen_t$, that corresponds to the number of times the option on the right was chosen during the trial. Here, $baselineEV_t$ and $baselineU_t$ are the difference between the expected value / uncertainty of the right and the left option at the first trial within the horizon. As such, these regressors capture the impact the initial information displayed on screen has on subsequent choices. $\Delta EVchosen_t$ and $\Delta Uchosen_t$ capture the difference between the initial baseline and the information presented at the current trial based on the choices the animal has made. That is, these regressors capture the update of outcome expectation and uncertainty between the right and the left option compared to the first choice based on the consequent rewards the animals experienced.

In our complete feedback condition, the animals can also learn about the reward they would have gotten, had they chosen the other option. This is not captured by $\Delta EVchosen_t$ and $\Delta Uchosen_t$ as these regressors only take the experienced (i.e., obtained) reward into account. To see how the unobtained reward affects choices, we included the regressors $\Delta EVunchosen_t$ and $\Delta Uunchosen_t$. These regressors are computed as the difference between the full outcome expectation and uncertainty (based on both the obtained and unobtained reward), and the outcome expectation and uncertainty for the obtained reward only. Just as with the $\Delta EVchosen_t$ and $\Delta Uchosen_t$, these regressors are also again constructed as the difference between the right and left option, and with the baseline subtracted.

To fit these models, we used STAN (https://mc-stan.org) and brms with the default priors [210, 211]. For each model, we ran 12 chains, each with 1000 iterations

after a warm-up of 1000 samples. We allowed all regressors to vary by condition (partial and complete) and animal (3 animals) as fixed effects. We modelled testing sessions as random effects with different Gaussians for each animal. That is, for each regressor and each animal we estimated the Gaussian distribution that session-level regressors are most likely drawn from. Group-level estimates of the coefficients were obtained by averaging across animals and/or conditions. To determine statistical significance, we counted the number of samples of the posterior that are greater/smaller than 0.

3.4.6 MRI data acquisition and pre-processing

Imaging data were acquired using a 3T clinical MRI scanner and an 8-cm-diameter four-channel phased-array receiver coil in conjunction with a radial transmission coil (Windmiller Kolster Scientific, Fresno, CA). Structural images were collected under general anaesthesia, using a T1-weighted MP-RAGE sequence (resolution = 0.5 × 0.5 × 0.5mm, repetition time (TR) = 2.05 s, echo time (TE) = 4.04ms, inversion time (TI) = 1.1s, flip angle = 8°). Three structural images per subject were averaged. Intramuscular injection of 10 mg/kg ketamine, 0.125–0.25 mg/kg xylazine, and 0.1 mg/kg midazolam were used to induce anaesthesia. Functional MRI data were collected while the subjects performed the task with a gradient-echo T2* echo planar imaging (EPI) sequence (resolution = 1.5 × 1.5 × 1.5mm, interleaved slice acquisition, TR = 2.28s, TE = 30ms, flip angle = 90°). To help image reconstruction, a proton-density-weighted image was acquired using a gradient-refocused-echo (GRE) sequence (resolution = 1.5 × 1.5 × 1.5 mm, TR = 10 ms, TE = 2.52 ms, flip angle = 25°) at the end of the session.

fMRI data were pre-processed according to a dedicated nonhuman primate fMRI pre-processing pipeline [76, 148, 169]combining FMRIB Software Library (FSL), Advanced Normalization Tools (ANTs), and Magnetic Resonance Comparative Anatomy Toolbox (MrCat; https://github.com/ neuroecology/MrCat) tools. In brief, T2* EPI data were reconstructed using an offline SENSE algorithm (Windmiller Kolster Scientific, Fresno, CA). Time-varying spatial distortions due to body

movement were corrected by non-linear registration (restricted to the phase encoding direction) of each slice and each volume of the time series to a reference low noise EPI image for each session. The distortion corrected and aligned session-wise images were first registered to the animal structural image and then to a group specific template in CARET macaque F99 space. Finally, the functional images were temporally filtered (high-pass temporal filtering, 3-dB cutoff of 100s) and spatially smoothed (Gaussian spatial smoothing, full-width half maximum of 3m).

3.4.7 fMRI analysis

We conducted our fMRI analysis using a hierarchical GLM (FSLREF). Specifically, we first fitted each individual session (in session space) using FSL's fMRI Expert Analysis Toolbox (FEAT). We then warped the session-level whole-brain maps into F99 standard space, before combining them using FEAT's FLAME 1+2 random effects procedure. Here, we used contrast to obtain separate estimates for the partial and complete sessions, the difference between partial and complete sessions, and their average. To determine statistical significance, we used a cluster-based approach with standard thresholding criteria of $z > 2.3$ and $p < 0.05$. To increase power, we ran this cluster-correction only in an *a priori* mask of the frontal cortex that was previously used [169].

On the session level, we included 58 regressors for the partial feedback sessions, and 73 (including the same 58 regressors as in the partial feedback condition) for the complete feedback sessions. On top of these regressors we also included nuisance regressors that indexed head motion and volumes with excessive noise. All regressors were convolved with an HRF that was modelled as a gamma function (mean lag = 3, standard deviation = 1.5 s), convolved with a boxcar function of 1 s.

The two main periods of the task we were interested in were when the stimuli first appeared on screen, and when the outcome appeared on subsequent choices in the long horizon trials. At stimulus onset we included a constant and regressors for the expected value of the chosen and unchosen options, and also regressors for the uncertainty of the chosen and unchosen option. To allow us to examine

the effects of these five regressors on first choices in short and long horizons, and subsequent choices within the long horizon, we up each regressor by horizon and choice number (first choice short horizon, first choice long horizon, second choice long horizon, third choice long horizon, and fourth choice long horizon) for a total of 25 regressors. At outcome we included another constant, the expected value of the chosen and unchosen options, the reward obtained on this trial, the absolute value of the prediction error of this trial (|reward - expected value|), and the update in uncertainty on this trial. Again, all of these regressors were split up by horizon and choice number, for a total of 30 regressors at outcome. On top of these regressors of interest we also included 3 control regressors: the log response time at stimulus onset, a constant at decision, and the response side (left or right) at decision. In the complete feedback condition, we included additional regressors: at outcome, we added regressors for the reward of the unchosen option, the absolute prediction error for the unchosen option, and the update in uncertainty for the unchosen option. Splitting these regressors up by horizon and choice within a horizon yields an additional 15 regressors.

Having split up all regressors this way into choice horizon and number of choices within a horizon, we used planned contrasts combining them again to answer our questions of interest. At stimulus onset we were only interested in first choices, as this allowed us to compare whether the animals represented expected value and uncertainty differently depending on condition (partial or complete feedback) and/or choice horizon (long and short). We thus constructed contrasts adding up and subtracting the first choices on long and short horizons for the constant, the expected value and the uncertainty. At outcome we were interested in reward effects and updates to the expected value of stimuli. As this should happen not just after first choices in a horizon but all choices, we used contrast to construct (weighted) averages of our regressors combining all choices within horizons. Moreover, to look at the effect of (signed) prediction errors, we use contrasts that subtract the expectation from the reward.

To visualise the cluster-corrected effects we find in our mask of the frontal cortex, we use an atlas of the macaque brain [154] to identify the regions where we observe activity. We then create ROIs by calculating the overlap of the anatomical region according to the atlas (dilated with a kernel of 3x3x3 voxels), and the functional activation we found. By extracting the average t-statistics in this region we are able to visualise the effects we found, and also examine the individual components that contributed the effects (e.g., the reward and outcome expectation for prediction errors).

To best describe the localisation of orbitofrontal activities, we considered 3 orbital subdivisions based on their respective position on the orbital surface. Lateral to the lateral orbitofrontal sulcus is the lateral OFC; medial to the medial orbitofrontal sulcus is the medial OFC. In between the two sulci is a region we referred to as the central OFC. Such parcellation resembles subdivisions considered in humans and rodents [203, 212, 213], although alternative labels have been proposed [202].

To best describe the localisation of cingulate activities we considered a dissociation between anterior and mid-cingulate subdivisions as proposed by Vogt and colleagues [214, 215].

4

General mechanisms of sustained engagement in macaques

Contents

Abstract

Staying engaged with a task is necessary to maintain goal-directed behaviours. Engagement varies depending on the specific task at hand but it also intrinsically fluctuates widely and continually in daily activities. This intrinsic component of

engagement is difficult to isolate behaviourally or neurally in controlled experiments with humans. By contrast, animals spontaneously move between periods of complete task engagement and disengagement, even in experimental settings. We, therefore, looked at behaviour in macaques in a series of four tasks while recording fMRI signals. We identified consistent patterns of behaviour predictive of impending task disengagement. This made it possible to build models capturing task-independent engagement and to link it to neural activity. Across all tasks, we identified common patterns of neural activity linked to impending task disengagement in mid-cingulate gyrus. By contrast, activity centred in perigenual anterior cingulate cortex (pgACC) was associated with maintenance of task performance. Importantly, we were able to careful control for task-specific factors such as the reward history, and other motivational effects, such as response vigour, as indexed by response time, when identifying neural activity associated with task engagement. Moreover, we showed pgACC activity had a causal link to task engagement; transcranial ultrasound stimulation of pgACC, but not of control regions, changed task engagement/disengagement patterns.

4.1 Introduction

Everyone experiences how they are more or less engaged with tasks that need doing throughout the day. While some of our motivation is clearly linked to specific tasks and incentives, we also find ourselves from time to time either demotivated or full of vigour regardless of the task at hand. Furthermore, while we might stay disengaged for longer periods of time, motivation can also suddenly collapse for periods of time leading to distracted behaviour such as checking one's phone. Moreover, in some people, such periods of demotivation are especially prominent; apathy – sustained periods of motivational collapse – is a core, transdiagnostic feature of psychological and neurological illnesses [216, 217].

Such fluctuations occur even though engagement must be sustained across extended periods of time for many goal-directed behaviours to succeed. Additionally, ideally, when performing a task, it is important to stay engaged independently of the

specifics of the task at hand. Important insights into the mechanisms of motivation have been gained by investigating motivation changes occurring in response to specific external factors such as reward incentives or other feedback [218]. However, motivation is also subject to intrinsic fluctuation and must be maintained despite adverse external factors. Likewise, sometimes motivation is lost despite sufficient incentives. It has been proposed that staying motivated requires cognitive resources that are depleted by effort and that can be restored by taking breaks [219].

While motivation is often considered a unitary entity, to understand it more fully, we need to consider the possible existence of separable components. Changes in response vigour [220] and speed [88, 148, 221–226] occur as motivation waxes and wanes. However, variation in response vigour and timing may occur only if a person decides to carry on performing a task. Therefore, a separate and potentially even more fundamental process is deciding whether or not to engage in the task at all or to pause and disengage completely. Importantly, these measures differ from attention lapses as indexed by erroneous responses that have also previously been studied in the context of motivation [227].

In the present study, we focus on general mechanisms of task engagement and disengagement across a series of four different tasks while recording brain activity using fMRI. In this way, we can identify neural activity changes in moments when an agent spontaneously and completely disengages from a task independently of the concurrent specific, external task demands. We used macaque monkeys to examine these issues for several reasons. The social and other demands of human neuroimaging experiments usually ensure that human participants exhibit continuous task execution; their performance scores may fluctuate but human participants rarely give up and spontaneously stop altogether in the same manner that they do frequently when outside the laboratory. That is, to make human participants stop doing a task in the laboratory, the difficulty of the task needs to be increased, while social demands guarantee that they stay engaged otherwise. Macaques, however, while engaged for the majority of the experiments, repeatedly and reproducibly both disengage and re-engage for some periods of time during daily testing in the

laboratory, even for simple task that do not require much effort [190, 228]. While this is generally a great nuisance for the researchers, for our study it is fortunate as it allowed us to construct and fit models to these disengagements and link them to their neural substrates. Using data from four diverse decision-making tasks allows us to find behavioural and neural signatures that are task-general. Importantly, these disengagements are not part of the task design but occur spontaneously despite the reward incentives provided by the tasks. Moreover, by controlling for variation in extrinsic experimental factors, such as reward levels, we can capture engagement and disengagement due to task-independent factors. Intrinsic motivation has previously been linked to satiation (for example, cumulative reward, [229] or time spent on task, e.g. [230]). By also controlling for these factors, we aim to capture the intrinsically fluctuating aspect of task-independent engagements and disengagement [231].

While motivation is continually fluctuating during extended activity [232], disengagements are all or none events. For example, one might feel more or less motivated to do a chore throughout the day – which we call the general engagement - and occasionally disengaging from one's work altogether when motivation drops. We examined neural activity related to both slow fluctuations and sudden pauses in engagement, both in aggregate and separately. To do this we used a new approach that considers the distribution of tasks engagements and disengagements to estimate continual variation in a general state. Such a state tracks the current level of engagement above and beyond the current trial. This allowed us to identify events when animals suddenly and 'surprisingly' disengage in an otherwise engaged state. By contrast, we can also identify 'expected' disengagements in a state of low general engagement. This allowed us to examine the neural activity concurrent with both general engagement and expected disengagements as well as surprising disengagements. Importantly, we contrasted these model-derived estimates of engagement with other distinct aspects of motivation such as changes in response vigour indexed by reaction time. This made it possible to dissociate signals leading to task engagement or disengagement from neural activity related to variation in motivation to execute a specific action quickly. By using a whole brain imaging

technique such as fMRI, we can seek neural correlates of engagement throughout the brain during all four tasks. This is important as the neural circuits linked to task engagement/disengagement are not well defined. However, we note that areas of anterior cingulate cortex (ACC) and adjacent medial frontal cortex have been linked to intrinsically motivated behaviours [233], mood fluctuation /citeVinckier2018, and neural activity has been reported to change in some related situations [228, 234], particularly when driven by endogenous factors such as satiety [235].

Our fMRI analysis identified one important area of activity change in perigenual ACC (pgACC) that was prominent across all four tasks. We therefore used neurostimulation data which causally manipulates this region to test its importance for task engagement: Specifically, one of the datasets used in our analysis had stimulated pgACC using transcranial ultrasound stimulation (TUS), and thus allowed us to compare the effect of pgACC stimulation against other control regions. Not only did we examine the impact of TUS on pgACC and compare it to sham TUS but in addition we also examined the impact of TUS to the basal forebrain (BF). BF TUS leads to changes in motivation-related influences on action timing [148] and so it provides an especially strong comparison with pgACC TUS. In addition, we examined the impact of TUS of an additional control region in the parietal operculum (POp) – a region in which task-related and task-initiation related activity had not been observed – to control for general cortical stimulation effects.

4.2 Results

We combined data from four different reward-based decision-making tasks [76, 148, 169, 236, 237], two of which were discussed in detail in Chapters 2 and 3. The tasks covered a range of different paradigms: simple stimulus-response mapping, incentivised exploration/exploitation, incentivised delayed responses, and novel value inference (see original publications for additional details). In each case, the animals occasionally disengaged from the task and stopped responding before re-engaging after some time. For the purpose of our analysis, we define disengagements as responding after 3 s or later, or not responding at all during a trial, i.e. it timed

out. However, for one of our tasks that incentivised late responses [148, 237], we only counted trials as disengaged where the animal did not respond at all (see Fig 4.1 for details for all tasks). We binarised trials into ones where the animals are engaged or disengaged (Fig 4.2A). Overall, we started with 17 datasets in 13 animals but excluded six datasets from five animals that disengaged in less than 5% of trials on average across sessions (Fig Fig 4.1). The two animals that participated in [76] also participated in [148, 237], which left eleven datasets from nine unique animals.

Our aim was to use disengagements to construct variables that, on a trial-by-trial basis, capture different aspects of task-independent disengagement, i.e. regardless of the specific details of each task. We then used these variables in an fMRI analysis to identify their neural correlates.

To contrast engagement/disengagement with variation in motivation related to response vigour and speed, we repeated the same analysis using response times (RTs). For this control analysis we only used data on engaged trials (we did not analyse the trials classified as disengagements in which, by definition, no response or delayed response is made; see Fig 4.1). For this analysis, we used data from 13 (unique) animals because we now had sufficient data from more animals to include in the analysis. However, we avoided considering data from one of the previous tasks [148, 236, 237] because the animals performing it were sometimes incentivised to respond late as part of the task design and thus RTs do not provide the simple measure of vigour in the same way as in other tasks.

4.2.1 Behavioural results

For each task, we constructed separate regression models that accounted for the extrinsic variables that could be measured in each experiment by the investigators. These models included regressors such as the task stimuli encountered, the responses made, the rewards animals received, and the trial number (see *Methods* for the specific models for each task). Using these models, we can account for variance in task-engagement and disengagement that is due to extrinsic factors. These regressors are, of course, the ones that are usually the focus of any analysis of

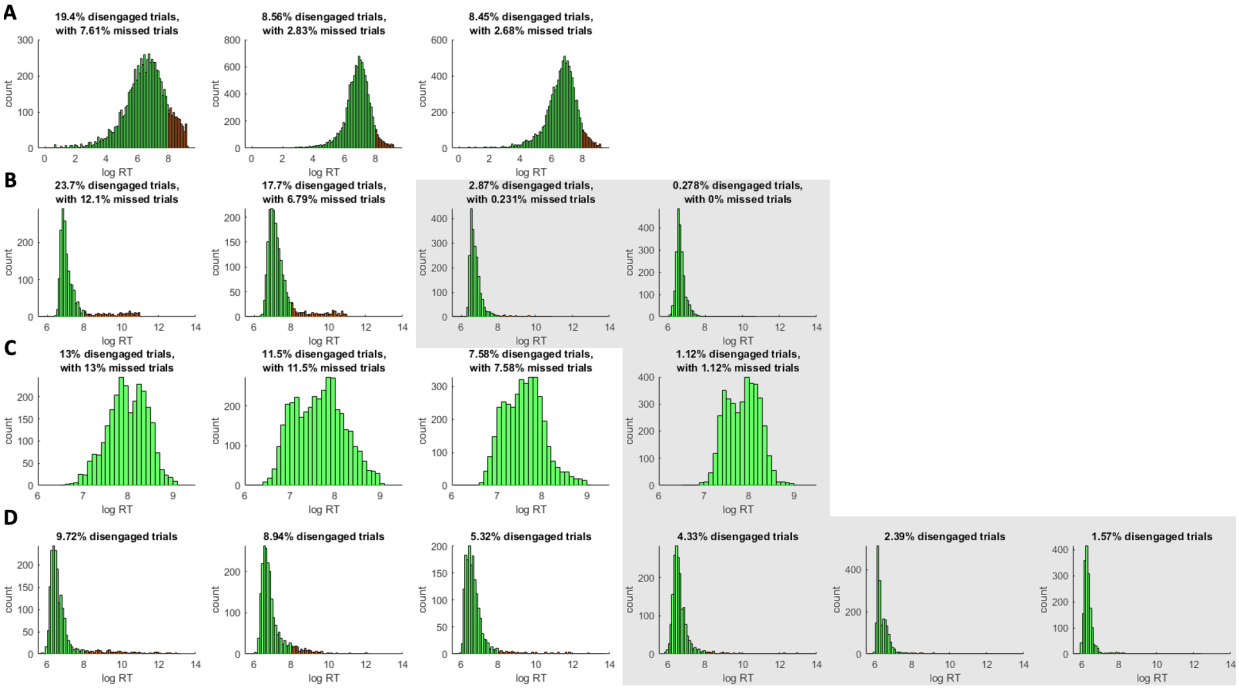


Figure 4.1: **Response time histograms**. For each monkey, we binarised trials into engaged (green) and disengaged trials (red). Trials were coded as disengaged if the monkey responded slower than 3s ($\log(3000\text{ ms}) \approx 8$) or did not respond at all (missed trials). Animals that had disengaged on less than 5% of trials were excluded from the study (grey background) (**A**) Shows the data from Chapter 3 [236]. In this experiment a trial timed out if the animal did not respond within 10s. The percentage of 'disengaged trials' in the panel title also contain these 'missed trials'. (**B**) shows the data from Bongioanni et al. [76]. Here a trial times out if the animal did not respond within 60 s. 2 of the 4 animals were excluded from the study as they did not have enough disengagement trials to qualify. (**C**) shows the data from Khalighinejad et al. [148, 237]. In this task, unlike the other 3 tasks, animals had an incentive to delay their response and respond later in a trial to potentially obtain more reward. As such, we did not count trials with RT $>$ 3s as disengaged trials for this task, and only used the missed trials in which the animals did not respond at all to code disengaged trials. 1 animal had to be excluded because it did not have enough overall disengaged trials. (**D**) shows the data from Chapter 2 [169]. In this task trials never timed-out if the animal did not respond, but instead the stimuli stayed on the screen until the animal re-engaged with the task. As such, there are no missed trials for this task. 3 of the 6 animals had to be excluded because the total number of disengaged trials was less than 5%.

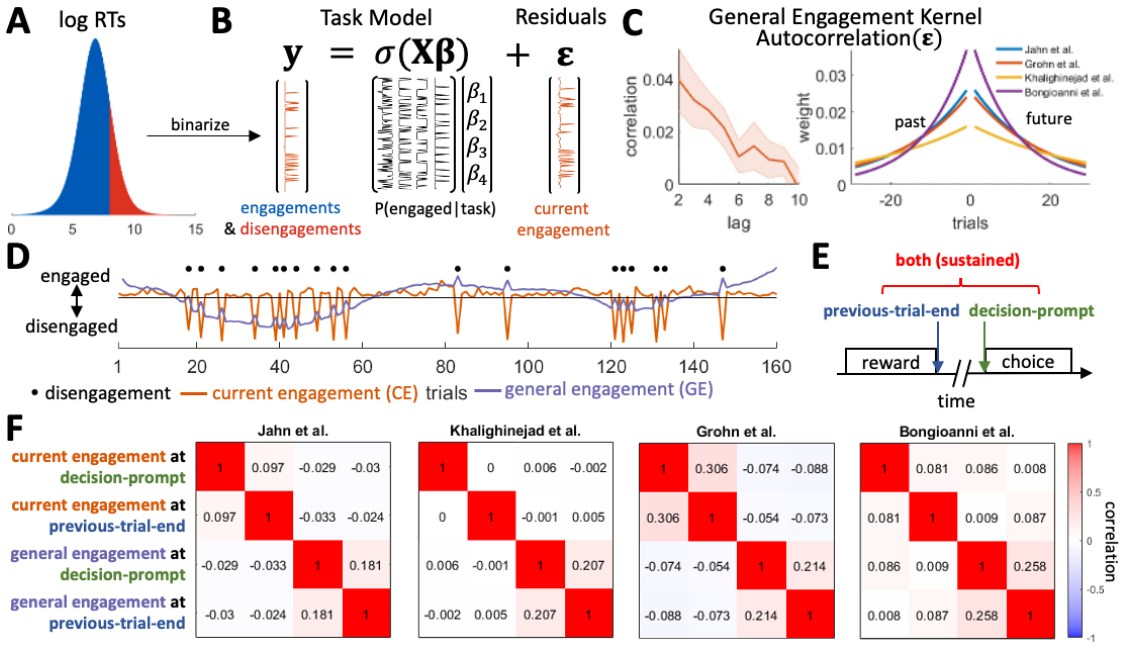


Figure 4.2: **Behavioural results and fMRI design.** (Continued on the following page.)

a neural data set. However, by regressing out the variance due to all extrinsic factors (i.e. taking the residual error of the regression models) we are left with the components of task-engagement and disengagement that are due to what is normally considered residual fluctuations in behaviour that typically receive little investigative attention (Fig 4.2B). However, these residuals also capture task engagement and disengagement that is dependent on intrinsic variation. As such, they capture the intrinsic level of *current engagement* (CE).

