Neural mechanisms of reward-guided behavioural flexibility

Sam Hall-McMaster

DPhil Thesis

Department of Experimental Psychology

Magdalen College, University of Oxford

Hilary Term, 2020



Neural mechanisms of reward-guided behavioural flexibility

Sam Hall-McMaster

Magdalen College, Oxford **Hilary Term, 2020**

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy at the University of Oxford

Abstract

Pursuing goals requires us to be flexible. When engaged in goal-directed behaviour, such flexibility is guided by the prospect of reward. The first part of this thesis examines how reward prospect influences flexible updating of task rules, needed to switch between tasks. Chapter 2 investigates how sequential changes in reward prospect influence rule updating across three behavioural experiments. This chapter shows inconclusive evidence for the hypothesis that sequential increases in reward prospect promote flexible rule updating. Chapter 3 examines the neural representation of task rules using multivariate pattern analyses of electroencephalographic data. This shows that high reward prospect boosts the neural encoding of task rules on trials where rules must be flexibly updated but not on trials where rules are repeated. The second part of this thesis examines how reward prospect is used to guide decisions about when to switch from option to another. Chapter 4 examines the theoretical implications of using reward properties of specific alternatives to guide switch decisions. Simulations show that decisions to switch based on specific alternatives tend to produce earlier switch decisions and higher reward overall than switch decisions based on average reward rates. Chapter 5 uses multivariate pattern analyses of electroencephalographic data to investigate how people decide when to switch to a specific option, in the absence of direct sensory cues to indicate its changing value. This reveals preliminary evidence that people track the hidden reward prospect for switching and use this information to regulate switch behaviour. Chapter 6 discusses implications of these results and bridges across cognitive control and decision making studies.

This work was funded by scholarships from the Rutherford Foundation and William Georgetti Trusts of New Zealand.

Approximate word count: 50,000 words.

Acknowledgements

I can think of no better team to supervise this thesis than Nick Myers and Mark Stokes. Thank you Nick for spending countless hours fielding my questions, reading countless drafts, offering snacks and looking out for my mental health during this DPhil. Thank you Mark for letting me pursue my wacky reward ideas in the lab, providing a much needed critical eye to bring my ideas under tighter experimental control and your reassurance throughout this DPhil.

Thank you to Fabrice Luyckx and Paul Muhle-Karbe for unofficially supervising this research and your friendship. Thank you to all members of the Stokes lab, especially Frida Printzlau and Dante Wasmuht for being such incredible DPhil mates. We've climbed this mountain together.

Thank you to those who have shared this DPhil adventure with me, especially Josh Calder-Travis, Alex Temple-McCune, Keno Jüchems, Michael Wolff, Sage Boettcher, Giedrė Čepukaitytė, Payal Bhavsar, Wenjun Huang, Mirjam Hazenbosch, Victoire Déjean, Aaron Singh Hundle and Naveen George. Thank you to the FMRIB course tutors for welcoming me into the team, especially Frederik Lange, Istvan Huszar, Amy Howard and Ludo Griffanti.

Thank you to the Rutherford Foundation and William Georgetti Trusts of New Zealand for funding this research and allowing me to undertake it in this great bastion of science.

Thank you to my family - Suzanne, Ken, Grace, Nana and Sylvie - for making me feel so close despite being so far from home. Thank you for all your love during the tough personal events that coincided with this DPhil. This thesis is for you.

Finally, thank you to Maud Grol for your unparalleled support. I arrived in Oxford to chase scientific discovery but I will leave having discovered something - and someone - much more valuable.

Statement of Contribution

Science is an intrinsically collaborative endeavour and this thesis would not have been possible without the help of scientists inside and outside my supervisory team. The task switching studies in chapters 2 and 3 of this thesis were designed in collaboration with Dr. Paul Muhle-Karbe, Dr. Nicholas Myers and Associate Professor Mark Stokes. In addition to my role in developing these designs, I collected the data for these projects and took the lead on analysis and write up. Analyses and write up were guided through expert feedback and assistance from Paul, Nick and Mark. The theoretical review presented in chapter 4 was developed from a brief review of foraging studies I submitted in my Transfer of Status report. This initial writing was developed through discussions with Fabrice Luyckx. Knowing that this work would be included in my thesis, I took the lead on writing it up for publication and creating simulations that were added to the chapter. The subsequent electroencephalographic experiment in chapter 5 was designed together with Nick Myers and Mark Stokes, including initial discussions with Dr. Nils Kolling and Dr. Laurence Hunt. I collected the data for this project and have taken the lead on analysis and writing, with on-going guidance from Nick and Mark.