If motivation is indeed drifting across trials, then we should be able to observe clustering in the residuals. To this end, we examined its autocorrelation. If engagement and disengagement were solely determined by extrinsic task features, then the residuals would not be autocorrelated over trials. However, in our data we find a persistent autocorrelation of CE, the residuals (Fig 4.2C left; significant for lags < 10 at $p < 0.05$ with Bonferroni correction; we exclude lag $= 1$ because in some tasks repeated disengagements were impossible, as the experiment waited for the animal to re-engage before continuing). In other words, periods of engagement and disengagement are temporally clustered.

Figure 4.2: **Behavioural results and fMRI design.** (**A**) We binarised animal's RTs into trials in which they were engaged or disengaged. On disengaged trials the animals took longer than 3 s to respond, or did not respond at all (i.e. the trial timed out). Fig 4.1 shows the individual RT distributions for each animal. (**B**) To control for the influence on motivation exerted by extrinsic task event-related factors, we constructed separate logistic regression models for each of our four tasks. Each model contained task-specific regressors (see Methods for details) as well as regressors coding for the previous five rewards/non-rewards the animals received at the end of each trial, the current cumulative reward, and the trial number. By regressing out the effects these variables have on engagement, we were left with the residuals. These residuals contain the fluctuations in motivation that are intrinsic as opposed to those that are due to extrinsic factors related to task structure and task events. We refer to this index as the intrinsic level of current engagement (CE). (**C**) (Left) We find a persistent autocorrelation of the residual fluctuations suggesting that intrinsic CE – engagements and disengagements – are temporally clustered. Shaded error represents the standard error of the mean across data sets. The average correlations for each lag from 2 to 10 are: 0.040, 0.032, 0.028, 0.022, 0.010, 0.014, 0.010, 0.009, and -0.002. (Right) By fitting exponential kernels to the index of the intrinsic CE (the residual fluctuations) separately for each of the four tasks, we can also capture this autocorrelation. (**D**) The same kernels can then be used to smooth the estimate of the intrinsic CE (orange line, shown for an example session) on each trial in each task. As a result, an estimate is obtained of the slowly fluctuating general engagement (GE) of an animal that can be made available for each trial (purple line, shown for an example session). (**E**) To capture effects of motivation in a similar manner in our neural analyses of all four tasks, we time-locked to two events in each trial that all our four tasks have in common: the end of the reward delivery in the previous trial, and the onset of the decision-prompt in the current trial. The rationale for looking at both of these time-points is that it's not a priori obvious when, during a trial, motivational effects should be most prominent; arguably motivation might be expected to produce sustained activity patterns that are observable at both time-points. (**F**) Even after their hemodynamic convolution with the relatively fast hemodynamic response function observed in macaques [238, 239], there is limited correlation between these regressors in all four tasks. Note also that time-shifted regressors (similar regressors but time-locked to previous-trial-end or current trial decision-prompt) are relatively uncorrelated because the task-designs ensured sufficient time intervals between the end of one trial and the beginning of the next in all four tasks.

We can use this autocorrelation to estimate the motivational state the animal is in on any given trial. We refer to this variable as the *general engagement* (GE). While CE corresponds to the residual fluctuations in Fig 4.2B, GE reflects the motivational state associated with a weighted average of CE on the current and surrounding trials: If the animal disengages on previous/future trials, we can assume it is also, to some degree, in a disengaged state currently. Conversely, if it is engaged on these trials, we can assume it is also, to some degree, in an engaged state currently. To this end, we fit exponential kernels to the residual fluctuations (Fig 4.2C right shows the fitted kernel for each of the four tasks). These kernels capture the extent to which the intrinsic motivation on a trial, as indexed by the residual fluctuations, is related to the intrinsic motivation on trials before and after it. Smoothing the residual fluctuations (CE; orange line in Fig 4.2D) by these kernels allows us to obtain an estimate of a continuously varying GE (blue line in Fig 4.2D) on each trial.

We can also combine the estimates of CE and GE to obtain two derived quantities that are used in first stages of the neural analysis as contrasts. First, we can average the current CE index with the continuously varying GE index to obtain an estimate of a third variable we refer to as *overall engagement* (OE). OE provides an overarching estimate of engagement on any trial as it uses both the engagement on the current trial (as given by CE) and of the surrounding trials (as given by GE) to index engagement, and so it is a useful starting point for neural analyses; as explained in more detail below, we can first identify areas in which activity is related to OE and then we can examine whether the activity tracks CE, the more slowly varying GE, or both quantities. Thus, CE and GE can also be thought of as the separated trial and state components of an overarching model that indexes OE. Second, we can subtract the model-derived estimate of GE from the CE level to identify *engagement shifts* (ES) when an animal's motivation suddenly collapses and there is abrupt task disengagement; the animal may be disengaged on the current trial even though the events that normally surround a disengagement were not observed. This allows us to examine CE that is unexpected given the current level of GE; i.e. it allows us to identify trials with low engagement in an otherwise

highly engaged state. Importantly, for the purpose of our neural analysis, both MS and OE can be constructed by subtracting/adding CE and GE on the contrast-level within a single general linear model.

We repeated an analogous, control analysis of RTs – an index of motivational change in relation to response vigour as opposed to task engagement. However, this analysis was performed on engaged trials only; responses were only made, and RTs were only measurable on engage trials (Fig 4.3A-C). We again find that, after having regressed out the variance in RTs due to task-manipulations, the error in RT estimates is autocorrelated over trials (significant for lags < 8 at $p < 0.05$ with Bonferroni correction). We refer to these residual fluctuations as *trial vigour*. By fitting exponential kernels to trial vigour, we again obtain estimates of a general *state vigour* on each trial. The GE and general state vigour estimates are analogous state-related variables but they are only weakly correlated (Fig 4.3D) and thus reflect different potential motivational processes. Just as for ES and OE, we can also subtract trial vigour and state vigour to obtain similar contrasts to use in our neural analysis. Once again these vigour-related variables were uncorrelated with our key task engagement/disengagement related variables of interest.

4.2.2 FMRI results

As in the behavioural analyses, we constructed a separate neural regression model for each task that captured all aspects of the extrinsic task variables (see *Methods* for the specific models). In addition to these task-specific models, we also included regressors that captured the motivational factors that we identified in our behavioural analysis (Fig 4.2C). Because the neural activity we are interested in is related to overarching engagement that is not necessarily associated with any one event that occurred during the task, we time-locked our regressors to two separate points within each trial that all four tasks had in common: (1) we time-locked to the decision-prompt on each trial when monkeys were asked to make a choice, and (2) we time-locked to the end of the outcome-period of the previous trial [187]. This ensured we had a measure of activity when task-specific performance and learning in a trial had been

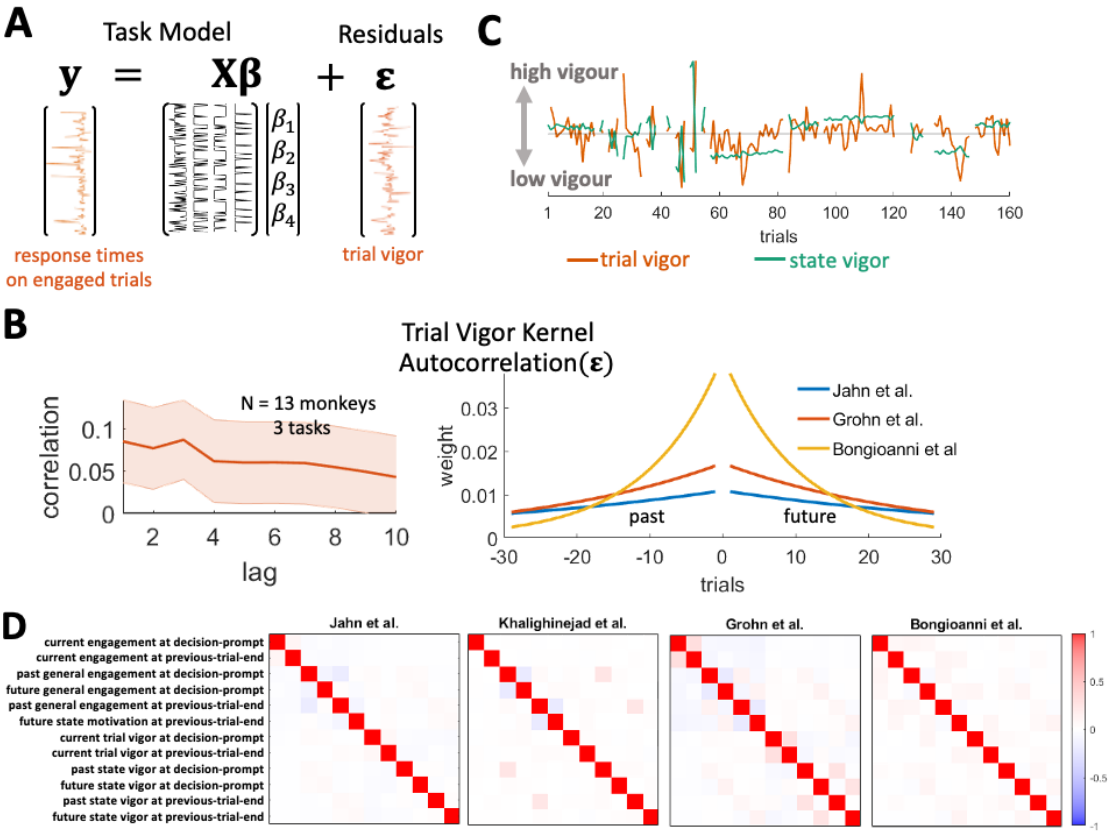


Figure 4.3: **Vigour as indexed by RTs and whole-brain GLMs.** (**A**) Just like for disengagement and engagement as shown in Figure 1, we also constructed separate regression models for the RTs on each task. By regressing out the task here, we were again left with an estimate of task-intrinsic motivation (trial vigour). (**B**) The residuals of our regression models for RTs are autocorrelated over trials (left; significant for lags < 8). We also again fitted exponential kernels to the residuals of each task (right). (**C**) Smoothing the residuals using these kernels gave us an estimate of the motivational state an animal is currently in (state vigour). This estimate of the motivational state is distinct from the one shown in Figure 1 as it is based on the residual of the RT regressions and not on the engagement/disengagement regressions. (**D**) Correlations of all regressors in the convolutional models used in the whole brain analysis. Our estimates of the current engagement (CE) and general engagement (GE) (Fig 4.2) and vigour (this Fig) are not correlated and thus capture different aspects of motivation. We also show the state regressors split up by the past and future state, which can be obtained by using only half of the kernels (i.e. the half directed towards the past or future). By splitting up the state regressors this way we can combine them again on the contrast level to obtain the overall motivational states. We can, however, also subtract them on the contrast level to examine potential differences between past and future levels of engagement/vigour.

concluded and potential preparatory activity for the coming trial was beginning while also ensuring that the measurement was taken in the same way across all tasks; the same two time points could be defined in an identical manner for all four tasks. Moreover, the previous-trial-end and the following decision-prompt are far enough apart in time to ensure that regressors time-locked to each event are relatively uncorrelated even after convolution with the macaque's fast hemodynamic response function [238, 239] (Fig 4.2F). We hypothesised that general motivation-related activity – our signals of interest – should be found at both time points. In our analysis we, therefore, included regressors for both CE and GE at both time-points, and use contrasts to also estimate OE and ES. Moreover, we also included our analogous control estimates of the trial vigour level and the state vigour at both of the same time-points (Fig 4.3). Importantly, as we can only estimate *trial vigour* and *state vigour* on engaged trials, these regressors are zeroed out on disengaged trials.

We combined the results of these session-level regressions separately for each data set per animal using fixed effects. In a final step, we combined the data from all data sets on a third level using random effects. This allows us to examine the neural correlates of task-independent engagement across tasks and animals. To examine the effects of engagement/disengagement we used eleven data sets from nine animals across four tasks while controlling for vigour. Statistical significance was determined using a standard cluster-based thresholding criteria of $z > 2.3$ and $p < 0.05$ [153]. Significant clusters for our contrasts of interest are shown as white outlines in Fig 4.4. Additionally, we also show the non-cluster-corrected z-statistics at a lower threshold of $z > 1.5$ in Fig 4.4 to give a more complete picture of the results. Moreover, in the supplementary analyses we report analyses for vigour-related effects using a larger sample of data from thirteen animals across three tasks, as discussed above.

When we examined neural activity related to CE (Fig 4.4A), we saw a large overlap between activity at previous-trial-end (Fig 4.4A left) and decision-prompt (Fig 4.4A middle), with activity at decision-prompt being slightly more lateral. Combining these estimates allowed us to examine regions that show activity both at previous-trial-end and decision-prompt (Fig 4.4A right), which suggests that

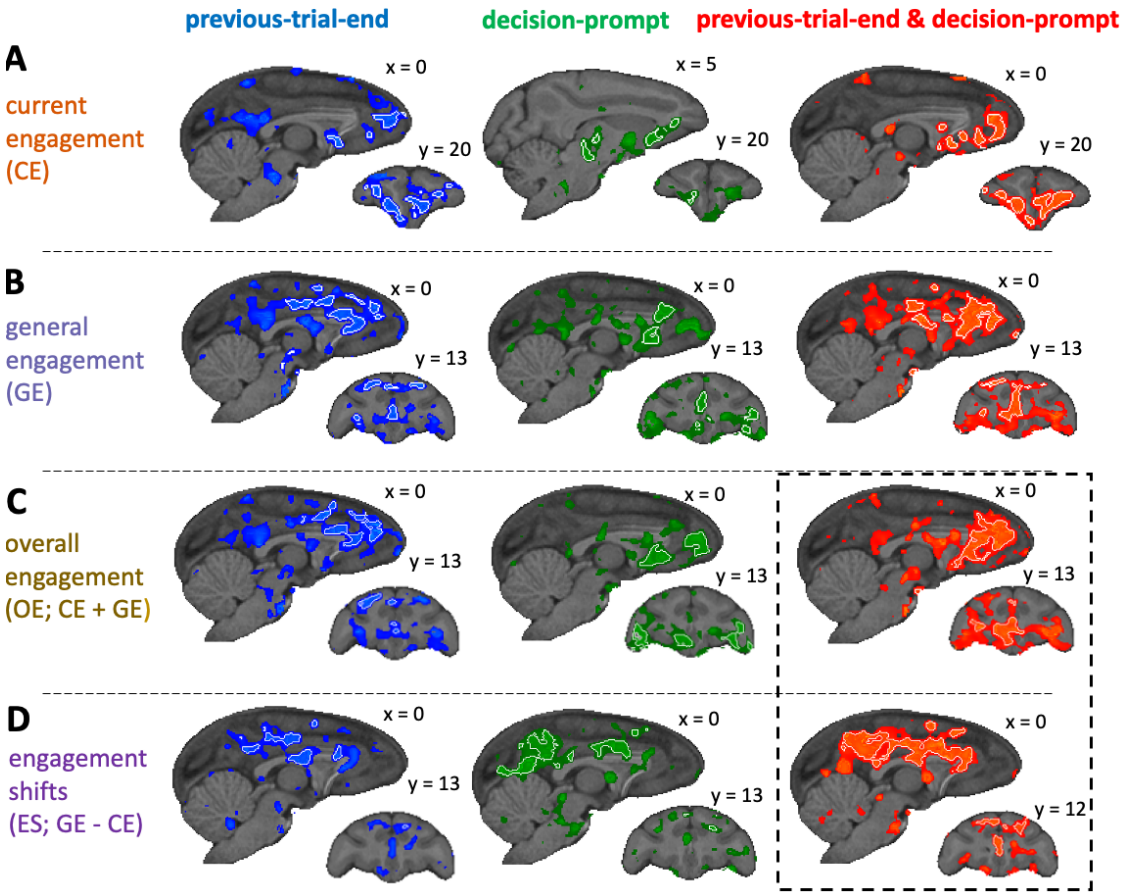


Figure 4.4: **Neural activity associated with engagement.** Whole-brain activity is shown for different contrasts (top to bottom), time-locked to different events (left to right). Activity with $z > 1.5$ shown superimposed, with white outlines indicating significant clusters at $z > 2.3$. (**A**) For CE we observed activity in regions spanning pgACC (area 32), sgACC (area 25), and OFC (areas 12 and 13), both at previous-trial-end and decision-prompt and when looking at both time-points combined. (**B**) For GE we observe activity throughout anterior and mid cingulate gyrus (including pgACC and supracallosal gACC), and frontopolar cortex. (**C**) For OE we observed activation most prominently in pgACC but extending into adjacent sgACC and dACC, and also OFC areas 13 and 47/12o when animals are engaging with the current trial while also being in an overall engaged state. (**D**) For ES, we observed activity in the supracallosal cingulate cortex (including supracallosal gACC) when animals disengaged from the trial despite otherwise being in an engaged state.

it is sustained throughout this task period and not linked to any particular task event (Fig 4.2E). While there was widespread activity in the brain, within frontal cortex, pgACC (area 32), ventromedial PFC (areas 25 and 14), and the larger orbitofrontal network (areas 12 and 13) were particularly active. For a full table of cluster locations and descriptions see Table C.1.

Similarly, when we examined neural activity related to GE (Fig 4.4B), we again saw a large overlap between activity at previous-trial-end (Fig 4.4B left) and decision-prompt (Fig 4.4B middle). Combining both time-points again yielded regions that show sustained activity (Fig 4.4B right). While the activity again included pgACC (area 32) prominently, there was somewhat less ventromedial PFC and OFC activity and instead more focus on anterior supracallosal ACC gyrus (gACC; area 24) as well anterior dorsal ACC sulcus. Moreover, we found a significant cluster in frontopolar cortex (area 10o). For a full table of cluster locations and descriptions see Table C.2.

To identify regions that were active when the animals had a high overall motivational level, we combined our estimates of CE and GE into OE (Fig 4.4C). At the end of the previous trial, activity was prominent in pgACC (area 32) and extended caudally into gACC (area 24) and into dorsal ACC sulcus (rostral cingulate zone) (Fig 2C left). At decision-prompt, activity was again seen in pgACC (area 32), but otherwise more orbitofrontal (area 47/12o) (Fig 4.4C middle). When combining activity at previous-trial-end and decision-prompt to find areas that were active throughout the whole task-period and across CE and GE, we observed a prominent and extensive area centred on pgACC (area 32), but extending into adjacent dorsal ACC sulcus (dACC; note that this area is sometimes refer to as mid-cingulate cortex or rostral cingulate zone) and subgenual ACC (sgACC; area 25) and also, albeit to a more limited extent in orbitofrontal cortex (OFC) in area 13 and the sub-region bordering ventrolateral prefrontal cortex – 47/12o – , and striatum (Fig 4.4C right). For a full table of cluster locations and descriptions see Table C.2.

We also looked for effects of ES, i.e. the difference between GE and CE (Fig 4.4D). Such activity was prominent when animals disengaged on the current trial while otherwise having been in an engaged state and likely to soon return again to

an engaged state. In other words, the analysis identifies 'surprising' disengagements, where the disengagement is not preceded or followed by other disengagements. Again, similar regions were active when time-locking to previous-trial-end (Fig 4.4D left) and decision-prompt (Fig 4.4D middle). When we time-locked to both previous-trial-end and decision-prompt, activity was prominent throughout mid, supracallosal cingulate gyrus (area 24) (Fig 4.4D right) extending into poster cingulate cortex and the precunous. For a full table of cluster locations and descriptions see Table C.2.

Overall, while we saw some small differences between the focus of activation between previous trial end and decision prompt, none of the frontal effects were statistically different in a comparison between the two. All statistically significant differences we found were in more posterior parts of the brain, suggesting that the frontal circuit activity carrying motivational information is particularly sustained.

To further examine the factors driving engagement on the whole-brain level, we focused on activity that was present both at previous-trial-end and decision prompt (Fig 4.4 right column) as this activity is most likely due to sustained motivation. There we focused on OE-related and ES related activity (Fig 4.4 dotted lines), and extracted the BOLD time course from regions of interest (ROIs) we placed in grey matter within the areas of functional activity. Specifically, we defined the ROIs as the overlap between functional activity and anatomically defined regions [154], and looked separately at the effects of CE and GE in the timecourse.

We observed that activity related to CE and GE appears similar in pgACC, OFC and the striatum (Fig 4.5A-C middle rows). Activity related to GE extended over a window of approximately 30s – approximately 15s before and 15s after the current trial. In contrast, activity related to current CE level was prominent before and on the trial itself. However, activity tracking both the more phasic current CE level and the more tonic GE was observed across all areas in which OE effects were observed, namely pgACC (area 32), OFC (area 13), and striatum (Fig 4.5A-C). Finally, to confirm that OE effects in each region were not driven by activity recorded just in one task, we extracted the t-statistics in these ROIs from the whole-brain analysis and examined them for differences by task (Fig 4.5A-C bottom rows). Effects in

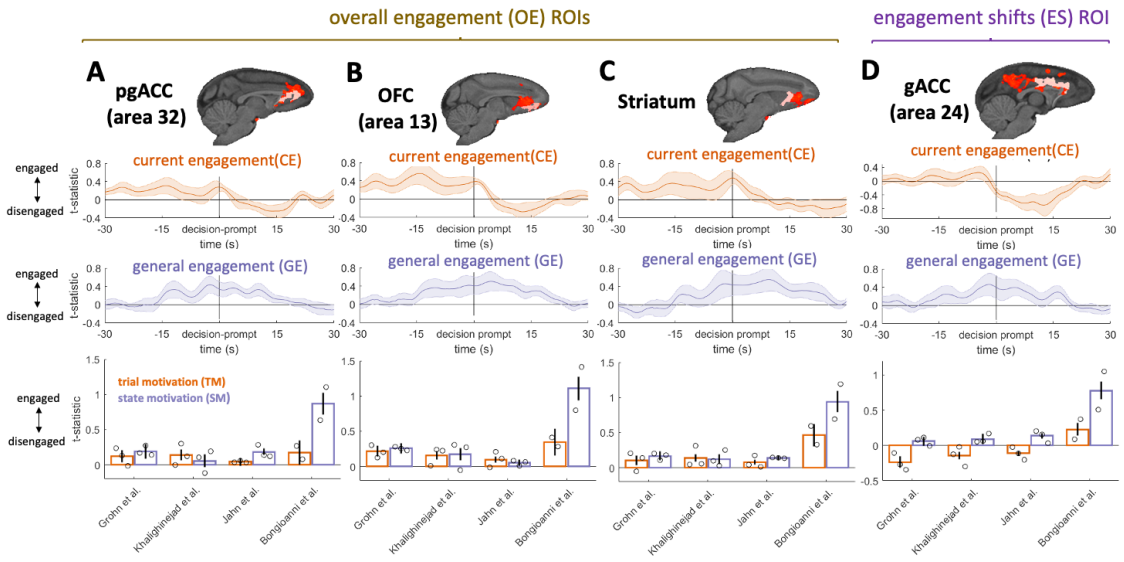


Figure 4.5: **Current engagement and general engagement timecourses in ROIs.** We extracted timecourses from ROIs placed in anatomically defined regions within our significant OE and ES clusters for activity both at previous-trial-end and decision-prompt. Significant clusters are shown in red with ROIs shown in light red (top). We then visualised the CE and GE timecourses in these regions time-locked to decision-prompt (middle rows). Shaded error bars represent standard errors of the mean across sessions. We also extracted the t-statistics associated with CE and GE from our whole-brain analysis in the same ROIs to visualise effects for each task separately (bottom row). Bars represent task-means and small dots represent individual animal means. (**A-C**) Extracted CE timecourses from pgACC, OFC, and striatum show sustained activity before and during the trial. By contrast, GE timecourses show sustained activity both before and after the trial. Effects are consistent across all four tasks (bottom). (**D**) Extracted CE timecourses from supracallosal gACC are sustained decreased during and after the current trial (i.e. disengaged), while GE timecourses are sustained increased before and after the current trial (i.e. engaged). Effects are consistent across three of the four tasks, with CE having the opposite (positive) sign in the fourth task (bottom).

the same direction were present in all four tasks and ROIs, although they were especially prominent in a task that required animals to make novel decisions [76].

Extracting the timecourse from the gACC ROI placed within the significant ES cluster (Fig 4.5D) demonstrated that there was both a decrease in activity that was related to CE – an effect that began shortly before trial onset but which was then sustained for some time afterwards – and an increase in activity related to GE (Fig 3D middle). To confirm that the effect was not driven by any one particular task, we extracted the t-statistics in the ROIs identified by the whole-brain analysis and

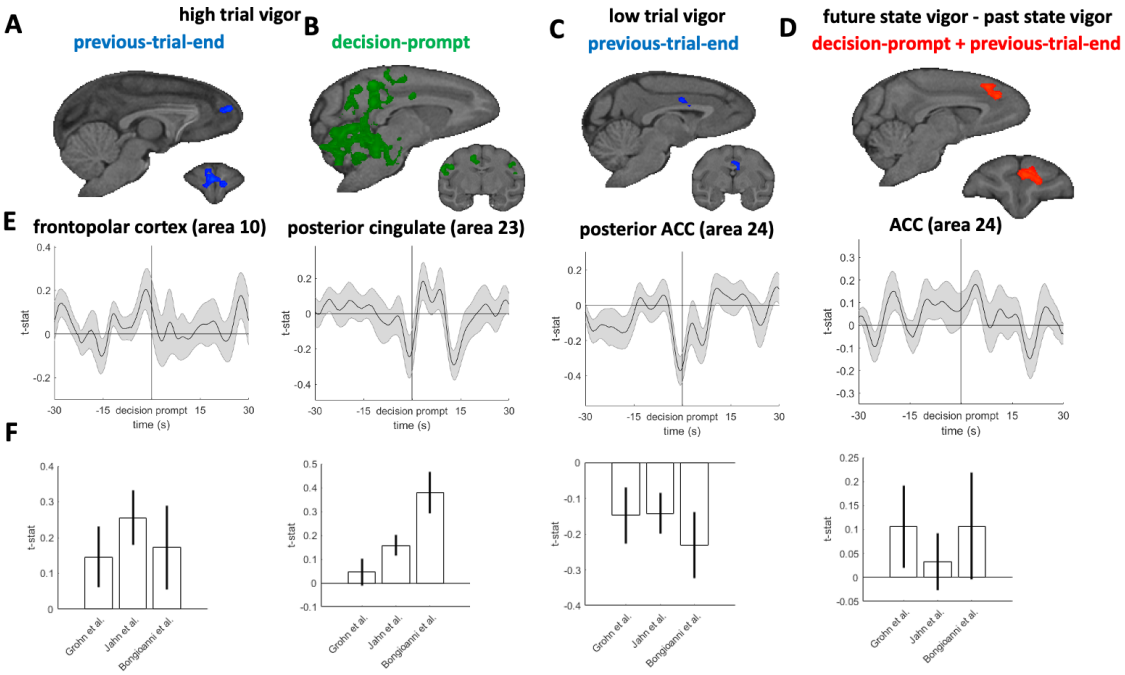


Figure 4.6: **Transient effects of vigour.** (**A**). At the end of the previous trial we find activity in the frotopolar cortex when the current trial has high trial vigour. (**B**) However, at decision-prompt of the trial itself high trial vigour is associated with activity in motoric areas in posterior cingulate and cerebellum. (**C**) Low vigour trials are preceded by activity in posterior ACC. (**D**) We also observed activity in ACC/preSMA when animals have higher state vigour in the future than in the past. (**E**) To examine how extended these effects are we visualise their timecourses in ROIs we placed. Unlike the more sustained effects of motivation as indexed by engagements/disengagements (Figs 4.4 and 4.5), effects of vigour are more contained to the current trial. (**F**) Splitting the effects within the ROIs up by task confirms that the signs of the effects are consistent across different tasks.

examined them by task. We found broadly similar effects in three tasks although the current CE effect was different in the fourth task (Fig 4.5E). The ES contrast also clearly revealed posterior cingulate cortex and precuneus, a region that has previously been implicated in decisions to disengage with foraging [240].

Finally, we note that these results were specific to task engagement/disengagement as opposed to response vigour: when we looked at the latter, we were unable to see similar patterns of neural activity to those shown in Figs 4.4 and 4.5 (see Figs 4.6 and 4.7 for vigour results). If anything, vigour activity was weaker overall and more transiently related to either decision prompt or after end trial. However, we found a small cluster of activity related to a future relative increase in vigour (Fig 4.7).

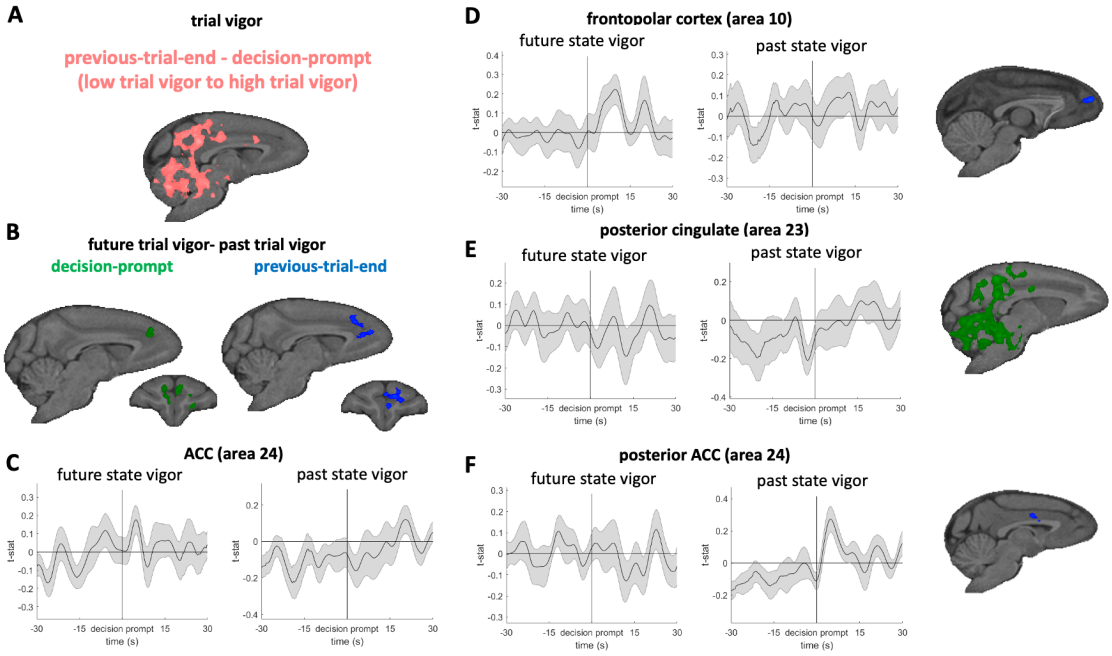


Figure 4.7: **Transient vigour effects and future and past vigour timecourses.** (**A**) To examine whether effects of trial vigour were transiently different between previous-trial-end and decision-prompt we examined a contrast of the difference in activity between the two time-points in our tasks. The whole-brain results show the same regions that were found when examining the previous trial end and decision-prompt separately (Fig 4.6BC). (**B**) When splitting up the activity we found relating to a higher future state vigour than past state vigour (Fig 4.6D) by decision-prompt and previous-trial-end, we observe the same regions. (**C**) Splitting up the timecourse extracted by future and past state vigour shows that activity related to future state vigour is above baseline, while activity for past state vigour is below baseline. (**D-F**) We also examined the activity in all other vigour-related ROIs for effects of past/future state vigour. (**D**) In frontopolar cortex we found increased activity after decision-prompt if animals had more vigour in the future. (**E**) In posterior cingulate we found activity preceding decision-prompt if the animals previously were in a low-vigour state. (**F**) In posterior ACC we found increased activity after decision-prompt if the animals previously were in a high-vigour state.

4.2.3 TUS results

Our fMRI analysis identified OE activity in pgACC (Fig 4.4C). A study we used in the fMRI analysis also causally manipulated activity in pgACC using transcranial ultrasound stimulation (TUS) [148] (Fig 4.8A). Thus, we next sought a causal test of pgACC's importance for task engagement. In addition to examining pgACC TUS data, we were also able to examine the impact of TUS in other regions: in the

dataset, BF and POp, were also stimulated, and it also include a sham condition [148]. BF is a useful control region because BF activity is associated with the timing of individual actions and BF TUS and cholinergic manipulation (BF is a source of many cholinergic projects) have been shown to alter the timing of individual actions [148, 237], potentially linking it to motivation by affecting response times. By contrast, POp was not associated with general task engagement/disengagement nor with performance of the specific task and so POp TUS acted as a general control for cortical stimulation. The TUS wave frequency was set to 250 kHz resonance frequency. TUS was applied in 30 ms bursts that were generated every 100 ms for a total period of 40 s. The procedure was then immediately repeated for another 40 s in the same area but in the other hemisphere. All TUS was applied prior to the behavioural task. Sustained TUS trains have previously been shown to exert an impact a sustained subsequent impact on neural activity and behaviour and therefore make it possible to examine the effect of neural disruption in the absence of any concomitant auditory effects that might be associated with the delivery of each TUS pulse [58, 76, 205, 241, 242].

Because the task during TUS had a strict response deadline [148], we only classified trials as disengaged if the animals failed to respond altogether during a trial, just as we did for the previous analyses. To examine the effect TUS had on the time spent disengaged, we classified each timepoint in each session as engaged or disengaged, calculated the total time disengaged per session, and averaged this measure over sessions. We observed that animals disengaged less overall after pgACC stimulation than during the control conditions ($\chi^2\,(1) = 99.01$; $p < 0.001$; Fig 4.8B).

To further examine this effect, we next calculated the time spent disengaged for each stimulation site as animals progressed through the session. This analysis showed that the significant difference was primarily driven by early disengagements being more frequent in the control conditions than after pgACC stimulation, whereas late disengagements were equally common throughout all conditions (Fig 4.8C). Indeed, when repeating the analysis in Fig 4.8B for the first 20 min of a session we observe a significant effect of stimulation condition ($\chi^2\,(1) = 301.31$; $p < 0.001$)

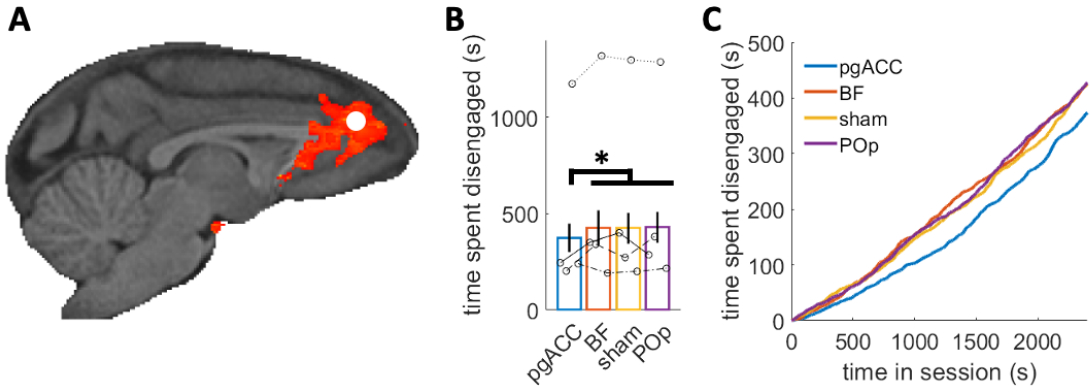


Figure 4.8: **TUS effects on disengagement.** (**A**) One of the tasks [148]. used in our analysis also causally manipulated activity in pgACC using TUS. The stimulation site is shown as a white circle superimposed on the significant OE cluster. The dataset also stimulated at two other regions which we use as controls (not shown), and also included a sham condition. (**B**) The total time spent disengage during the task was significantly lower after pgACC stimulation than when one of the control regions was stimulated, or in the sham condition. Bar represent condition means, small dots represent individual subject means, and error bars represent SEM. After pgACC stimulation animals spend on average 373.6 s disengaged during a session, while they spent 425.0 s, 424.8 s, and 428.0 s disengaged after BF, sham, or POp stimulation respectively. (**C**) The total time spent disengaged by time in the experiment reveals that after pgACC stimulation, animals are more engaged early on during the task, whereas during the latter half of sessions animals were equally engaged regardless of stimulation site.

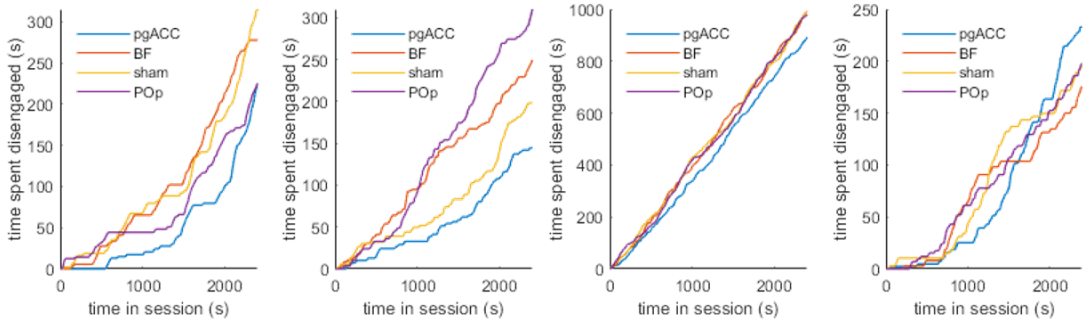


Figure 4.9: **Time spent engaged after TUS split up by animal.** The same analysis as in Fig 4.8C but split up for each of the four animals. The effect of more engagement early on can be seen in each animal.

but the effect is not significant when only analysing the last 20 min of a session ($\chi^2\,(1) = 1.48$; $p = 0.224$). This effect of more engagement early on can also be observed in each animal individually (Fig 4.9).

4.3 Discussion

Motivation fluctuates throughout daily activity leading to inattention and changes in response vigour. Ultimately, however, people and animals may give up on a task completely and either remain inactive or pursue an entirely different course of behaviour. While the process of error monitoring and subsequent adjustment of behaviour has received considerable attention [243–245], less is known about the processes that drive complete task disengagement. This is despite the obvious relevance such mechanisms have to the understanding of apathy – a prominent feature of psychological and neurological illnesses [216]. Although the social demands of the research setting mean that human participants rarely give up on a task completely when they are participating in an experiment, it is not unusual for macaques to move between periods of task disengagement and then re-engagement. In the current investigation we identified such periods and found that they manifested in similar ways in the behaviour across eleven macaques performing four different cognitive tasks in the MRI scanner. When animals were strongly engaged in any task and unlikely to disengage, then a broad region of increased activity spanning several areas, but which was especially prominent in pgACC, was found. Activity was weakest on trials when the animals' motivation collapsed and they disengaged from performance. The effects were apparent even when we controlled for RT suggesting that pgACC activity was related to task disengagement rather than any change in response timing [148], response control [243, 246], or any change in response vigour that might lead to changes in RT [220].

That pgACC was linked to engagement across tasks suggests that it is concerned with the intrinsic motivation to perform any task rather than with a specific feature of a particular task. This conclusion is further strengthened by the observation that the link between pgACC activity and task engagement was found after regressing out any influence that specific task events might have had on neural activity. In fact, for all analyses, we extensively regressed out task parameters to remove all the variance linked to task features and reward history, so that we were able

to examine how fluctuations in the residual, task feature-unrelated activity was linked to fluctuations in engagement.

In all tasks included in the analysis, we could distinguish between activity related to task engagement on a given trial (current engagement; CE) – whether the animal was engaged or disengaged on the trial itself – and the more general state (general engagement; GE) surrounding the trial. Moreover, activity change was not just apparent at the time of responding but it was present and built up over a longer preceding time period. Timecourse analyses revealed elevated signals approximately 30 s before and 30 s after the trial in question. The slowly evolving pgACC signal might reflect the parallel slow evolution of motivational factors and their independence of specific task events.

Importantly, the corresponding pgACC region of the human brain [161] has been linked to the predisposition to initiate foraging behaviour and in determining that the prospect of potential future outcomes mean that it is worth initiating a sequence of behaviour despite potential costs [60, 183]. The pgACC is unusual in that it is one of only two cortical regions that projects strongly to the striosomal compartment of the basal ganglia, in anterior striatum, which in turn is distinguished by a number of anatomical features including projection to the dopaminergic midbrain [224, 247, 248]. As a result, pgACC is well placed to regulate fundamental aspects of motivated behaviour under the control of dopamine [249].

Not only was activity in pgACC predictive of task engagement but TUS-induced alteration of pgACC activity led to consistent patterns of changed task engagement in the four macaques that participated in an additional TUS study. After the application of TUS, macaques were less likely to disengage from a task. Normally, when animals were in the control condition, in the first half of a 20 min testing session, macaques disengaged from the task for approximately 3 min. After pgACC TUS, however, animals often worked continually without disengaging or only took a break for approximately 2 min on average. Importantly, the effects were specific to pgACC TUS and were not observed after TUS to two control brain regions. First, similar effects not seen when applying TUS to an anterior parietal control

region, POp, in which there was no task-related or task engagement-related activity. Second, and perhaps even more importantly, such effects were not seen when TUS was applied to the cholinergic basal forebrain (BF) even though it has previously been shown that BF TUS and systemic cholinergic manipulation change the timing of animals' decisions to act [148, 237].

Relatively few behavioural experiments have focused on the macaque pgACC and previous behavioural analysis approaches have not allowed identification of clear changes in task engagement [148] of the sort that we were able to identify here. Moreover the effects we report here were stronger in the first half of the data set recorded after TUS. However, it has been reported that electrical microstimulation of the macaque pgACC during a cost/benefit decision making task led to fewer decisions to pay higher costs (enduring air puffs) to obtain higher rewards (more juice) [88]. If pgACC is not only responsible for setting the general willingness to endure costs for benefits during choices but also responsible for setting the general level of engagement, then our results and these previous findings can be reconciled. However, it is important to note that TUS is unlikely to recreate patterned excitation of specific neurons that can be induced by microstimulation but rather it may be more likely to disrupt the endogenous activity patterns within a brain region [205, 242]. In the rat, optogenetic inhibition of the projections from the homologue of pgACC [250] – often called the prelimbic cortex – to the striosome similarly leads rats to be more likely to pay the cost of engaging in a trial in order to obtain a reward [251]. This occurs because pgACC outputs synapse with inhibitory interneurons in the striosome which, in turn, connect with striatal projection neurons. Thus, disrupting pgACC leads to the release of striatal projection neurons from inhibition. As noted, striosomal projection neurons are distinguished by their unique anatomical connections to regions such as the dopaminergic midbrain. In summary, pgACC TUS or pgACC optogenetically mediated inhibition in monkeys and rats respectively make animals more likely to engage in an effortful task to obtain reward or to take a costly action to obtain reward. Both interventions may resemble one another in leading to the release of striatal projection neurons from inhibition.

The pgACC region studied here not only has a homologue in rodents but also in humans [161, 250]. In humans, coupling between pgACC activity and striatal activity has been linked to disinhibition of effortful choices; first, it was more prominent when the costs of a course of action were high but it was still pursued and second it was more prominent in individuals who were inclined to pursue such courses of action [183]. Individual variation in pgACC activity has also been reported to covary with how influenced each person is by the prospect of future reward despite the need to engage in a sequence of decisions [60]. It also tracks how well people have been performing simple tasks and how they are likely to evaluate their performance [252, 253].

The idea that animals make decisions to engage or disengage with one behaviour or another or simply to do nothing at all is consistent with a growing body of work on foraging decision making and their neural correlates [60, 254]. It also suggests alternative ways of thinking about situations in which people and animals appear to lack motivation. In particular, the engagement shift (ES) activity in supracallosal cingulate gyrus (area 24, also called mid-cingulate cortex) might normally, in less constrained situations than in the current experiment, lead to sudden deliberate decisions to disengage, rather than simply reflecting slowly waning motivation. While this is only speculation, it is nonetheless noteworthy that ES specifically activated the supracallosal cingulate gyrus in a region adjacent to one that has been linked to switching and foraging activity in the past in humans, macaques, and rodents [39, 58–60, 183, 190, 254] and which is distinct from pgACC. Overall, our results suggest slowly drifting motivational fluctuation where low pgACC activity is linked to low motivational states and repeated giving up, while sudden choices to give up surprisingly during high motivational state are triggered by sudden supracallosal gACC activity. The engagement shift (ES) was also the only contrast that clearly revealed posterior cingulate cortex and precuneus, a region that has previously been implicated in decisions to disengage with foraging [240], further suggesting that ES might be linked to deliberative actions to specifically disengage in a trial, rather than a slow collapsing of overall motivation.

Overall, our findings not only suggest a pgACC mechanism for intrinsic motivation but also emphasise the multifaceted nature of motivation itself. Specifically, we could dissociate the motivation to engage versus give up with the task at hand, from the motivation of how vigorously to continue working while engaged. However, our ES index suggests that even giving up might not be solely determined by one motivational factor. In fact, in our study, animals might give up because of an overall change in intrinsic motivation (OE) or because they deliberately, but transiently, want to do something else (ES). We suggest that future work should embrace this complexity and instead of looking for one singular source for motivational processes distinguish between different types of motivation and their multiple links to behaviour.

Importantly, engagement-related activity was not confined to pgACC but was also noticeable in a posterior part of the lateral orbitofrontal sulcus. This region has been identified with credit assignment – the linking of specific choices to specific outcomes [205, 238] – but it is also notable that cortex in the same region or nearby is the second cortical region in the macaque, in addition to pgACC, that projects to the striosome and in which stimulation is known to affect cost-benefit decision making [224, 247].

While the current study has taken some of the first steps needed to identify the neural mechanisms mediating the impact of generalised intrinsic motivational factors on engagement in complex cognitive and behavioural processes, some questions remain unanswered. Notably while pgACC and posterior lateral orbitofrontal sulcus were less active when motivational collapse and task disengagement occurred, a more posterior mid-cingulate gyrus region (area 24) was most active during collapse (Fig 4.5). As well as attempting to understanding the key elements that determine the multifaceted relationships between specific task features, types of motivation, brain activity and the cellular mechanisms at play in pgACC and beyond, future important steps should examine the effect of stimulation in area 24.

4.4 Materials and Methods

4.4.1 Subjects

13 rhesus macaques across 17 data sets were included in the four studies considered. All procedures were conducted under licenses from the United Kingdom (UK) Home Office in accordance with the UK Animals (Scientific Procedures) Act 1986 and with the European Union guidelines (EU Directive 2010/63/EU).

4.4.2 Data collection

The fMRI data were acquired in a horizontal 3 Tesla MRI scanner with a full-size bore using a four-channel, phased-array, receive-only radio-frequency coil in conjunction with a local transmission coil (Windmiller Kolster Inc, Fresno, USA). The animals were head-fixed in a sphinx position in an MRI-compatible chair (Rogue Research, CA). fMRI data were acquired using a gradient-echo T2* echo planar imaging (EPI) sequence with the following parameters: 1.5 $\times$ 1.5 $\times$ 1.5 mm resolution, 36 axial interleaved slices with no gap, TR of 2280 ms, TE of 30 ms and 130 volumes per run. Proton-density-weighted images using a gradient-refocused echo (GRE) sequence (TR = 10 ms, TE = 2.52 ms) were acquired as reference for offline image reconstruction.

4.4.3 Behavioural task-models

We used data from four different tasks [76, 148, 169, 236, 237], including the tasks discussed in Chapters 2 and 3. Details of each task can be found in the original publications or chapter of this thesis. In all tasks monkeys had to respond to stimuli on screen that were rewarded, while their neural activity was recorded using fMRI. Briefly, in the task described in Chapter 3 [236], an exploration-exploitation task with different time horizons was used. On some trials, monkeys had to make one-off choices between two stimuli on screen based on the information presented. On other trials, they had to choose between the same options repeatedly, which enabled them to learn more about the value of the options. During the task described in Chapter 2, [169] a single option was presented on screen. By manipulating the

reward associated with the option, as well as the location of the option on the screen, different kinds of surprises were induced. In the study of Bongioanni and colleagues [76], monkeys had to choose between two options that varied among two dimensions, reward amount and reward probability. They presented the monkeys with novel stimuli that they had not encountered before but the value of which they should be able to infer based on previously observed stimuli. Khalighinejad and colleagues [148, 237] showed monkeys a single stimulus that contained information about the reward amount and the inter-trial-interval length. The longer monkeys waited to respond, the more the reward probability increased, which was also displayed as a feature of the stimulus, but at the price of loosing time as the experiment did not have a fixed number of trials but was limited to 40 min. This allowed them to study how monkeys decide when to make a response.

To regress out the effect task-manipulations have on engagements/disengagements, we used logistic regressions to control for these effects. For all tasks, we included regressors for the rewards the animals obtained on the previous 5 trials, the current trial number, and the cumulative reward the animals received so far during a session. Additionally, we included task-specific regressors for each task that are based on the models used in the original analyses of the tasks:

For study 1 (Chapter 3) [236] we included a regressor coding for repetition bias (whether the animal has responded on the same side on the previous trial), a regressor coding for the choice horizon (short or long), and a regressor coding for the current choice number within a horizon. Moreover, we used the Bayesian model described by Jahn's and colleagues 25 to estimate the expected reward and the expected uncertainty on each trial. We then included the sum of the expected reward of both stimuli, and the sum of the uncertainty of both stimuli as regressors as well as the absolute difference in expected reward and uncertainty between the two stimuli. Finally, we allowed these 4 latter regressors to vary by horizon as interaction terms.

For study 2 (Chapter 2) [169] we included a regressor coding for whether the stimulus is on the left or the right side of the screen, a regressor coding for whether

the stimulus switched sides, and a regressor coding for whether the monkeys received 2 drops of juice on the last trial.

For study 3 [76] we included regressors for the absolute additive value difference, the absolute multiplicative value difference, the total additive value, and the total multiplicative value. Additive and multiplicative value here refer to adding or multiplying reward magnitude and probability (further details can be found in the original publication). Moreover, we also included a regressor capturing a repetition bias (responding on the same side as on the previous trial).

For study 4 [148, 237] we included regressors for the current reward magnitude, the length of the upcoming inter-trial-interval, and the speed of the dots on screen.

All models were run separately for each monkey. For each monkey, we allowed all regressors to also vary as random slopes by session. We then took the difference between the model prediction and observed behaviour as our measure of CE.

4.4.4 Autocorrelation and kernels

To calculate the autocorrelation of our measure of intrinsic motivation we shift the timeseries for each session of each monkey by lags from 2-10 and compute the correlation for each (we leave out lag=1 because for some of our experiments two disengagements cannot occur after each other because of the task design). We then separately average the sessions of each monkey, before finally averaging over monkeys.

To test whether the autocorrelation is significantly larger than 0, we randomly permute the data of each session and repeat the above procedure 10000 times on the permuted data. We then determine the p-value as the number of times the average autocorrelation over monkeys is smaller than the permuted average. Because we are testing lags from 2-10, we use a p-value of $0.05/9 = 0.0056$. For RTs we use a p-value of $0.05/10 = 0.005$ because we are testing lags 1-10.

To compute the motivational state, we fitted an exponential kernel to our measure of intrinsic motivation. Specifically, we found the free parameter α that minimised the squared distance between the function $\alpha(1-\alpha)^{|d|}N$ and the data, where for each trial, d indexes all past and future trials of a session, leaving out the current

trial, i.e. $d = \text{first trial}, \ldots, -2, -1, 1, 2, \ldots, \text{last trial}$, and N is a normalisation factor that makes the weights sum up to 1, i.e. $N = \sum \left(\alpha(1-\alpha)^{|d|}\right)^{-1}$. We compute α separately for each of our 4 tasks. For study 1 [236] and study 4 [148, 237] we used $d = \text{first trial}, \ldots, -2, 2, \ldots, \text{last trial}$—leaving out trials -1 and 1—because disengagements do not occur concurrently because of the task designs.

We then used the fitted value of α to smooth the data, thus obtaining a state estimate on each trial. By using only the half of the kernel that is directed towards the past/future, i.e. $d = \text{first trial}, \ldots, -2, -1$ and $d = 1, 2, \ldots, \text{last trial}$, we were also able obtain separate state estimates of the past and future GE, which we used as regressors in the whole brain analysis.

When fitting the kernel to RTs we are only using engaged trials. Therefore, the timeseries is interrupted when a disengagement happens, which also breaks the autocorrelation. For RTs we therefore only use consecutive chunks that are uninterrupted by disengagements to fit the kernel, i.e. we set $d = \text{earliest trial that is engaged}, \ldots, -2, -1, 1, 2, \ldots, \text{latest trial that is engaged}$.

4.4.5 Whole-brain analyses

EPI data were prepared for analysis following a dedicated nonhuman primate fMRI processing pipeline using tools from FSL [149], Advanced Normalization Tools (ANTs) [150], and the Magnetic Resonance Comparative Anatomy Toolbox (MrCat; https://github.com/neuroecology/MrCat).

Like for our behavioural analysis, we also created separate neural regression models for each task. Apart from these task-specific regressors (further outlined below), we also included the same regressors across-tasks. For all tasks, we included regressors for the current level of the intrinsic CE (computed as described in the behavioural task-models section), the past GE, and the future SM (computed as described in the autocorrelation and kernels section). We included all of these regressors twice, once time-locked to the end of the reward delivery of the previous trial, and once time-locked to the onset of the decision-prompt. Moreover, we also included regressors for the trial vigor, and the past and future state vigor,

again time-locked both to the end of the previous trial's reward delivery and the decision-prompt. The correlation between these 12 regressors in shown in Fig 4.3D.

To compute overall estimates of GE and state vigor, we created contrasts that summed up the past and future GE, and the past and future state vigor. Moreover, to estimate OE we added a contrast that summed up CE and GE, and to estimate ES we added a contrast that subtracted CE and GE. Similar contrasts were included for vigor. Finally, we also included contrasts that subtracted the past and future GE, and the past and future state vigor.

Additionally, we also included some control regressors that were the same for all 4 tasks. We included intercepts time-locked to the beginning of the reward delivery, the end of the reward delivery, the onset of the decision-prompt, and when decisions were made. We also included the current trial number, the cumulative reward so far, and the seconds since the beginning of the experiment, all time-locked to the end of the previous trial's reward-delivery, and to the onset of the decision prompt. Moreover, we also included confound regressors to index head motion and volumes with excessive noise.

Task-specific regressors were based on the models used in the original papers. The regressors we included were:

For study 1 (Chapter 3) [236] we included an intercept time-locked to the onset of the wait-stimulus. We also included regressors for the expected reward of the chosen stimulus, the expected reward of the unchosen stimulus, the uncertainty of the chosen stimulus, and the uncertainty of the unchosen stimulus, all time locked to the wait-stimulus. These quantities were calculated according to the Bayesian model described in the original paper. At decision, we included a regressor for the response side. At the beginning of the reward delivery, we included regressors for the amount of reward received, the expected reward of the chosen stimulus, the expected reward of the unchosen stimulus, the uncertainty of the chosen stimulus, and the uncertainty of the unchosen stimulus, all again according to the Bayesian model. Some sessions also included a horizon manipulation, such that animals had to either make one-off decisions, or decide among the same options multiple times

while learning new information about the options throughout. For these sessions, we included a regressor at the decision-prompt whether the trial was a short or a long horizon trial. Furthermore, in some sessions animals received feedback about the reward of the unchosen stimulus, whereas in others they did not. For the sessions that included this feedback, we also included a regressor for the amount of reward of the unchosen option, time-locked to reward delivery.

For study 2 (Chapter 2) [169] we included a regressors for the response side and whether the stimulus had switched sides at decision. At reward delivery, we included regressors for the current reward amount, and the reward amount of the previous 5 trials as separate regressors. We also included a regressor for whether the reward was 2 drops of juice, and a regressor for whether the previous reward was 2 drops of juice. Finally, we also included a regressor for whether the current trial was an error and no reward would be delivered, time-locked to when the reward would otherwise be delivered.

For study 3 [76] we included regressors for the absolute additive value difference, the absolute multiplicative value difference, the total additive value, and the total multiplicative value, all time-locked to decision-prompt. These regressors are further described in the original paper. We also included a regressor for the response side at decision, and a regressor for the reward amount at reward delivery.

For study 4 [148, 237] we included regressors for the current reward magnitude, the upcoming inter-trial-interval duration, and the dot-speed, all time-locked to stimulus presentation. We also included regressors for the last trial's reward amount, and the number of dots on screen when the last trial's response was made, also time-locked to stimulus presentation. At decision, we included a regressor for the number of dots currently on screen. Finally, we included a regressor for the reward amount at reward delivery.

We used a hierarchical GLM approach to combine data from monkeys and sessions: We first fitted each session individually using the appropriate regression model (as described above), and then warped the resulting statistical maps into F99 standard space. There, on a second hierarchical level, we combined data individually

for each monkey using fixed effects and pre-planned contrasts over regressors that were shared across models. Finally, on a third hierarchical level, we combined data from all monkeys using random effects, as implemented in the FLAME 1+2 procedure from FLS [149]. To test for statistical significance, we used a standard cluster-based thresholding criteria of $z > 2.3$ and $p < 0.05$ [153].

Analyses were run in FSL's fMRI Expert Analysis Tool (FEAT). Regressors were z-scored and convolved with a hemodynamic response function (HRF), which was modelled as a gamma function (lag = 3 , sd = 1.5) convolved with a boxcar function of duration 1s.

4.4.6 ROI analyses and timecourses

To define ROIs, we calculated the overlap between the cluster-corrected t-statistic map from the whole-brain analysis and anatomically defined regions based on an atlas [154], which we dilated with a kernel of 3x3x3 voxels. We then warped these ROIs into session-space using the nonlinear deformation field.

To visualise the BOLD timecourse of a regressor we re-ran the convolutional whole-brain analysis for each session of each monkey in FEAT, leaving out the 12 regressors of interest we described above but including all other task-relevant and nuisance regressors. We then extracted the average residual of this whole-brain analysis from each ROI. Next, we upsample the timecourse by a factor of 10 using spline interpolation. Because we are interested in temporally extended effects of motivation, we then smooth the upsampled timecourse with a moving average filter of 5 s.

4.4.7 TUS stimulation and analysis

TUS stimulation was conducted with a single-element ultrasound transducer (H115-MR, diameter 64 mm, Sonic Concept, Bothell, WA, USA) with region-specific coupling cones filled with degassed water and sealed with a latex membrane (Durex). The ultrasound wave frequency was set to the 250 kHz resonance frequency and 30 ms bursts of ultrasound were generated every 100 ms (duty cycle 30%) with a digital

function generator (Handyscope HS5, TiePie engineering, Sneek, the Netherlands). Overall, the stimulation lasted for 40 s. A 75-Watt amplifier (75A250A, Amplifier Research, Souderton, PA) was used to deliver the required power to the transducer. For further details see [148].

To calculate the time spent disengaged, we classified each trial in each session as engaged or disengaged in the same way we did for the data sets for the behavioural and fMRI analysis. We then calculated the total time spent disengaged for each session, and tested whether there was a significant difference between the sessions in which pgACC was stimulated or the control conditions (BF, POp, or sham stimulation). In this model we assumed that the data was Poisson distributed with a log link function, and we also included a random intercept for each animal to control for different baseline effects.

To visualise where in a session differences between conditions emerged, we also calculated the cumulative sum of the time spent disengaged for each second of each session, and then averaged this sum over sessions for each condition.

5
General discussion

Contents

This thesis focused on how primate brains actively seek out new information through strategic exploration, learn from surprising experiences and all the while keeping the necessary task engagement to successfully adapt to changing task demands.

I have identified a circuit in the OFC that both learns from unusual reward experiences when necessary, and uses hypothetical outcomes which can prevent unnecessary exploration, enabling improved future exploitation of complex environments. This was balanced by cingulate circuits that fuelled value driven exploration when necessary and more generally motivated sustained engagement across tasks. Interestingly, parts of the cingulate were also active during periodic disengagements, which could either be as a reaction to such abrupt collapse or triggering them for the purposes of change.

5.1 Linking surprise and volatility

In both Chapters 2 and 3 I showed neural activity related to surprise. In in Chapter 2 I demonstrated that rare reward events are only being monitored in OFC in learnable/volatile sessions. OFC seems to only pay attention to rare rewards if the environment they occur in is sufficiently complex and worth tracking. Unlike scalar prediction errors, however, rare reward surprise is not used to update an estimate of average value. Instead, rare reward surprise signals an unusual noteworthy event. Thus, the finding that such a mechanism is only active in volatile environments is not due to an increased need to update value but rather most likely due to an increased tendency to assign credit to unusual states. In other words, the increased volatility of the environment boosts a system that detects unusual features of the environment. Rare amounts of reward are once such—but by no means the only—possible feature.

This findings can be contrasted with classic scalar reward prediction-errors [1]. In Chapter 1 I contextualise these canonical prediction errors by describing Rescorla-Wanger learning, which links them to updating the reward expectation. However, both of the tasks used in Chapter 2 and 3 also included specific conditions during which learning was not needed. Nonetheless prediction errors signals persisted. In Chapter 2 the majority of sessions employed static reward-schedules, and we might reasonably assume is that the overtrained monkeys expect such a schedule as the default prior when starting a new session. For the horizon-task used in Chapter 3, there is no reason to update expectations after short-horizon outcomes, or after the last outcome in a series of long-horizon choices. However, in both tasks prediction error signals persistently occur even in the absence of a need to learn. Indeed, this observation is consistent with other studies that expose overtrained animals repeatedly to the same conditions: while both the average expected reward and the full distribution of rewards should be familiar to the animal based on past experience, prediction error signals are still being found [44, 45].

Multiple explanations for this are possible. It could be the case that while the reward-environments the animals encounter during the tasks are static, this is not the case for ecologically more plausible environments the animals might encounter

in the wild. Thus, the animals never cease to learn because of a strong prior for some level of volatility in the environment. This could be complemented by a more stable value representation elsewhere in the brain. Alternatively, it could be the case that the prediction error signals are not being used to update a value estimate. Instead, the mismatch signal between value and expectation could be automatic, and/or is being used to track other properties of reward, such as its distribution [45].

5.2 Ecological task designs

My failure to find rare reward surprise in stable sessions in Chapter 2 also has interesting implications for task design. A task that is too simple such as the design used during stable sessions in Chapter 2, i.e., a single option associated with three different reward amounts with fixed probabilities, might not be best to detect effects of interest as certain systems might only become active when the task is sufficiently complex. It has been suggested that tasks in the laboratory should strive for more ecological validity [255], which entails that tasks have a higher number of dimensions that an animal might potentially track. While some of these dimensions might appear unnecessary to answer the question of interest, e.g. volatility for detecting rare reward events, they ensure that the animal engages with the task, which in turn boosts the signals of interest.

While not discussed much in Chapter 2, and indeed not surviving cluster-correction on the whole brain level, I also found some evidence for scaler reward prediction error signals in the OFC for this task (Fig 2.11D), and the prediction error signal had the same reversed sign as the prediction error signals for the chosen and the unchosen option found in Chapter 3. Again, the weakness of these signals in OFC in Chapter 2 could be attributed to the simplicity of the task used in Chapter 2 compared to Chapter 3.

However, the findings in Chapter 4 provide a small caveat to this explanation. The behavioural engagement and disengagement patterns of the simple task from Chapter 2 were not noticeably different than for the other three more complex tasks (Fig 4.1). Nonetheless, it seems reasonable to assume that while the animals did

not behaviourally disengage more frequently for the simple task, the task did not require them to mentally stay engaged but could be 'solved' by simple responses to the stimulus. Examining whether the overall complexity of a task, such as the complexity of the required responses or the complexity of the reward schedules, alters basic neural signals would be an interesting area of investigation. While only serving as anecdotal evidence, I found the strongest engagement signals for arguably the most complicated out of the four tasks [76] analysed in Chapter 4 (Fig 4.5A).

5.3 Value signals in cingulate cortex

In Chapter 3 I found a difference in the signs for the expected value signals on first choices depending on whether the animals would receive complete feedback (which was associated with a positive encoding of value) or partial feedback (which was associated with a negative encoding of value) in the mid cingulate as can be seen in Fig 3.7. The location for this value difference overlaps with the surprising disengagement signals found in Chapter 4 (although the disengagement effect is much more widespread) as shown in Fig 4.5B. While the default choice of an animal is to be value-seeking, I showed in Chapter 3 that animals can suppress this tendency to become more exploratory if required. Such behaviour might be thought of as suppressing the default response of the animal to seeing a stimulus with a higher expected value, which would explain the negative sign for value in the partial feedback condition. This provides a possible link between the findings in Chapters 3 and 4: I observed increased activity in mid cingulate cortex for deviations from the default response, despite the default being different things such as seeking value (Chapter 3) or to engage with a task (Chapter 4). In particular, I found the strongest results in Chapter 4 when the disengagement occurred in an engaged state, which indicates that being engaged was indeed the default behaviour for a period of trials pre- and succeeding the disengagement. Such an interpretation frames disengagements as deliberate choices where mid cingulate cortex signals to strategically disengage from the task after having been

previously engaged, presumably allowing the animal to recuperate strength to then continue engaging with the task.

This interpretation of results also implies that mid cingulate cortex dynamically changes its activity patterns based on the task demands, as it represents value with a positive sign in the complete feedback condition. That is, in the absence of a default choice to suppress for the purpose of learning, the region reverts to supporting that default choice or at least being more active when that default is more valuable, which in Chapter 3 is to be value-seeking. Thus, the region might allocate resources to a specific behaviour as demanded by the task. While again being only very weak evidence, it is noteworthy that the negative sign for engagement was only present in three out of the four tasks analysed in Chapter 4, while the sign was positive for the most complex task [76] (Fig 4.5B). This could suggest that for this task mid cingulate gyrus was so pre-occupied with regulating some other complex behaviour that no resources were available to strategically disengage, which is why activity reverted back to signalling regular task engagement.

5.4 Future work

All the studies presented in Chapters 2 - 4 have natural follow-up experiments. The U-shaped reward distribution used in Chapter 2 would make an interesting addition to most reward-based decision-making task that investigate which features of value are encoded in the brain, such as the task by Rothenhoefer and colleagues [54] discussed in Chapter 1, or tasks investigating distributional reinforcement learning [45]. Moreover, this could be complimented by simultaneously varying other features of reward, such as its identify [51]. As with other reward-features the brain represents, such as the reward distribution [45], it also yet remains to be shown whether there is a behavioural advantage that comes with such a representation.

While the basic horizon-task [27] has been extensively studied in humans, the task-version that was presented in Chapter 3 and that also included a condition in which counterfactual feedback was presented has not been tested in humans yet. As such, it would be of interest to investigate whether human value signals also change

sign depending on the feedback type, or whether this effect is related to the specific exploration strategy used by macaques. Moreover, it remains to be seen whether macaques generally do not employ directed exploration or whether the absence of such an effect was specific to the data set described in Chapter 3.

The general engagement patterns identified in Chapter 4 could also be replicated in a purposefully designed task. For instance, to further investigate the exact nature of the engagements and disengagements, participants could either be predictably or unpredictably queried to disengage from a task. The difference between such expected and unexpected pauses in engagement could shed light whether a region is involved in strategic pauses to recuperate resources or in monitoring unplanned pauses. Other natural follow-up studies could investigate how intrinsic engagement and disengagement interact with other task factors, and whether different types of motivation beyond vigour and engagement could be behaviourally defined. Additionally, it is worth examining how neural population dynamics and neural coding changes with different levels of task engagement.

Appendices

A

FMRI cluster locations for Chapter 2

Description	Cluster index	Local maxima	Max z-stat	F99 coordinates (mm)			Cluster size (# voxels)
				X	Y	Z	
Right orofacial motor cortex, extends into primary and secondary somatosensory cortex and also extends into anterior insula	1	1	6.62	25.7	1.01	6.04	4906
		2	6.36	25.2	0	7.55	
		3	6.17	23.1	-3.52	8.55	
		4	5.99	20.6	-3.52	11.6	
		5	5.99	27.7	-3.02	8.05	
		6	5.83	23.1	-3.52	10.6	
Left orofacial motor cortex, extends into primary and secondary somatosensory cortex	2	1	5.83	-26.7	-0.503	8.55	2736
		2	5.77	-25.1	-1.51	10.6	
		3	5.68	-21.6	-2.51	11.6	
		4	5.52	-26.2	-4.02	7.04	
		5	5.39	-26.2	-3.52	9.05	
		6	5.3	-25.7	-3.02	6.54	
Right ventrolateral striatum, extending towards basal forebrain	3	1	5.13	16.1	1.01	-2.01	1845
		2	4.98	15.6	1.51	-3.02	
		3	4.89	13.6	1.01	-1.01	
		4	4.38	14.6	2.52	-1.51	
		5	4.33	17.6	-1.01	-2.01	
		6	4.32	16.1	0.503	-0.503	
Left ventrolateral striatum, extending towards basal forebrain	4	1	6.11	-12.1	3.02	-4.02	1320
		2	5.08	-13.6	1.01	-4.02	
		3	4.91	-14.1	2.52	-4.53	
		4	4.59	-13.6	0	-3.52	
		5	4.47	-15.6	1.51	-4.02	
		6	4.43	-14.6	0	-3.02	

Table A.1: Clusters, z-statistics, and coordinates for sRPE in the VOI.

Description	Cluster index	Local maxima	Max z-stat	F99 coordinates (mm)			Cluster size (# voxels)
				X	Y	Z	
Anterolateral striatum, extending into lateral orbitofrontal cortex (area 12/47o)	1	1	3.53	-13.1	6.04	2.52	1353
		2	3.48	-13.1	5.53	0.503	
		3	3.33	-16.6	9.56	5.53	
		4	3.23	-10.1	3.02	4.53	
		5	3.21	-21.1	11.1	2.52	
		6	3.2	-11.6	4.02	4.02	

Table A.2: Clusters, z-statistics, and coordinates for RRE in the VOI in changing/learnable sessions.

Description	Cluster index	Local maxima	Max z-stat	F99 coordinates (mm)			Cluster size (# voxels)
				X	Y	Z	
Primary motor cortex, extends into supplementary motor area	1	1	4.91	0	-8.55	22.1	7237
		2	4.65	-13.1	-11.6	19.6	
		3	4.64	-1.01	-8.55	21.6	
		4	4.58	0.503	-11.6	21.6	
		5	4.48	0	-10.6	22.6	
		6	4.38	-14.6	-10.1	19.6	
Lateral prefrontal cortex (area 46, area 12/47)	2	1	4.91	-20.6	11.6	11.6	1464
		2	4.85	-21.6	13.6	11.6	
		3	4.57	-18.1	14.1	13.1	
		4	4.29	-19.6	13.6	11.6	
		5	4.2	-21.6	13.6	12.6	
		6	4.2	-20.1	10.1	12.1	
Insula and secondary somatosensory cortex	3	1	4.38	17.1	-9.56	7.55	1432
		2	4.37	16.6	-8.55	7.55	
		3	4.14	18.1	-12.6	8.55	
		4	4.05	23.1	-9.56	9.05	
Posterior cingulate cortex (area 23)	4	1	4.01	4.02	-14.1	13.6	983
		2	3.96	4.02	-12.6	14.1	
		3	3.72	2.01	-12.6	14.1	
		4	3.61	2.01	-14.1	11.1	
		5	3.56	2.01	-14.6	10.1	
		6	3.04	4.53	-6.04	16.1	

Table A.3: Clusters, z-statistics, and coordinates for VS in the VOI.

Description	Cluster index	Local maxima	Max z-stat	F99 coordinates (mm)			Cluster size (# voxels)
				X	Y	Z	
Anterolateral striatum extending into the anterior insula/posterior OFC	1	1	3.53	-15.6	5.53	-1.01	764
		2	3.33	-16.1	5.53	-4.02	
		3	3.07	-10.6	3.02	3.52	
		4	3.03	-11.1	2.52	2.52	
		5	3.02	-13.1	4.02	3.02	
		6	2.83	-9.05	4.53	4.02	

Table A.4: Clusters, z-statistics, and coordinates for RRE in the VOI for the contrast comparing changing/learnable and equiprobable sessions.

Description	Cluster index	Local maxima	Max z-stat	F99 coordinates (mm)			Cluster size (# voxels)
				X	Y	Z	
Brainstem	1	1	6.72	1.01	-17.1	-20.1	7238
		2	6.52	2.52	-15.6	-19.1	
		3	6.42	1.51	-24.6	-22.1	
		4	6.1	5.03	-25.7	-18.6	
		5	6.07	1.51	-23.6	-23.6	
		6	6.06	1.01	-19.6	-18.1	
Right orofacial motor cortex, extends into primary and secondary somatosensory cortex and also extends into anterior insula	2	1	6.83	25.7	1.01	6.04	5256
		2	6.08	26.2	-2.01	6.54	
		3	5.93	23.1	-2.51	11.1	
		4	5.9	21.1	-3.02	10.1	
		5	5.7	26.7	-3.52	6.54	
		6	5.69	27.2	-2.51	6.54	
Dorsal occipital	3	1	4.36	16.6	-40.2	16.6	3151
		2	4.21	16.1	-40.2	17.6	
		3	4.19	17.6	-39.7	16.1	
		4	4.15	14.6	-40.2	17.6	
		5	4.09	19.1	-37.2	15.6	
		6	3.94	20.6	-41.7	7.04	
Left orofacial motor cortex, extends into primary and secondary somatosensory cortex and also extends into anterior insula	4	1	6.51	-25.7	-2.51	9.56	2999
		2	6.37	-22.1	-3.02	10.6	
		3	5.94	-21.6	-3.52	9.56	
		4	5.66	-26.7	-3.02	7.04	
		5	5.65	-24.6	-4.53	7.04	
		6	5.62	-23.1	-3.02	11.1	
Left ventral striatum extends into basal forebrain and insula	5	1	6.59	-13.1	2.52	-4.02	2799
		2	5.3	-14.1	1.01	-3.52	
		3	4.98	-12.1	2.52	-3.02	
		4	4.83	-11.1	2.01	-5.53	
		5	4.74	-15.1	0.5	-3.52	
		6	4.59	-15.6	-1.01	-3.02	
Left dopaminergic midbrain extending into thalamus	6	1	4.89	-5.03	-10.6	-2.51	2234
		2	4.42	-7.04	-12.1	-7.04	
		3	4.31	-4.53	-10.6	-4.53	
		4	4.3	-7.04	-11.1	-7.04	
		5	4.28	-5.03	-17.1	-0.5	
		6	4.21	-7.04	-11.1	-5.03	
Right ventral striatum extends into basal forebrain and insula	7	1	6.02	15.6	1.01	-2.01	2058
		2	4.48	15.1	1.51	-0.5	
		3	4.46	13.1	1.51	-1.01	
		4	4.39	13.1	2.01	-3.02	
		5	4.3	17.1	-0.5	-1.01	
		6	4.28	16.1	-2.01	0	
Right dopaminergic midbrain extending into LGN and hippocampus	8	1	4.17	12.1	-14.6	-5.53	1410
		2	4.12	10.1	-14.1	-7.54	
		3	3.99	13.1	-18.1	-4.02	
		4	3.75	11.6	-16.1	-6.04	
		5	3.66	13.1	-15.1	-4.02	
		6	3.42	14.1	-18.6	-1.01	
Parieto-occipital region / superior parietal lobule	9	1	5.22	-2.51	-35.2	20.1	1240
		2	5.15	-1.51	-34.7	19.6	
		3	5.03	0	-35.7	19.1	
		4	4.74	-2.01	-36.2	19.1	
		5	4.55	0	-35.2	20.1	
		6	4.29	-0.5	-35.7	18.1	
Thalamus	10	1	3.75	5.03	-11.1	1.01	1212
		2	3.61	0	-5.53	5.03	
		3	3.31	5.53	-9.05	-5.53	
		4	3.29	0.503	-5.53	3.52	
		5	3.24	-4.02	-8.55	7.55	
		6	3.19	2.01	-9.05	3.02	
Medial occipital lobe	11	1	4.02	-3.52	-44.3	-3.52	1192
		2	3.86	-5.53	-40.2	-4.02	
		3	3.7	-8.05	-41.2	-5.53	
		4	3.52	-4.53	-41.2	-5.03	
		5	3.51	-6.54	-40.2	-5.03	
		6	3.48	-0.5	-46.3	-3.52	

Table A.5: Clusters, z-statistics, and coordinates for sRPE in the whole-brain analysis.

Description	Cluster index	Local maxima	Max z-stat	F99 coordinates (mm)			Cluster size (# voxels)
				X	Y	Z	
Anterolateral striatum, extending into lateral orbitofrontal cortex (area 12/47o)	1	1	3.63	-13.1	6.04	2.01	1472
		2	3.55	-13.1	6.04	3.02	
		3	3.49	-13.1	5.53	0.503	
		4	3.37	-16.6	10.1	5.53	
		5	3.28	-21.1	11.1	2.52	
		6	3.23	-10.6	3.02	4.02	

Table A.6: Clusters, z-statistics, and coordinates for RRE in the whole-brain analysis in changing/learnable sessions.

Description	Cluster index	Local maxima	Max z-stat	F99 coordinates (mm)			Cluster size (# voxels)
				X	Y	Z	
Intraparietal sulcus	1	1	5.09	12.1	-22.1	22.1	11050
		2	4.9	-14.1	-11.1	19.6	
		3	4.86	-13.1	-12.1	20.1	
		4	4.82	0	-9.05	21.6	
		5	4.77	-1.01	-8.55	22.1	
		6	4.69	0	-22.6	23.1	
Calcarine fissure / visual cortex	2	1	5.83	12.6	-35.2	0.503	10322
		2	5.44	13.1	-36.7	0.503	
		3	5.01	13.1	-34.2	2.52	
		4	4.98	12.6	-36.2	-0.5	
		5	4.93	12.1	-33.7	2.01	
		6	4.67	11.6	-33.7	1.01	
Ventrolateral PFC	3	1	4.85	-20.6	12.1	11.6	1548
		2	4.58	-20.1	13.6	12.1	
		3	4.56	-18.1	15.1	13.1	
		4	4.39	-21.6	14.1	13.1	
		5	4.2	-21.6	13.6	12.1	
		6	4.1	-19.6	9.05	13.1	
Posterior insula	4	1	4.71	16.6	-8.55	7.55	1518
		2	4.2	16.6	-11.1	7.55	
		3	4.19	18.1	-13.1	8.05	
		4	4.02	22.6	-10.1	9.05	
		5	3.84	18.1	-11.6	8.05	
		6	3.09	25.7	-9.05	10.6	

Table A.7: Clusters, z-statistics, and coordinates for VS in the whole-brain analysis.

B

FMRI cluster locations for Chapter 3

Partial + complete: horizon at first choice

Cluster Index	Z	x	y	z	Description
1	4.96	21.60	17.10	6.54	Right lOFC extending into dlPFC
1	4.10	9.05	3.52	22.60	
1	3.85	18.60	6.54	18.10	
1	3.77	12.10	14.60	13.60	
1	3.71	11.60	3.52	22.60	
1	3.67	15.60	5.53	16.60	
2	3.52	0.00	12.10	10.60	ACC extending into the right striatum
2	3.42	-0.50	14.10	10.10	
2	3.28	9.56	6.05	-1.01	
2	3.22	-0.50	12.10	11.60	
2	3.21	0.50	10.60	6.04	
2	3.17	0.50	9.05	5.53	
3	3.79	-11.60	19.60	12.10	Left dlPFC extending into the OFC
3	3.53	-7.04	20.10	8.55	
3	3.50	-7.04	20.10	9.56	
3	3.05	-16.60	15.60	16.10	
3	3.02	6.54	20.10	6.54	
3	3.00	-7.54	20.10	7.04	

Complete - partial: main effect at first choice

Cluster Index	Z	x	y	z	Description
1	3.30	-8.55	15.10	14.10	MCC bilateral
1	3.28	-7.54	15.10	14.10	
1	3.23	-4.53	15.10	14.60	
1	3.21	-2.51	11.10	16.10	
1	3.05	1.01	18.10	15.60	
1	2.94	2.01	15.10	15.60	

Complete - partial: ERchosen at first choice

Cluster Index	Z	x	y	z	Description
1	3.30	-8.55	15.10	14.10	Left dlPFC extending into left area 44
1	3.28	-7.54	15.10	14.10	
1	3.23	-4.53	15.10	14.60	
1	3.21	-2.51	11.10	16.10	
1	3.05	1.01	18.10	15.60	
1	2.94	2.01	15.10	15.60	
2	3.89	3.02	16.60	14.60	Right MCC extending to left MCC
2	3.53	1.01	14.60	16.10	
2	3.05	0.00	14.60	13.60	
2	2.93	-1.51	15.10	12.10	

Complete: inverted chosen PE at outcome

Cluster Index	Z	x	y	z	Description
1	4.15	19.10	9.05	-3.52	OFC
1	3.99	14.10	9.05	-2.51	
1	3.57	14.10	13.60	3.52	
1	3.47	24.10	13.60	3.52	
1	3.44	15.10	14.10	4.02	
1	3.38	20.10	13.60	3.52	

Complete: inverted unchosen PE at outcome

Cluster Index	Z	x	y	z	Description
1	3.45	2.52	15.10	-0.50	mOFC
1	3.38	-2.01	18.10	-1.51	
1	2.94	-5.53	15.10	2.01	
1	2.89	-5.53	11.60	1.51	
1	2.87	-5.53	12.60	2.01	
1	2.78	-1.01	21.60	0.00	

Table B.1: Tables showing the peaks of all significant clusters found within our frontal masks that are reported in the main text. Coordinates are given in the F99 standard space.

C

FMRI cluster locations for Chapter 3

Cluster Index	F99 coordinates Z	x	y	z	# voxels	description
4	5.14	27.2	-1.51	-10.6	19261	right temporal pole, right insula, bilateral striatum, bilateral area 11, 12 and 13, and area 25, 32, and 14
	4.56	27.7	-3.02	-9.05		
	4.33	16.1	17.1	8.05		
	4.22	-0.503	6.54	-0.503		
	4.2	6.54	18.6	7.55		
	4.17	22.6	2.52	-15.1		
3	4.28	29.2	-16.6	5.03	3923	superior temporal sulcus
	4.2	23.6	-16.1	0		
	4.11	25.2	-18.1	0.503		
	3.99	28.2	-7.04	7.04		
	3.95	24.1	-12.6	-5.53		
	3.81	27.2	-3.02	-1.01		
2	4.32	-21.1	-4.02	-1.01	2251	left insula extending into left claustrum
	4.15	-18.6	-2.01	-1.51		
	4.06	-21.6	-12.1	-8.05		
	4.06	-22.6	-4.53	-4.53		
	3.95	-22.6	-12.6	-8.05		
	3.95	-18.6	-9.05	1.01		
1	5.19	18.1	-8.55	19.1	1933	right motor cortex
	4.67	11.6	-12.6	20.1		
	4.45	13.6	-10.1	17.6		
	4.42	17.1	-9.56	17.6		
	4.36	11.6	-14.1	21.1		
	4.34	15.6	-9.56	19.1		

Table C.1: Significant neural clusters for CE at previous-trial-end and decision-prompt combined.

		F99 coordinates				
Cluster Index	Z	x	y	z	# voxels	description
4	4.65	4.02	10.1	4.02	15951	right Striatum, right OFC (area 13), frontopolar cortex (10), ACC (32 and 24), SMA and preSMA
	4.54	6.04	14.1	21.1		
	4.51	7.55	26.7	8.55		
	4.47	-1.01	21.1	13.6		
	4.47	1.01	5.03	9.56		
	4.47	11.6	18.1	14.1		
3	6.52	11.1	-13.1	-11.6	10981	medial temporal lope, hippocampus and entorhinal cortex
	6	11.1	-11.6	-12.1		
	5.06	24.1	-6.04	0.503		
	4.99	22.1	-10.1	2.52		
	4.99	24.1	-7.54	-2.51		
	4.93	26.2	-21.1	13.6		
2	4.86	-22.1	-5.53	-4.02	3822	left superior temporal lobe extending into left insula
	4.82	-22.1	-7.04	-5.03		
	4.79	-11.6	4.02	-6.54		
	4.63	-20.6	-4.02	2.01		
	4.63	-14.1	5.03	-6.04		
	4.61	-22.6	-7.54	-2.51		
1	5.02	16.6	7.04	-9.05	3177	right superior temporal lope extending into right insula
	4.99	18.1	6.04	-7.54		
	4.89	17.1	7.04	-8.05		
	4.83	12.1	8.55	-2.01		
	4.73	23.6	4.53	-8.55		
	4.41	17.6	8.05	-11.6		

Table C.2: Significant neural clusters for GE at previous-trial-end and decision-prompt combined.

		F99 coordinates				
Cluster Index	Z	x	y	z	# voxels	description
3	6.56	9.05	-11.6	-10.6	32606	right striatum, area 25, anterior area 24, area 32, right area 13, right hippocampus, extending into right temporal pole, right amygdala, right insula
	6.26	22.6	-9.05	1.51		
	6.06	24.6	-14.6	2.52		
	5.72	25.7	-20.6	13.6		
	5.63	13.6	8.55	-2.01		
	5.58	10.6	-10.1	-11.1		
2	5.23	-13.1	10.1	-0.503	4643	left insula, left temporal pole
	5.02	-10.6	3.52	-5.53		
	4.95	-14.1	7.04	-6.04		
	4.84	-15.1	5.53	-4.53		
	4.66	-13.6	8.05	-3.02		
	4.63	-7.04	1.01	-8.55		
1	6.78	-21.6	-2.01	-2.51	3696	left temporal pole and insula
	5.84	-22.1	-6.54	-2.01		
	5.81	-22.6	-5.03	-4.53		
	5.74	-23.6	-6.54	-1.51		
	5.57	-23.6	-2.01	-0.503		
	5.41	-22.6	-6.54	1.01		

Table C.3: Significant neural clusters for OE at previous-trial-end and decision-prompt combined.

Cluster Index	F99 coordinates Z	x	y	z	# voxels	description
3	4.49	1.91E-06	-23.6	18.1	23196	cingulate and dorsomedial frontal cortex
	4.47	0.503	-25.1	17.1		
	4.4	6.54	-18.6	13.6		
	4.37	1.01	5.53	11.1		
	4.33	2.01	-24.1	17.1		
	4.31	3.02	-21.6	14.6		
2	4.24	25.7	-33.7	-5.03	1917	right extrastriate visual association cortex
	4.1	22.6	-38.7	1.01		
	4.08	21.6	-37.7	0.503		
	4.02	18.6	-39.7	-11.1		
	3.97	17.6	-38.2	-9.05		
	3.96	25.7	-30.2	1.91E-06		
1	3.99	13.1	-38.2	10.6	1833	right extrastriate visual association cortex
	3.95	13.1	-37.7	8.05		
	3.89	13.1	-34.7	9.56		
	3.83	12.6	-36.2	10.1		
	3.82	13.1	-37.7	9.05		
	3.77	15.1	-38.2	11.6		

Table C.4: Significant neural clusters for ES at previous-trial-end and decision-prompt combined.

References

[1] Wolfram Schultz, Peter Dayan, and Pendleton Read Montague. "A neural substrate of prediction and reward." In: *Science (New York, N.Y.)* 275 (5306 Mar. 1997), pp. 1593–9. DOI: 10.1126/science.275.5306.1593.

[2] Miriam C. Klein-Flügge, Alessandro Bongioanni, and Matthew F.S. Rushworth. "Medial and orbital frontal cortex in decision-making and flexible behavior". In: *Neuron* (June 2022). DOI: 10.1016/J.NEURON.2022.05.022.

[3] Mark Laubach et al. "What, If Anything, Is Rodent Prefrontal Cortex?" In: *eNeuro* 5 (5 Sept. 2018), pp. 315–333. DOI: 10.1523/ENEURO.0315-18.2018.

[4] Kentaro Miyamoto et al. "Causal neural network of metamemory for retrospection in primates". In: *Science* 355 (6321 Jan. 2017), pp. 188–193. DOI: 10.1126/SCIENCE.AAL0162/SUPPL_FILE/MIYAMOTO.SM.PDF.

[5] Seng Bum Michael Yoo et al. "The neural basis of predictive pursuit". In: *Nature Neuroscience 2020 23:2* 23 (2 Jan. 2020), pp. 252–259. DOI: 10.1038/s41593-019-0561-6.

[6] Leon J. Kamin. "Selective association and conditioning". In: ed. by N.J. Mackintosh and W.K. Honig. Dalhousie University Press, June 1969, pp. 42–64.

[7] Moshe Bar et al. "Top-down facilitation of visual recognition". In: *Proceedings of the National Academy of Sciences of the United States of America* 103 (2 Jan. 2006), pp. 449–454. DOI: 10.1073/PNAS.0507062103/SUPPL_FILE/07062FIG9.PDF.

[8] Kosuke Akatsuka et al. "The effect of stimulus probability on the somatosensory mismatch field". In: *Experimental Brain Research* 181 (4 Aug. 2007), pp. 607–614. DOI: 10.1007/S00221-007-0958-4/FIGURES/7.

[9] Nachum Ulanovsky, Liora Las, and Israel Nelken. "Processing of low-probability sounds by cortical neurons". In: *Nature Neuroscience 2003 6:4* 6 (4 Mar. 2003), pp. 391–398. DOI: 10.1038/nn1032.

[10] Cynthia A. Erickson and Robert Desimone. "Responses of Macaque Perirhinal Neurons during and after Visual Stimulus Association Learning". In: *Journal of Neuroscience* 19 (23 Dec. 1999), pp. 10404–10416. DOI: 10.1523/JNEUROSCI.19-23-10404.1999.

[11] Sarah J. Blakemore, Susan J. Goodbody, and Daniel M. Wolpert. "Predicting the Consequences of Our Own Actions: The Role of Sensorimotor Context Estimation". In: *Journal of Neuroscience* 18 (18 Sept. 1998), pp. 7511–7518. DOI: 10.1523/JNEUROSCI.18-18-07511.1998.

[12] Marta Kutas and Steven A. Hillyard. “Reading Senseless Sentences: Brain Potentials Reflect Semantic Incongruity”. In: *Science* 207 (4427 1980), pp. 203–205. DOI: 10.1126/SCIENCE.7350657.

[13] Paul W. Glimcher. “Decisions, uncertainty, and the brain : the science of neuroeconomics”. In: (2003), p. 375.

[14] David W. Stephens and J. R. (John R.) Krebs. “Foraging Theory”. In: (1987), p. 263.

[15] David W. Stephens. “On economically tracking a variable environment”. In: *Theoretical Population Biology* 32 (1 Aug. 1987), pp. 15–25. DOI: 10.1016/0040-5809(87)90036-0.

[16] Sara J. Shettleworth et al. “Tracking a fluctuating environment: a study of sampling”. In: *Animal Behaviour* 36 (1 Feb. 1988), pp. 87–105. DOI: 10.1016/S0003-3472(88)80252-5.

[17] Staffan Tamm. “Tracking varying environments: sampling by hummingbirds”. In: *Animal Behaviour* 35 (6 Dec. 1987), pp. 1725–1734. DOI: 10.1016/S0003-3472(87)80065-9.

[18] Donald L. Kramer and Daniel M. Weary. “Exploration versus exploitation: a field study of time allocation to environmental tracking by foraging chipmunks”. In: *Animal Behaviour* 41 (3 Mar. 1991), pp. 443–449. DOI: 10.1016/S0003-3472(05)80846-2.

[19] Aleksandr Aronovich Feldbaum. “Optimal control systems”. In: (1965), p. 452.

[20] James G. March. “Exploration and Exploitation in Organizational Learning”. In: *https://doi.org/10.1287/orsc.2.1.71* 2 (1 Feb. 1991), pp. 71–87. DOI: 10.1287/ORSC.2.1.71.

[21] Juha Uotila et al. “Exploration, exploitation, and financial performance: analysis of SandP 500 corporations”. In: *Strategic Management Journal* 30 (2 Feb. 2009), pp. 221–231. DOI: 10.1002/SMJ.738.

[22] Christopher John Cornish Hellaby Watkins. “Learning from Delayed Rewards”. University of Cambridge, 1989.

[23] R. Duncan Luce. *Individual choice behavior*. John Wiley, 1959.

[24] Richard S. Sutton and Andrew G. Barto. “Reinforcement Learning: An Introduction”. In: (1998).

[25] Nathaniel D. Daw et al. “Cortical substrates for exploratory decisions in humans”. In: *Nature* 441 (7095 June 2006), pp. 876–879. DOI: 10.1038/nature04766.

[26] Charles Findling et al. “Computational noise in reward-guided learning drives behavioral variability in volatile environments”. In: *Nature Neuroscience 2019 22:12* 22 (12 Oct. 2019), pp. 2066–2077. DOI: 10.1038/s41593-019-0518-9.

[27] Robert C. Wilson et al. “Orbitofrontal cortex as a cognitive map of task space”. In: *Neuron* (2014). DOI: 10.1016/j.neuron.2013.11.005.

[28] Clemence Almeras, Valerian Chambon, and Valentin Wyart. “Competing cognitive pressures on human exploration in the absence of trade-off with exploitation”. In: *bioarxiv* (2022). DOI: 10.31234/OSF.IO/9QPUZ.

[29] Ethan S. Bromberg-Martin and Okihide Hikosaka. "Midbrain dopamine neurons signal preference for advance information about upcoming rewards." In: *Neuron* 63 (1 July 2009), pp. 119–126. DOI: 10.1016/J.NEURON.2009.06.009/ATTACHMENT/9850476C-3B83-43E5-9700-B0DA2BAC4818/MMC1.PDF.

[30] Ethan S. Bromberg-Martin and Ilya E. Monosov. "Neural circuitry of information seeking". In: *Current opinion in behavioral sciences* 35 (Oct. 2020), p. 62. DOI: 10.1016/J.COBEHA.2020.07.006.

[31] Jacqueline Gottlieb et al. "Information seeking, curiosity and attention: computational and neural mechanisms". In: *Trends in cognitive sciences* 17 (11 Nov. 2013), p. 585. DOI: 10.1016/J.TICS.2013.09.001.

[32] Erie D. Boorman et al. "How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action". In: *Neuron* 62 (5 June 2009), pp. 733–743. DOI: 10.1016/j.neuron.2009.05.014.

[33] David Badre et al. "Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration". In: *Neuron* 73 (3 Feb. 2012), p. 595. DOI: 10.1016/J.NEURON.2011.12.025.

[34] Wojciech K. Zajkowski, Malgorzata Kossut, and Robert C. Wilson. "A causal role for right frontopolar cortex in directed, but not random, exploration". In: *eLife* 6 (Sept. 2017). DOI: 10.7554/ELIFE.27430.

[35] Gary Aston-Jones and Jonathan D. Cohen. "An Integrative Theory Of Locus Coeruleus-Norepinephrine Function: Adaptive Gain and Optimal Performance". In: *Annual Review of Neuroscience* 28 (1 July 2005), pp. 403–450. DOI: 10.1146/annurev.neuro.28.061604.135709.

[36] Siddhartha Joshi et al. "Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex". In: *Neuron* 89 (1 Jan. 2016), p. 221. DOI: 10.1016/J.NEURON.2015.11.028.

[37] Marieke Jepma and Sander Nieuwenhuis. "Pupil diameter predicts changes in the exploration-exploitation trade-off: Evidence for the adaptive gain theory". In: *Journal of Cognitive Neuroscience* 23 (7 July 2011), pp. 1587–1596. DOI: 10.1162/JOCN.2010.21548.

[38] Magda Dubois et al. "Human complex exploration strategies are enriched by noradrenaline-modulated heuristics". In: *eLife* 10 (Jan. 2021), pp. 1–34. DOI: 10.7554/ELIFE.59907.

[39] D. Gowanlock R. Tervo et al. "The anterior cingulate cortex directs exploration of alternative strategies". In: *Neuron* 109 (11 June 2021), 1876–1887.e6. DOI: 10.1016/J.NEURON.2021.03.028.

[40] Robert A Rescorla and Allan R Wagner. "A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement". In: *Classical Conditioning II Current Research and Theory* 21 (6 1972), pp. 64–99. DOI: 10.1101/gr.110528.110.

[41] Timothy E J Behrens et al. "Learning the value of information in an uncertain world". In: *Nature Neuroscience* 10 (9 Sept. 2007), pp. 1214–1221. DOI: 10.1038/nn1954.

[42] Payam Piray and Nathaniel D. Daw. "A model for learning based on the joint estimation of stochasticity and volatility". In: *Nature Communications 2021 12:1* 12 (1 Nov. 2021), pp. 1–16. DOI: 10.1038/s41467-021-26731-9.

[43] P. Read Montague, Peter Dayan, and Terrence J. Sejnowski. "A framework for mesencephalic dopamine systems based on predictive Hebbian learning". In: *Journal of Neuroscience* 16 (5 Mar. 1996), pp. 1936–1947. DOI: 10.1523/JNEUROSCI.16-05-01936.1996.

[44] Neir Eshel et al. "Dopamine neurons share common response function for reward prediction error". In: *Nature Neuroscience 2016 19:3* 19 (3 Feb. 2016), pp. 479–486. DOI: 10.1038/nn.4239.

[45] Will Dabney et al. "A distributional code for value in dopamine-based reinforcement learning". In: *Nature 2020 577:7792* 577 (7792 Jan. 2020), pp. 671–675. DOI: 10.1038/s41586-019-1924-6.

[46] Will Dabney et al. "Implicit Quantile Networks for Distributional Reinforcement Learning". In: *35th International Conference on Machine Learning, ICML 2018* 3 (June 2018), pp. 1774–1787. DOI: 10.48550/arxiv.1806.06923.

[47] Moritz Moeller et al. "An association between prediction errors and risk-seeking: Theory and behavioral evidence". In: *PLOS Computational Biology* 17 (7 July 2021), e1009213. DOI: 10.1371/JOURNAL.PCBI.1009213.

[48] Kimberlee D'Ardenne et al. "BOLD responses reflecting dopaminergic signals in the human ventral tegmental area". In: *Science* 319 (5867 Feb. 2008), pp. 1264–1267. DOI: 10.1126/SCIENCE.1150605/SUPPL_FILE/DARDENNE.SOM.PDF.

[49] Mathias Pessiglione et al. "Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans". In: *Nature 2006 442:7106* 442 (7106 Aug. 2006), pp. 1042–1045. DOI: 10.1038/nature05051.

[50] James D. Howard and Thorsten Kahnt. "Identity prediction errors in the human midbrain update reward-identity expectations in the orbitofrontal cortex". In: *Nature Communications 2018 9:1* 9 (1 Apr. 2018), pp. 1–11. DOI: 10.1038/s41467-018-04055-5.

[51] James D. Howard and Thorsten Kahnt. "To be specific: The role of orbitofrontal cortex in signaling reward identity". In: *Behavioral neuroscience* 135 (2 Mar. 2021), pp. 210–217. DOI: 10.1037/BNE0000455.

[52] Armin Lak, William R. Stauffer, and Wolfram Schultz. "Dopamine neurons learn relative chosen value from probabilistic rewards". In: *eLife* 5 (Oct. 2016). DOI: 10.7554/ELIFE.18044.001.

[53] Wolfram Schultz, Wiliam R. Stauffer, and Armin Lak. "The phasic dopamine signal maturing: from reward via behavioural activation to formal economic utility". In: *Current Opinion in Neurobiology* 43 (Apr. 2017), pp. 139–148. DOI: 10.1016/J.CONB.2017.03.013.

[54] Kathryn M. Rothenhoefer et al. "Rare rewards amplify dopamine responses". In: *Nature Neuroscience 2021 24:4* 24 (4 Mar. 2021), pp. 465–469. DOI: 10.1038/s41593-021-00807-7.

[55] Moritz Moeller, Sanjay Manohar, and Rafasl Bogacz. "Uncertainty–guided learning with scaled prediction errors in the basal ganglia". In: *PLOS Computational Biology* 18 (5 May 2022), e1009816. DOI: 10.1371/JOURNAL.PCBI.1009816.

[56] Philippe N. Tobler, Christopher D. Fiorillo, and Wolfram Schultz. "Adaptive coding of reward value by dopamine neurons". In: *Science* 307 (5715 2005), pp. 1642–1645. DOI: 10.1126/science.1105370.

[57] David Meder et al. "Simultaneous representation of a spectrum of dynamically changing value estimates during decision making". In: *Nature Communications* 8 (1 Dec. 2017), p. 1942. DOI: 10.1038/s41467-017-02169-w.

[58] Elsa F. Fouragnan et al. "The macaque anterior cingulate cortex translates counterfactual choice value into actual behavioral change". In: *Nature Neuroscience 2019 22:5* 22 (5 Apr. 2019), pp. 797–808. DOI: 10.1038/s41593-019-0375-6.

[59] Nils Kolling, Marco Wittmann, and Matthew F.S. Rushworth. "Multiple neural mechanisms of decision making and their competition under changing risk pressure". In: *Neuron* 81 (5 Mar. 2014), pp. 1190–1202. DOI: 10.1016/J.NEURON.2014.01.033/ATTACHMENT/A9984A25-8961-453B-8E1A-778EAD3A0ADE/MMC1.PDF.

[60] Nils Kolling et al. "Prospection, Perseverance, and Insight in Sequential Behavior". In: *Neuron* (2018). DOI: 10.1016/j.neuron.2018.08.018.

[61] Clay B. Holroyd et al. "Human midcingulate cortex encodes distributed representations of task progress". In: *Proceedings of the National Academy of Sciences of the United States of America* 115 (25 June 2018), pp. 6398–6403. DOI: 10.1073/PNAS.1803650115/SUPPL_FILE/PNAS.1803650115.SAPP.PDF.

[62] Erie D. Boorman, Matthew F. Rushworth, and Tim E. Behrens. "Ventromedial prefrontal and anterior cingulate cortex adopt choice and default reference frames during sequential multi-alternative choice". In: *The Journal of neuroscience : the official journal of the Society for Neuroscience* 33 (6 Feb. 2013), pp. 2242–2253. DOI: 10.1523/JNEUROSCI.3022-12.2013.

[63] Laurence T Hunt et al. "Mechanisms underlying cortical activity during value-guided choice". In: *Nature Neuroscience* 15 (3 Jan. 2012), pp. 470–476. DOI: 10.1038/nn.3017.

[64] Xiao-Jing Wang. "Probabilistic Decision Making by Slow Reverberation in Cortical Circuits". In: *Neuron* 36 (5 Dec. 2002), pp. 955–968. DOI: 10.1016/S0896-6273(02)01092-9.

[65] Xiao Jing Wang. "Decision Making in Recurrent Neuronal Circuits". In: *Neuron* 60 (2 Oct. 2008), pp. 215–234. DOI: 10.1016/J.NEURON.2008.09.034.

[66] Kong Fatt Wong and Xiao Jing Wang. "A Recurrent Network Mechanism of Time Integration in Perceptual Decisions". In: *Journal of Neuroscience* 26 (4 Jan. 2006), pp. 1314–1328. DOI: 10.1523/JNEUROSCI.3733-05.2006.

[67] MaryAnn P. Noonan et al. "Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex." In: *Proceedings of the National Academy of Sciences of the United States of America* 107 (47 Nov. 2010), pp. 20547–52. DOI: 10.1073/pnas.1012246107.

[68] Nathalie Camille et al. "Ventromedial Frontal Lobe Damage Disrupts Value Maximization in Humans". In: *Journal of Neuroscience* 31 (20 May 2011), pp. 7527–7532. DOI: 10.1523/JNEUROSCI.6527-10.2011.

[69] MaryAnn P. Noonan et al. "Contrasting Effects of Medial and Lateral Orbitofrontal Cortex Lesions on Credit Assignment and Decision-Making in Humans". In: *The Journal of Neuroscience* 37 (29 July 2017), pp. 7023–7035. DOI: 10.1523/JNEUROSCI.0692-17.2017.

[70] Adam P. Steiner and A. David Redish. "Behavioral and neurophysiological correlates of regret in rat decision-making on a neuroeconomic task". In: *Nature Neuroscience 2014 17:7* 17 (7 June 2014), pp. 995–1002. DOI: 10.1038/nn.3740.

[71] Giorgio Coricelli et al. "Regret and its avoidance: a neuroimaging study of choice behavior". In: *Nature Neuroscience 2005 8:9* 8 (9 Aug. 2005), pp. 1255–1262. DOI: 10.1038/nn1514.

[72] Nathalie Camille et al. "The involvement of the orbitofrontal cortex in the experience of regret". In: *Science* 304 (5674 May 2004), pp. 1167–1170. DOI: 10.1126/SCIENCE.1094550/SUPPL_FILE/CAMILLE_SOM.PDF.

[73] Benedetto De Martino et al. "Confidence in value-based choice". In: *Nature Neuroscience 2012 16:1* 16 (1 Dec. 2012), pp. 105–110. DOI: 10.1038/nn.3279.

[74] Thomas H.B. FitzGerald, Ben Seymour, and Raymond J. Dolan. "The Role of Human Orbitofrontal Cortex in Value Comparison for Incommensurable Objects". In: *The Journal of Neuroscience* 29 (26 July 2009), p. 8388. DOI: 10.1523/JNEUROSCI.0717-09.2009.

[75] Nadescha Trudel et al. "Polarity of uncertainty representation during exploration and exploitation in ventromedial prefrontal cortex". In: *Nature Human Behaviour 2020 5:1* 5 (1 Aug. 2020), pp. 83–98. DOI: 10.1038/s41562-020-0929-3.

[76] Alessandro Bongioanni et al. "Activation and disruption of a neural mechanism for novel choice in monkeys". In: *Nature 2021 591:7849* 591 (7849 Jan. 2021), pp. 270–274. DOI: 10.1038/s41586-020-03115-5.

[77] Georgios K. Papageorgiou et al. "Inverted activity patterns in ventromedial prefrontal cortex during value-guided decision-making in a less-is-more task". In: *Nature Communications 2017 8:1* 8 (1 Dec. 2017), pp. 1–14. DOI: 10.1038/s41467-017-01833-5.

[78] Marco K. Wittmann et al. "Global reward state affects learning and activity in raphe nucleus and anterior insula in monkeys". In: *Nature Communications 2020 11:1* 11 (1 July 2020), pp. 1–17. DOI: 10.1038/s41467-020-17343-w.

[79] Yael Niv. "Learning task-state representations". In: *Nature Neuroscience 2019 22:10* 22 (10 Sept. 2019), pp. 1544–1553. DOI: 10.1038/s41593-019-0470-8.

[80] Pietro Vertechi et al. "Inference-Based Decisions in a Hidden State Foraging Task: Differential Contributions of Prefrontal Cortical Areas". In: *Neuron* 106 (1 Apr. 2020), p. 166. DOI: 10.1016/J.NEURON.2020.01.017.

[81] Yanhe Liu, Yu Xin, and Ning long Xu. "A cortical circuit mechanism for structural knowledge-based flexible sensorimotor decision-making". In: *Neuron* 109 (12 June 2021), 2009–2024.e6. DOI: 10.1016/J.NEURON.2021.04.014.

[82] Jingfeng Zhou et al. "Rat Orbitofrontal Ensemble Activity Contains Multiplexed but Dissociable Representations of Value and Task Structure in an Odor Sequence Task". In: *Current Biology* 29 (6 Mar. 2019), 897–907.e3. DOI: 10.1016/J.CUB.2019.01.048.

[83] Jingfeng Zhou et al. "Evolving schema representations in orbitofrontal ensembles during learning". In: *Nature 2020 590:7847* 590 (7847 Dec. 2020), pp. 606–611. DOI: 10.1038/s41586-020-03061-2.

[84] Alexandra O. Constantinescu, Jill X. O'Reilly, and Timothy E.J. Behrens. "Organizing conceptual knowledge in humans with a gridlike code". In: *Science* 352 (6292 June 2016), pp. 1464–1468. DOI: 10.1126/SCIENCE.AAF0941/SUPPL_FILE/CONSTANTINESCU.SM.PDF.

[85] Christian F. Doeller, Caswell Barry, and Neil Burgess. "Evidence for grid cells in a human memory network". In: *Nature 2010 463:7281* 463 (7281 Jan. 2010), pp. 657–661. DOI: 10.1038/nature08704.

[86] Miriam C. Klein-Flügge et al. "Multiple associative structures created by reinforcement and incidental statistical learning mechanisms". In: *Nature Communications 2019 10:1* 10 (1 Oct. 2019), pp. 1–15. DOI: 10.1038/s41467-019-12557-z.

[87] Mark E Walton, David M Bannerman, and Matthew F S Rushworth. "The role of rat medial frontal cortex in effort-based decision making." In: *The Journal of neuroscience : the official journal of the Society for Neuroscience* 22 (24 Dec. 2002), pp. 10996–1003. DOI: 10.1523/JNEUROSCI.22-24-10996.2002.

[88] Ken-ichi Amemori and Ann M Graybiel. "Localized microstimulation of primate pregenual cingulate cortex induces negative decision-making". In: *Nature Neuroscience* 15 (5 May 2012), pp. 776–785. DOI: 10.1038/nn.3088.

[89] Maria K. Eckstein et al. "The Interpretation of Computational Model Parameters Depends on the Context". In: *bioRxiv* (June 2022), p. 2021.05.28.446162. DOI: 10.1101/2021.05.28.446162.

[90] Hugo Malagon-Vina et al. "Fluid network dynamics in the prefrontal cortex during multiple strategy switching". In: *Nature Communications 2018 9:1* 9 (1 Jan. 2018), pp. 1–13. DOI: 10.1038/s41467-017-02764-x.

[91] Yael Niv. "Cost, benefit, tonic, phasic: what do response rates tell us about dopamine and motivation?" In: *Annals of the New York Academy of Sciences* 1104 (2007), pp. 357–376. DOI: 10.1196/ANNALS.1390.018.

[92] Yael Niv, Nathaniel D Daw, and Peter Dayan. "How fast to work: Response vigor, motivation and tonic dopamine". In: *Advances in Neural Information Processing Systems* 18 (2005).

[93] Yael Niv et al. "Tonic dopamine: opportunity costs and the control of response vigor". In: *Psychopharmacology 2007 191:3* 191 (3 Oct. 2007), pp. 507–520. DOI: 10.1007/S00213-006-0502-4.

[94] Arif A. Hamid et al. "Mesolimbic dopamine signals the value of work". In: *Nature Neuroscience 2016 19:1* 19 (1 Nov. 2015), pp. 117–126. DOI: 10.1038/nn.4173.

[95] Alexandre Zénon, Sophie Devesse, and Etienne Olivier. "Dopamine Manipulation Affects Response Vigor Independently of Opportunity Cost". In: *Journal of Neuroscience* 36 (37 Sept. 2016), pp. 9516–9525. DOI: 10.1523/JNEUROSCI.4467-15.2016.

[96] Satoshi Ikemoto and Jaak Panksepp. "The role of nucleus accumbens dopamine in motivated behavior: A unifying interpretation with special reference to reward-seeking". In: *Brain Research Reviews* 31 (1 1999), pp. 6–41. DOI: 10.1016/S0165-0173(99)00023-5.

[97] John D. Salamone et al. "Dopamine, Behavioral Economics, and Effort". In: *Frontiers in Behavioral Neuroscience* 3 (SEP Sept. 2009). DOI: 10.3389/NEURO.08.013.2009.

[98] Joshua D. Berke. "What does dopamine mean?" In: *Nature Neuroscience 2018 21:6* 21 (6 May 2018), pp. 787–793. DOI: 10.1038/s41593-018-0152-y.

[99] Wolfram Schultz. "Behavioral Theories and the Neurophysiology of Reward". In: *Annual Review of Psychology* 57 (Nov. 2005), pp. 87–115. DOI: 10.1146/ANNUREV.PSYCH.56.091103.070229.

[100] Wolfram Schultz. "Updating dopamine reward signals". In: *Current Opinion in Neurobiology* 23 (2 Apr. 2013), pp. 229–238. DOI: 10.1016/J.CONB.2012.11.012.

[101] William R. Stauffer et al. "Dopamine Neuron-Specific Optogenetic Stimulation in Rhesus Macaques". In: *Cell* 166 (6 Sept. 2016), 1564–1571.e6. DOI: 10.1016/j.cell.2016.08.024.

[102] Chun Yun Chang et al. "Brief optogenetic inhibition of dopamine neurons mimics endogenous negative reward prediction errors". In: *Nature Neuroscience 2016 19:1* 19 (1 Dec. 2015), pp. 111–116. DOI: 10.1038/nn.4191.

[103] Andrew S. Hart et al. "Phasic Dopamine Release in the Rat Nucleus Accumbens Symmetrically Encodes a Reward Prediction Error Term". In: *Journal of Neuroscience* 34 (3 Jan. 2014), pp. 698–704. DOI: 10.1523/JNEUROSCI.2489-13.2014.

[104] John P. O'Doherty et al. "Temporal difference models and reward-related learning in the human brain". In: *Neuron* 38 (2 Apr. 2003), pp. 329–337. DOI: 10.1016/S0896-6273(03)00169-7.

[105] John O'Doherty et al. "Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning". In: *Science* 304 (5669 Apr. 2004), pp. 452–454. DOI: 10.1126/SCIENCE.1094285/SUPPL_FILE/O_DOHERTY.PDF.

[106] Todd A. Hare et al. "Dissociating the Role of the Orbitofrontal Cortex and the Striatum in the Computation of Goal Values and Prediction Errors". In: *Journal of Neuroscience* 28 (22 May 2008), pp. 5623–5630. DOI: 10.1523/JNEUROSCI.1309-08.2008.

[107] Philippe N. Tobler et al. "Reward Value Coding Distinct From Risk Attitude-Related Uncertainty Coding in Human Reward Systems". In: *Journal of Neurophysiology* 97 (2 Feb. 2007), pp. 1621–1632. DOI: 10.1152/jn.00745.2006.

[108] Miriam C. Klein-Flügge et al. "Dissociable Reward and Timing Signals in Human Midbrain and Ventral Striatum". In: *Neuron* 72 (4 Nov. 2011), p. 654. DOI: 10.1016/J.NEURON.2011.08.024.

[109] Grit Hein et al. "How learning shapes the empathic brain". In: *Proceedings of the National Academy of Sciences of the United States of America* 113 (1 Jan. 2016), pp. 80–85. DOI: 10.1073/PNAS.1514539112/SUPPL_FILE/PNAS.201514539SI.PDF.

[110] Robb B. Rutledge et al. "Testing the Reward Prediction Error Hypothesis with an Axiomatic Model". In: *Journal of Neuroscience* 30 (40 Oct. 2010), pp. 13525–13536. DOI: 10.1523/JNEUROSCI.1747-10.2010.

[111] Madoka Matsumoto et al. "Medial prefrontal cell activity signaling prediction errors of action values". In: *Nature Neuroscience 2007 10:5* 10 (5 Apr. 2007), pp. 647–656. DOI: 10.1038/nn1890.

[112] Hyojung Seo and Daeyeol Lee. "Temporal Filtering of Reward Signals in the Dorsal Anterior Cingulate Cortex during a Mixed-Strategy Game". In: *Journal of Neuroscience* 27 (31 Aug. 2007), pp. 8366–8377. DOI: 10.1523/JNEUROSCI.2369-07.2007.

[113] Otto Leif Tinklepaugh. "An experimental study of representative factors in monkeys". In: *Journal of Comparative Psychology* 8 (3 June 1928), pp. 197–236. DOI: 10.1037/H0075798.

[114] Kathryn A. Burke et al. "The role of the orbitofrontal cortex in the pursuit of happiness and more specific rewards". In: *Nature 2008 454:7202* 454 (7202 July 2008), pp. 340–344. DOI: 10.1038/nature06993.

[115] Elisabeth A. Murray and Peter H. Rudebeck. "Specializations for reward-guided decision-making in the primate ventral prefrontal cortex". In: *Nature Reviews Neuroscience 2018 19:7* 19 (7 May 2018), pp. 404–417. DOI: 10.1038/s41583-018-0013-4.

[116] Peter H. Rudebeck et al. "Specialized Representations of Value in the Orbital and Ventrolateral Prefrontal Cortex: Desirability versus Availability of Outcomes". In: *Neuron* 95 (5 Aug. 2017), 1208–1220.e5. DOI: 10.1016/j.neuron.2017.07.042.

[117] Elisabeth A. Murray et al. "Specialized areas for value updating and goal selection in the primate orbitofrontal cortex". In: *eLife* 4 (Dec. 2015). DOI: 10.7554/ELIFE.11695.

[118] Nina Lopatina et al. "Medial Orbitofrontal Neurons Preferentially Signal Cues Predicting Changes in Reward during Unblocking". In: *Journal of Neuroscience* 36 (32 Aug. 2016), pp. 8416–8424. DOI: 10.1523/JNEUROSCI.1101-16.2016.

[119] Nina Lopatina et al. "Lateral orbitofrontal neurons acquire responses to upshifted, downshifted, or blocked cues during unblocking". In: *eLife* 4 (Dec. 2015). DOI: 10.7554/ELIFE.11299.001.

[120] Masataka Watanabe. "Reward expectancy in primate prefrental neurons". In: *Nature 1996 382:6592* 382 (6592 Aug. 1996), pp. 629–632. DOI: 10.1038/382629a0.

[121] Anthony Dickinson. "Actions and habits: the development of behavioural autonomy". In: *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 308 (1135 Feb. 1985), pp. 67–78. DOI: 10.1098/RSTB.1985.0010.

[122] Yael Niv and Geoffrey Schoenbaum. "Dialogues on prediction errors". In: *Trends in Cognitive Sciences* 12 (7 July 2008), pp. 265–272. DOI: 10.1016/j.tics.2008.03.006.

[123] Justin R. Chumbley et al. "Surprise beyond prediction error". In: *Human Brain Mapping* 35 (9 2014), p. 4805. DOI: 10.1002/HBM.22513.

[124] Elsa Fouragnan, Chris Retzler, and Marios G. Philiastides. "Separate neural representations of prediction error valence and surprise: Evidence from an fMRI meta-analysis". In: *Human Brain Mapping* 39 (7 July 2018), pp. 2887–2906. DOI: 10.1002/HBM.24047.

[125] Céline. Amiez, Jean-Paul Joseph, and Emmanuel. Procyk. "Reward Encoding in the Monkey Anterior Cingulate Cortex". In: *Cerebral Cortex* 16 (7 July 2006), pp. 1040–1055. DOI: 10.1093/CERCOR/BHJ046.

[126] John M. Pearce and Geoffrey Hall. "A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli". In: *Psychological Review* 87 (6 Nov. 1980), pp. 532–552. DOI: 10.1037/0033-295X.87.6.532.

[127] Matthew R. Nassar et al. "An Approximately Bayesian Delta-Rule Model Explains the Dynamics of Belief Updating in a Changing Environment". In: *Journal of Neuroscience* 30 (37 Sept. 2010), pp. 12366–12378. DOI: 10.1523/JNEUROSCI.0822-10.2010.

[128] Matthew FS Rushworth, Rogier B. Mars, and Christopher Summerfield. "General mechanisms for making decisions?" In: *Current Opinion in Neurobiology* 19 (1 Feb. 2009), pp. 75–83. DOI: 10.1016/J.CONB.2009.02.005.

[129] Ethan S. Bromberg-Martin, Masayuki Matsumoto, and Okihide Hikosaka. "Dopamine in motivational control: rewarding, aversive, and alerting". In: *Neuron* 68 (5 Dec. 2010), p. 815. DOI: 10.1016/J.NEURON.2010.11.022.

[130] Wolfram Schultz. "Multiple reward signals in the brain". In: *Nature Reviews Neuroscience 2000 1:3* 1 (3 2000), pp. 199–207. DOI: 10.1038/35044563.

[131] Jan P. Gläscher and John P. O'Doherty. "Model-based approaches to neuroimaging: combining reinforcement learning theory with fMRI data". In: *Wiley Interdisciplinary Reviews: Cognitive Science* 1 (4 July 2010), pp. 501–510. DOI: 10.1002/WCS.57.

[132] Matthew F.S. Rushworth et al. "Frontal Cortex and Reward-Guided Learning and Decision-Making". In: *Neuron* 70 (6 June 2011), pp. 1054–1069. DOI: 10.1016/J.NEURON.2011.05.014.

[133] Antonio Rangel, Colin Camerer, and P. Read Montague. "A framework for studying the neurobiology of value-based decision making". In: *Nature Reviews Neuroscience* 9 (7 July 2008), pp. 545–556. DOI: 10.1038/nrn2357.

[134] Nathaniel D. Daw et al. "Model-based influences on humans' choices and striatal prediction errors". In: *Neuron* 69 (6 Mar. 2011), p. 1204. DOI: 10.1016/J.NEURON.2011.02.027.

[135] Robert C. Wilson and Yael Niv. "Is Model Fitting Necessary for Model-Based fMRI?" In: *PLOS Computational Biology* 11 (6 June 2015), e1004237. DOI: 10.1371/JOURNAL.PCBI.1004237.

[136] Armin Lak, William R. Stauffer, and Wolfram Schultz. "Dopamine prediction error responses integrate subjective value from different reward dimensions". In: *Proceedings of the National Academy of Sciences of the United States of America* 111 (6 Feb. 2014), pp. 2343–2348. DOI: 10.1073/PNAS.1321596111/SUPPL_FILE/PNAS.201321596SI.PDF.

[137] Javier A. Suarez et al. "Sensory prediction errors in the human midbrain signal identity violations independent of perceptual distance". In: *eLife* 8 (2019). DOI: 10.7554/ELIFE.43962.

[138] Yuji K. Takahashi et al. "Dopamine Neurons Respond to Errors in the Prediction of Sensory Features of Expected Rewards". In: *Neuron* 95 (6 Sept. 2017), 1395–1405.e3. DOI: 10.1016/j.neuron.2017.08.025.

[139] Erie D. Boorman et al. "Two Anatomically and Computationally Distinct Learning Signals Predict Changes to Stimulus-Outcome Associations in Hippocampus". In: *Neuron* 89 (6 Mar. 2016), pp. 1343–1354. DOI: 10.1016/j.neuron.2016.02.014.

[140] Sandra Iglesias et al. "Hierarchical prediction errors in midbrain and basal forebrain during sensory learning." In: *Neuron* 80 (2 Oct. 2013), pp. 519–30. DOI: 10.1016/j.neuron.2013.09.009.

[141] Kou Murayama et al. "Neural basis of the undermining effect of monetary reward on intrinsic motivation". In: *Proceedings of the National Academy of Sciences of the United States of America* 107 (49 Dec. 2010), pp. 20911–20916. DOI: 10.1073/PNAS.1013305107/SUPPL_FILE/PNAS.201013305SI.PDF.

[142] Matthew R. Nassar, Rasmus Bruckner, and Michael J. Frank. "Statistical context dictates the relationship between feedback-related EEG signals and learning". In: *eLife* 8 (Aug. 2019). DOI: 10.7554/ELIFE.46975.

[143] Jill X. O'Reilly et al. "Dissociable effects of surprise and model update in parietal and anterior cingulate cortex". In: *Proceedings of the National Academy of Sciences of the United States of America* 110 (38 Oct. 2013), E3660–E3669. DOI: 10.1073/PNAS.1305373110/SUPPL_FILE/PNAS.201305373SI.PDF.

[144] Daeyeol Lee et al. "Functional Specialization of the Primate Frontal Cortex during Decision Making". In: *Journal of Neuroscience* 27 (31 Aug. 2007), pp. 8170–8173. DOI: 10.1523/JNEUROSCI.1561-07.2007.

[145] Natalie Caspari et al. "Functional Similarity of Medial Superior Parietal Areas for Shift-Selective Attention Signals in Humans and Monkeys". In: *Cerebral cortex (New York, N.Y. : 1991)* 28 (6 June 2018), pp. 2085–2099. DOI: 10.1093/CERCOR/BHX114.

[146] Natalie Caspari et al. "Covert Shifts of Spatial Attention in the Macaque Monkey". In: *Journal of Neuroscience* 35 (20 May 2015), pp. 7695–7714. DOI: 10.1523/JNEUROSCI.4383-14.2015.

[147] Danielle C. Turner et al. "The role of the lateral frontal cortex in causal associative learning: exploring preventative and super-learning". In: *Cerebral cortex (New York, N.Y. : 1991)* 14 (8 Aug. 2004), pp. 872–880. DOI: 10.1093/CERCOR/BHH046.

[148] Nima Khalighinejad et al. “A Basal Forebrain-Cingulate Circuit in Macaques Decides It Is Time to Act”. In: *Neuron* 105 (2 Jan. 2020), 370–384.e8. DOI: 10.1016/J.NEURON.2019.10.030.

[149] Mark Jenkinson et al. “FSL”. In: *NeuroImage* 62 (2 Aug. 2012), pp. 782–790. DOI: 10.1016/J.NEUROIMAGE.2011.09.015.

[150] Brian B. Avants et al. “The Insight ToolKit image registration framework”. In: *Frontiers in Neuroinformatics* 8 (APR Apr. 2014). DOI: 10.3389/FNINF.2014.00044.

[151] Hauke Kolster et al. “Visual Field Map Clusters in Macaque Extrastriate Visual Cortex”. In: *Journal of Neuroscience* 29 (21 May 2009), pp. 7031–7039. DOI: 10.1523/JNEUROSCI.0518-09.2009.

[152] David C. Van Essen and Donna L. Dierker. “Surface-Based and Probabilistic Atlases of Primate Cerebral Cortex”. In: *Neuron* 56 (2 Oct. 2007), pp. 209–225. DOI: 10.1016/J.NEURON.2007.10.015.

[153] Keith. J. Worsley et al. “A Three-Dimensional Statistical Analysis for CBF Activation Studies in Human Brain:” in: *http://dx.doi.org/10.1038/jcbfm.1992.127* 12 (6 June 1992), pp. 900–918. DOI: 10.1038/JCBFM.1992.127.

[154] Colin Reveley et al. “Three-Dimensional Digital Template Atlas of the Macaque Brain”. In: *Cerebral Cortex (New York, NY)* 27 (9 2017), p. 4463. DOI: 10.1093/CERCOR/BHW248.

[155] William R. Thompson. “On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples”. In: *Biometrika* 25 (3/4 Dec. 1933), p. 285. DOI: 10.2307/2332286.

[156] Elise Payzan-LeNestour and Peter Bossaerts. “Risk, Unexpected Uncertainty, and Estimation Uncertainty: Bayesian Learning in Unstable Settings”. In: *PLoS Computational Biology* 7 (1 Jan. 2011). Ed. by Tim Behrens, e1001048. DOI: 10.1371/journal.pcbi.1001048.

[157] Cécile Garcia et al. “Balancing costs and benefits in primates: ecological and palaeoanthropological views”. In: *Philosophical Transactions of the Royal Society B* 376 (1819 Mar. 2021). DOI: 10.1098/RSTB.2019.0667.

[158] Christopher M. Warren et al. “The effect of atomoxetine on random and directed exploration in humans”. In: *PLoS ONE* 12 (4 Apr. 2017). DOI: 10.1371/JOURNAL.PONE.0176034.

[159] Robert C. Wilson et al. “Balancing exploration and exploitation with information and randomization”. In: *Current Opinion in Behavioral Sciences* 38 (Apr. 2021), pp. 49–56. DOI: 10.1016/J.COBEHA.2020.10.001.

[160] Patrick Friedrich et al. “Imaging evolution of the primate brain: the next frontier?” In: *NeuroImage* 228 (Mar. 2021), p. 117685. DOI: 10.1016/J.NEUROIMAGE.2020.117685.

[161] Franz-Xaver Neubert et al. “Connectivity reveals relationship of brain areas for reward-guided learning and decision making in human and monkey frontal cortex.” In: *Proceedings of the National Academy of Sciences of the United States of America* 112 (20 May 2015), E2695–704. DOI: 10.1073/pnas.1410767112.

[162] Jérôme Sallet et al. "The organization of dorsal frontal cortex in humans and macaques." In: *The Journal of neuroscience : the official journal of the Society for Neuroscience* 33 (30 July 2013), pp. 12255–74. DOI: 10.1523/JNEUROSCI.5108-12.2013.

[163] Hiroshi Abe and Daeyeol Lee. "Distributed Coding of Actual and Hypothetical Outcomes in the Orbital and Dorsolateral Prefrontal Cortex". In: *Neuron* 70 (4 May 2011), pp. 731–741. DOI: 10.1016/j.neuron.2011.03.026.

[164] Benjamin Y. Hayden, John M. Pearson, and Michael L. Platt. "Fictive Reward Signals in the Anterior Cingulate Cortex". In: *Science* 324 (5929 May 2009), pp. 948–950. DOI: 10.1126/science.1168488.

[165] Elsie Premereur, Peter Janssen, and Wim Vanduffel. "Functional MRI in Macaque Monkeys during Task Switching". In: *Journal of Neuroscience* 38 (50 Dec. 2018), pp. 10619–10630. DOI: 10.1523/JNEUROSCI.1539-18.2018.

[166] Kiyoshi Nakahara et al. "Functional MRI of macaque monkeys performing a cognitive set-shifting task". In: *Science (New York, N.Y.)* 295 (5559 Feb. 2002), pp. 1532–1536. DOI: 10.1126/SCIENCE.1067653.

[167] Kristen A. Ford et al. "BOLD fMRI activation for anti-saccades in nonhuman primates". In: *NeuroImage* 45 (2 Apr. 2009), pp. 470–476. DOI: 10.1016/J.NEUROIMAGE.2008.12.009.

[168] Peter M. Kaskan et al. "Learned Value Shapes Responses to Objects in Frontal and Ventral Stream Networks in Macaque Monkeys". In: *Cerebral cortex (New York, N.Y. : 1991)* 27 (5 May 2017), pp. 2739–2757. DOI: 10.1093/CERCOR/BHW113.

[169] Jan Grohn et al. "Multiple systems in macaques for tracking prediction errors and other types of surprise". In: *PLOS Biology* 18 (10 Oct. 2020), e3000899. DOI: 10.1371/JOURNAL.PBIO.3000899.

[170] Alizée Lopez-Persem et al. "Differential functional connectivity underlying asymmetric reward-related activity in human and nonhuman primates". In: *Proceedings of the National Academy of Sciences of the United States of America* 117 (45 Nov. 2020), pp. 28452–28462. DOI: 10.1073/PNAS.2000759117/SUPPL_FILE/PNAS.2000759117.SAPP.PDF.

[171] Brian Lau and Paul W. Glimcher. "Value representations in the primate striatum during matching behavior". In: *Neuron* 58 (3 May 2008), pp. 451–463. DOI: 10.1016/J.NEURON.2008.02.021.

[172] Laurence T. Hunt et al. "Triple dissociation of attention and decision computations across prefrontal cortex". In: *Nature Neuroscience 2018 21:10* 21 (10 Sept. 2018), pp. 1471–1481. DOI: 10.1038/s41593-018-0239-5.

[173] Sébastien Ballesta and Camillo Padoa-Schioppa. "Economic Decisions through Circuit Inhibition". In: *Current Biology* 29 (22 Nov. 2019), 3814–3824.e5. DOI: 10.1016/J.CUB.2019.09.027.

[174] Scott Mackey and Michael Petrides. "Quantitative demonstration of comparable architectonic areas within the ventromedial and lateral orbital frontal cortex in the human and the macaque monkey brains". In: *The European journal of neuroscience* 32 (11 Dec. 2010), pp. 1940–1950. DOI: 10.1111/J.1460-9568.2010.07465.X.

[175] Robert R. Hampton, Aaron Zivin, and Elisabeth A. Murray. "Rhesus monkeys (Macaca mulatta) discriminate between knowing and not knowing and collect information as needed before acting". In: *Animal cognition* 7 (4 Oct. 2004), pp. 239–246. DOI: 10.1007/S10071-004-0215-1.

[176] Hsiao Wei Tu, Alex A. Pani, and Robert R. Hampton. "Rhesus monkeys (Macaca mulatta) adaptively adjust information seeking in response to information accumulated". In: *Journal of comparative psychology (Washington, D.C. : 1983)* 129 (4 Nov. 2015), pp. 347–355. DOI: 10.1037/A0039595.

[177] Marion Bosc et al. "Checking behavior in rhesus monkeys is related to anxiety and frontal activity". In: *Scientific Reports 2017 7:1* 7 (1 Mar. 2017), pp. 1–9. DOI: 10.1038/srep45267.

[178] E. Procyk, Y. L. Tanaka, and J. P. Joseph. "Anterior cingulate activity during routine and non-routine sequential behaviors in macaques". In: *Nature Neuroscience 2000 3:5* 3 (5 May 2000), pp. 502–508. DOI: 10.1038/74880.

[179] Stephen Ferrigno et al. "Recursive sequence generation in monkeys, children, U.S. adults, and native Amazonians". In: *Science Advances* 6 (26 June 2020). DOI: 10.1126/SCIADV.AAZ1002/SUPPL_FILE/AAZ1002_SM.PDF.

[180] Lea Roumazeilles et al. "Social prediction modulates activity of macaque superior temporal cortex". In: *Science advances* 7 (38 Sept. 2021). DOI: 10.1126/SCIADV.ABH2392.

[181] Maya Zhe Wang and Benjamin Y. Hayden. "Monkeys are curious about counterfactual outcomes". In: *Cognition* 189 (Aug. 2019), p. 1. DOI: 10.1016/J.COGNITION.2019.03.009.

[182] Benjamin Y. Hayden, John M. Pearson, and Michael L. Platt. "Neuronal basis of sequential foraging decisions in a patchy environment". In: *Nature Neuroscience 2011 14:7* 14 (7 June 2011), pp. 933–939. DOI: 10.1038/nn.2856.

[183] Nils Kolling et al. "Neural mechanisms of foraging." In: *Science (New York, N.Y.)* 336 (6077 2012), pp. 95–98. DOI: 10.1126/science.1216930.

[184] Gary A. Kane et al. "Increased locus coeruleus tonic activity causes disengagement from a patch-foraging task". In: *Cognitive, Affective and Behavioral neuroscience* 17 (6 Dec. 2017), pp. 1073–1083. DOI: 10.3758/S13415-017-0531-Y.

[185] Keisetsu Shima and Jun Tanji. "Role for cingulate motor area cells in voluntary movement selection based on reward". In: *Science (New York, N.Y.)* 282 (5392 Nov. 1998), pp. 1335–1338. DOI: 10.1126/SCIENCE.282.5392.1335.

[186] Steven W Kennerley et al. "Optimal decision making and the anterior cingulate cortex". In: *Nature Neuroscience* 9 (7 July 2006), pp. 940–947. DOI: 10.1038/nn1724.

[187] René Quilodran, Marie Rothé, and Emmanuel Procyk. "Behavioral Shifts and Action Valuation in the Anterior Cingulate Cortex". In: *Neuron* 57 (2 Jan. 2008), pp. 314–325. DOI: 10.1016/J.NEURON.2007.11.031/ATTACHMENT/645C49A4-31A0-477A-829C-7D07DBEA421B/MMC1.PDF.

[188] Céline Amiez et al. "Modulation of feedback related activity in the rostral anterior cingulate cortex during trial and error exploration". In: *NeuroImage* 63 (3 Nov. 2012), pp. 1078–1090. DOI: 10.1016/J.NEUROIMAGE.2012.06.023.

[189] Jascha Achterberg et al. "A One-Shot Shift from Explore to Exploit in Monkey Prefrontal Cortex". In: *Journal of Neuroscience* 42 (2 Jan. 2022), pp. 276–287. DOI: 10.1523/JNEUROSCI.1338-21.2021.

[190] Frederic M. Stoll et al. "The Effects of Cognitive Control and Time on Frontal Beta Oscillations". In: *Cerebral Cortex* 26 (4 Apr. 2016), pp. 1715–1732. DOI: 10.1093/cercor/bhv006.

[191] Emmanuel Procyk et al. "Midcingulate Motor Map and Feedback Detection: Converging Data from Humans and Monkeys". In: *Cerebral cortex (New York, N.Y. : 1991)* 26 (2 Feb. 2016), pp. 467–476. DOI: 10.1093/CERCOR/BHU213.

[192] Steven W. Kennerley and Jonathan D. Wallis. "Evaluating choices by single neurons in the frontal lobe: outcome value encoded across multiple decision variables". In: *The European journal of neuroscience* 29 (10 May 2009), p. 2061. DOI: 10.1111/J.1460-9568.2009.06743.X.

[193] Chung Hay Luk and Jonathan D. Wallis. "Choice Coding in Frontal Cortex during Stimulus-Guided or Action-Guided Decision-Making". In: *Journal of Neuroscience* 33 (5 Jan. 2013), pp. 1864–1871. DOI: 10.1523/JNEUROSCI.4920-12.2013.

[194] Ulrike Basten et al. "How the brain integrates costs and benefits during decision making". In: *Proceedings of the National Academy of Sciences of the United States of America* 107 (50 Dec. 2010), pp. 21767–21772. DOI: 10.1073/PNAS.0908104107/SUPPL_FILE/PNAS.200908104SI.PDF.

[195] Marios G. Philiastides, Guido Biele, and Hauke R. Heekeren. "A mechanistic account of value computation in the human brain". In: *Proceedings of the National Academy of Sciences of the United States of America* 107 (20 May 2010), pp. 9430–9435. DOI: 10.1073/PNAS.1001732107/SUPPL_FILE/PNAS.201001732SI.PDF.

[196] Philippe Domenech, Sylvain Rheims, and Etienne Koechlin. "Neural mechanisms resolving exploitation-exploration dilemmas in the medial prefrontal cortex". In: *Science* 369 (6507 Aug. 2020). DOI: 10.1126/SCIENCE.ABB0184/SUPPL_FILE/ABB0184_DOMENECH_SM.PDF.

[197] Mehdi Khamassi et al. "Behavioral Regulation and the Modulation of Information Coding in the Lateral Prefrontal and Cingulate Cortex". In: *Cerebral cortex (New York, N.Y. : 1991)* 25 (9 Sept. 2015), pp. 3197–3218. DOI: 10.1093/CERCOR/BHU114.

[198] Caroline I. Jahn et al. "Dual contributions of noradrenaline to behavioural flexibility and motivation". In: *Psychopharmacology* 235 (9 Sept. 2018), p. 2687. DOI: 10.1007/S00213-018-4963-Z.

[199] Dougal G.R. Tervo et al. "Behavioral variability through stochastic choice and its gating by anterior cingulate cortex". In: *Cell* 159 (1 Sept. 2014), pp. 21–32. DOI: 10.1016/j.cell.2014.08.037.

[200] Mark J. Buckley et al. "Dissociable components of rule-guided behavior depend on distinct medial and prefrontal regions". In: *Science (New York, N.Y.)* 325 (5936 July 2009), pp. 52–58. DOI: 10.1126/SCIENCE.1172377.

[201] Steven W. Kennerley, Timothy E.J. Behrens, and Jonathan D. Wallis. "Double dissociation of value computations in orbitofrontal and anterior cingulate neurons". In: *Nature Neuroscience 2011 14:12* 14 (12 Oct. 2011), pp. 1581–1589. DOI: 10.1038/nn.2961.

[202] Peter H. Rudebeck and Elisabeth A. Murray. "The orbitofrontal oracle: cortical mechanisms for the prediction and evaluation of specific behavioral outcomes". In: *Neuron* 84 (6 Dec. 2014), pp. 1143–1156. DOI: 10.1016/J.NEURON.2014.10.049.

[203] Alicia Izquierdo, Robin K. Suda, and Elisabeth A. Murray. "Bilateral Orbital Prefrontal Cortex Lesions in Rhesus Monkeys Disrupt Choices Guided by Both Reward Value and Reward Contingency". In: *Journal of Neuroscience* 24 (34 Aug. 2004), pp. 7540–7548. DOI: 10.1523/JNEUROSCI.1921-04.2004.

[204] Mark E. Walton et al. "Separable Learning Systems in the Macaque Brain and the Role of Orbitofrontal Cortex in Contingent Learning". In: *Neuron* 65 (6 Mar. 2010), pp. 927–939. DOI: 10.1016/J.NEURON.2010.02.027.

[205] Davide Folloni et al. "Ultrasound modulation of macaque prefrontal cortex selectively alters credit assignment–related activity and behavior". In: *Science Advances* 7 (51 Dec. 2021), p. 7700. DOI: 10.1126/SCIADV.ABG7700.

[206] Michael J. Tobia et al. "Neural systems for choice and valuation with counterfactual learning signals". In: *NeuroImage* 89 (Apr. 2014), pp. 57–69. DOI: 10.1016/J.NEUROIMAGE.2013.11.051.

[207] Doris Pischedda, Stefano Palminteri, and Giorgio Coricelli. "The Effect of Counterfactual Information on Outcome Value Coding in Medial Prefrontal and Cingulate Cortex: From an Absolute to a Relative Neural Code". In: *Journal of Neuroscience* 40 (16 Apr. 2020), pp. 3268–3277. DOI: 10.1523/JNEUROSCI.1712-19.2020.

[208] Vikram S. Chib et al. "Evidence for a Common Representation of Decision Values for Dissimilar Goods in Human Ventromedial Prefrontal Cortex". In: *Journal of Neuroscience* 29 (39 Sept. 2009), pp. 12315–12320. DOI: 10.1523/JNEUROSCI.2575-09.2009.

[209] Alizée Lopez-Persem, Philippe Domenech, and Mathias Pessiglione. "How prior preferences determine decision-making frames and biases in the human brain". In: *eLife* 5 (Nov. 2016). DOI: 10.7554/ELIFE.20317.

[210] Paul Christian Bürkner. "brms: An R Package for Bayesian Multilevel Models Using Stan". In: *Journal of Statistical Software* 80 (Aug. 2017), pp. 1–28. DOI: 10.18637/JSS.V080.I01.

[211] Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual*. 2021.

[212] Thorsten Kahnt et al. "Connectivity-Based Parcellation of the Human Orbitofrontal Cortex". In: *Journal of Neuroscience* 32 (18 May 2012), pp. 6240–6250. DOI: 10.1523/JNEUROSCI.0257-12.2012.

[213] Juan Carlos Cerpa, Alain R. Marchand, and Etienne Coutureau. "Distinct regional patterns in noradrenergic innervation of the rat prefrontal cortex". In: *Journal of chemical neuroanatomy* 96 (Mar. 2019), pp. 102–109. DOI: 10.1016/J.JCHEMNEU.2019.01.002.

[214] Nicola Palomero-Gallagher et al. "Receptor architecture of human cingulate cortex: evaluation of the four-region neurobiological model". In: *Human brain mapping* 30 (8 Aug. 2009), pp. 2336–2355. DOI: 10.1002/HBM.20667.

[215] Sabrina van Heukelum et al. "Where is Cingulate Cortex? A Cross-Species View". In: *Trends in neurosciences* 43 (5 May 2020), pp. 285–299. DOI: 10.1016/J.TINS.2020.03.007.

[216] Masud Husain and Jonathan P. Roiser. "Neuroscience of apathy and anhedonia: a transdiagnostic approach". In: *Nature reviews. Neuroscience* 19 (8 Aug. 2018), pp. 470–484. DOI: 10.1038/S41583-018-0029-9.

[217] Jacqueline Scholl et al. "Should I stick with it or move on? The effect of apathy and compulsivity on planning and stopping in sequential decision making". In: *PLoS Biology* (2022).

[218] Mathias Pessiglione et al. "How the brain translates money into force: A neuroimaging study of subliminal motivation". In: *Science* (2007). DOI: 10.1126/science.1140459.

[219] Florent Meyniel et al. "Neurocomputational account of how the human brain decides when to have a break". In: *Proceedings of the National Academy of Sciences of the United States of America* (2013). DOI: 10.1073/pnas.1211925110.

[220] Yael Niv et al. "Tonic dopamine: opportunity costs and the control of response vigor". In: *Psychopharmacology* 191 (3 Apr. 2007), pp. 507–520. DOI: 10.1007/S00213-006-0502-4.

[221] Andreas Klaus, Joaquim Alves Da Silva, and Rui M. Costa. "What, If, and When to Move: Basal Ganglia Circuits and Self-Paced Action Initiation". In: *https://doi.org/10.1146/annurev-neuro-072116-031033* 42 (July 2019), pp. 459–483. DOI: 10.1146/ANNUREV-NEURO-072116-031033.

[222] Marcel Brass and Patrick Haggard. "The what, when, whether model of intentional action". In: *Neuroscientist* 14 (4 Aug. 2008), pp. 319–325. DOI: 10.1177/1073858408317417.

[223] Okihide Hikosaka. "The habenula: from stress evasion to value-based decision-making". In: *Nature Reviews Neuroscience 2010 11:7* 11 (7 Apr. 2010), pp. 503–513. DOI: 10.1038/nrn2866.

[224] Satoko Amemori et al. "Microstimulation of primate neocortex targeting striosomes induces negative decision-making". In: *European Journal of Neuroscience* 51 (3 Feb. 2020), pp. 731–741. DOI: 10.1111/EJN.14555.

[225] Nima Khalighinejad et al. "Human decisions about when to act originate within a basal forebrain-nigral circuit". In: *Proceedings of the National Academy of Sciences of the United States of America* 117 (21 May 2020), pp. 11799–11810. DOI: 10.1073/PNAS.1921211117/-/DCSUPPLEMENTAL.

[226] Nima Khalighinejad et al. "A habenula-insular circuit encodes the willingness to act". In: *Nature Communications 2021 12:1* 12 (1 Nov. 2021), pp. 1–12. DOI: 10.1038/s41467-021-26569-1.

[227] Michael Esterman et al. "In the Zone or Zoning Out? Tracking Behavioral and Neural Fluctuations During Sustained Attention". In: *Cerebral Cortex* 23 (11 Nov. 2013), pp. 2712–2723. DOI: 10.1093/CERCOR/BHS261.

[228] Aurore San-Galli et al. “Primate Ventromedial Prefrontal Cortex Neurons Continuously Encode the Willingness to Engage in Reward-Directed Behavior”. In: *Cerebral Cortex* 28 (1 Jan. 2018), pp. 73–89. DOI: 10.1093/CERCOR/BHW351.

[229] Takafumi Minamimoto, Giancarlo La Camera, and Barry J. Richmond. “Measuring and Modeling the Interaction Among Reward Size, Delay to Reward, and Satiation Level on Motivation in Monkeys”. In: *https://doi.org/10.1152/jn.90959.2008* 101 (1 Jan. 2009), pp. 437–447. DOI: 10.1152/JN.90959.2008.

[230] Maarten A.S. Boksem, Theo F. Meijman, and Monicque M. Lorist. “Mental fatigue, motivation and action monitoring”. In: *Biological Psychology* 72 (2 May 2006), pp. 123–132. DOI: 10.1016/j.biopsycho.2005.08.007.

[231] Murray Sidman and William C. Stebbins. “Satiation effects under fixed-ratio schedules of reinforcement”. In: *Journal of Comparative and Physiological Psychology* 47 (2 Apr. 1954), pp. 114–116. DOI: 10.1037/H0054127.

[232] Fabien Vinckier et al. “Neuro-computational account of how mood fluctuations arise and affect decision making”. In: *Nature Communications 2018 9:1* 9 (1 Apr. 2018), pp. 1–12. DOI: 10.1038/s41467-018-03774-z.

[233] Richard E. Passingham, Sara L. Bengtsson, and Hakwan C. Lau. “Medial frontal cortex: from self-generated action to reflection on one’s own performance”. In: *Trends in Cognitive Sciences* 14 (1 Jan. 2010), pp. 16–21. DOI: 10.1016/J.TICS.2009.11.001.

[234] Benjamin Chew et al. “A Neurocomputational Model for Intrinsic Reward”. In: *Journal of Neuroscience* 41 (43 Oct. 2021), pp. 8963–8971. DOI: 10.1523/JNEUROSCI.0858-20.2021.

[235] Sebastien Bouret and Barry J. Richmond. “Ventromedial and Orbital Prefrontal Neurons Differentially Encode Internally and Externally Driven Motivational Values in Monkeys”. In: *Journal of Neuroscience* 30 (25 June 2010), pp. 8591–8601. DOI: 10.1523/JNEUROSCI.0049-10.2010.

[236] Caroline Jahn et al. “Strategic exploration in the macaque’s prefrontal cortex”. In: *bioarxiv* (2022). DOI: https://doi.org/10.1101/2022.05.11.491468.

[237] Nima Khalighinejad et al. “Complementary roles of serotonergic and cholinergic systems in decisions about when to act”. In: *Current Biology* (2022).

[238] Bolton K.H. Chau et al. “Contrasting Roles for Orbitofrontal Cortex and Amygdala in Credit Assignment and Learning in Macaques”. In: *Neuron* 87 (5 Sept. 2015), pp. 1106–1118. DOI: 10.1016/J.NEURON.2015.08.018.

[239] Francisca P. Leite et al. “Repeated fMRI Using Iron Oxide Contrast Agent in Awake, Behaving Macaques at 3 Tesla”. In: *NeuroImage* 16 (2 June 2002), pp. 283–294. DOI: 10.1006/NIMG.2002.1110.

[240] David L. Barack, Steve W.C. Chang, and Michael L. Platt. “Posterior Cingulate Neurons Dynamically Signal Decisions to Disengage during Foraging”. In: *Neuron* 96 (2 Oct. 2017), 339–347.e5. DOI: 10.1016/J.NEURON.2017.09.048.

[241] Davide Folloni et al. "Manipulation of Subcortical and Deep Cortical Activity in the Primate Brain Using Transcranial Focused Ultrasound Stimulation". In: *Neuron* 101 (6 Mar. 2019), 1109–1116.e5. DOI: 10.1016/J.NEURON.2019.01.019/ATTACHMENT/B07EC44A-BC11-4175-81B5-5AA77CD95500/MMC1.PDF.

[242] Lennart Verhagen et al. "Offline impact of transcranial focused ultrasound on cortical activation in primates". In: *eLife* 8 (Feb. 2019). DOI: 10.7554/ELIFE.40541.

[243] Jeffrey D. Schall, Thomas J. Palmeri, and Gordon D. Logan. "Models of inhibitory control". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 372 (1718 Apr. 2017). DOI: 10.1098/RSTB.2016.0193.

[244] Markus Ullsperger, Claudia Danielmeier, and Gerhard Jocham. "Neurophysiology of performance monitoring and adaptive behavior". In: *Physiological Reviews* 94 (1 Jan. 2014), pp. 35–79. DOI: 10.1152/PHYSREV.00041.2012/ASSET/IMAGES/LARGE/Z9J0011426740006.JPEG.

[245] Markus Ullsperger et al. "Neural mechanisms and temporal dynamics of performance monitoring". In: *Trends in Cognitive Sciences* 18 (5 May 2014), pp. 259–267. DOI: 10.1016/J.TICS.2014.02.009.

[246] Liya Ma et al. "Macaque anterior cingulate cortex deactivation impairs performance and alters lateral prefrontal oscillatory activities in a rule-switching task". In: *PLOS Biology* 17 (7 July 2019), e3000045. DOI: 10.1371/JOURNAL.PBIO.3000045.

[247] Frank Eblen and Ann M. Graybiel. "Highly restricted origin of prefrontal cortical inputs to striosomes in the macaque monkey". In: *Journal of Neuroscience* 15 (9 Sept. 1995), pp. 5999–6013. DOI: 10.1523/JNEUROSCI.15-09-05999.1995.

[248] Fumino Fujiyama et al. "Exclusive and common targets of neostriatofugal projections of rat striosome neurons: a single neuron-tracing study using a viral vector". In: *European Journal of Neuroscience* 33 (4 Feb. 2011), pp. 668–677. DOI: 10.1111/J.1460-9568.2010.07564.X.

[249] Mark E Walton et al. "What Is the Relationship between Dopamine and Effort?" In: *Trends in Neurosciences* 42 (2 Feb. 2019), pp. 79–91. DOI: 10.1016/J.TINS.2018.10.001.

[250] Brent A Vogt. *Architecture, cytology and comparative organization of primate cingulate cortex*. 2009.

[251] Alexander Friedman et al. "A Corticostriatal Path Targeting Striosomes Controls Decision-Making under Conflict". In: *Cell* 161 (6 June 2015), pp. 1320–1333. DOI: 10.1016/j.cell.2015.04.049.

[252] Dan Bang and Stephen M. Fleming. "Distinct encoding of decision confidence in human medial prefrontal cortex". In: *Proceedings of the National Academy of Sciences of the United States of America* 115 (23 June 2018), pp. 6082–6087. DOI: 10.1073/PNAS.1800795115/-/DCSUPPLEMENTAL.

[253] Marco K Wittmann et al. "Predictive decision making driven by multiple time-linked reward representations in the anterior cingulate cortex." In: *Nature communications* 7 (Aug. 2016), p. 12327. DOI: 10.1038/ncomms12327.

[254] Nils Kolling et al. “Value, search, persistence and model updating in anterior cingulate cortex”. In: *Nature Neuroscience 2016 19:10* 19 (10 Sept. 2016), pp. 1280–1285. DOI: 10.1038/nn.4382.

[255] Simone G. Shamay-Tsoory and Avi Mendelsohn. “Real-Life Neuroscience: An Ecological Approach to Brain and Behavior Research”. In: *Perspectives on psychological science : a journal of the Association for Psychological Science* 14 (5 Sept. 2019), pp. 841–859. DOI: 10.1177/1745691619856350.