Table of Contents

Chapter 1. General introduction 1
Chapter 2. Sequential effects of reward on flexible control 31
Experiment 2.1 36
Experiment 2.2 46
Experiment 2.3 57
Chapter 3. Reward boosts neural coding of task rules to optimise flexible control 70
Experiment 3.1 72
Experiment 3.2 111
Chapter 4. Theoretical influence of option-specific properties on switch decisions 124
Simulations 133
Chapter 5. Neural encoding of hidden reward information and switch decisions 152
Experiment 5.1 153
Chapter 6. General discussion 182
References 191

Chapter 1. General introduction

Imagine crossing the street on your way to work when all of sudden a car appears out of nowhere. Do you sprint across the road or jump back towards the curb? What about deciding whether to pursue an academic path or pivot in a new direction? You did always dream about opening your very own bagel shop after all. The pursuit of goals, whether these are small or big, involves being flexible in how we control our behaviour. To make it to our destination, we need to react flexibly to incoming sensory information. To make weighty decisions about our careers, we need to evaluate different possibilities and be willing to chart a new course. This thesis investigates how motivation for good outcomes shapes the flexible control of behaviour.

What is behavioural flexibility?

Behavioural flexibility can take many forms. For instance, it might involve switching between tasks or it might involve switching between different choices. Different forms of flexibility have been investigated in distinct research traditions. One such tradition examines cognitive control, which refers to a broad collection of mental processes that direct thought and action in the service of goals (Amer et al., 2016; Braver, 2012). Mental processes that contribute to cognitive control include working memory (Jimura et al., 2010), impulse inhibition (Leotti & Wager, 2010) and switching between relevant tasks (Kleinsorge & Rinkenauer, 2012). One thing that is clear from the breadth of these processes is that flexibility can be considered from multiple angles. From a working memory perspective, flexibility might involve updating items held in mind or transforming items for upcoming action. From an inhibitory control perspective, flexibility might involve halting a motor response. The cognitive control studies in this thesis will focus on task switching as a marker

of behavioural flexibility. Typical task switching studies require participants to alternate between different cognitive operations, such as evaluating the magnitude or parity of a number target. One characteristic effect in such tasks is that switching between the two operations comes at a behavioural cost. Speed and accuracy are typically worse on trials where the rule must be flexibly updated relative to the previous trial (switch trials), compared with trials where the same rule is used relative to the previous trial (repeat trials). The difference in performance between switch and repeat trials (the switch cost), can therefore serve as an index of flexible behavioural control, where switch cost reductions indicate increased flexibility (Aarts et al., 2010; Fröber & Dreisbach, 2016; Kleinsorge & Rinkenauer, 2012; Shen & Chun, 2011).

Another tradition of investigating flexibility comes from foraging, a subfield of decision making research based on Optimal Foraging Theory (Stephens & Krebs, 1986). In this area, flexibility has been viewed as disengaging from a current option to search for alternative options (Kolling et al., 2012), where options tend to be items that can be selected to gain reward. The main difference between foraging and task switching is that foraging is interested in *when* people choose to initiate flexible behaviour, as opposed to rapid adjustments in behaviour based on explicit cues. Foraging also examines behavioural flexibility within a sequence of choices, as opposed to more independent trial-wise adjustments in task switching. Two distinct forms of choice have been used to examine flexibility in foraging studies. Stay-switch or patch leaving choices involve exploiting an option and deciding when to abandon it, as its reward rate depletes over time. Accept-reject choices involve deciding whether to accept an option and gain its reward, or whether to reject it, so that a more rewarding option might be encountered (Stephens & Krebs, 1986). While the foraging research in this thesis will primarily focus on stay-switch or patch-leaving

choices, a critical aspect across all foraging paradigms is that decisions to initiate flexible behaviour depend on the prospect of reward. Current options that offer a high reward prospect, relative to the alternatives, should drive option engagement (persistence). Current options that offer a low reward prospect, relative to the alternatives, should drive option disengagement (flexibility).

The cognitive control and decision making traditions outlined above provide two defining characteristics of flexible behavioural control that will be investigated in this thesis. Cognitive control research highlights the idea that flexible behavioural control often involves updating contextual information, such as new rules needed to guide behaviour (chapters 2 and 3). Decision making research highlights the idea that flexible behavioural control involves monitoring and evaluating the reward prospect for maintaining or adapting behaviour (chapters 4 and 5).

Which neural processes give rise to flexible behaviour?

We have now established two defining characteristics of behavioural flexibility: updating contextual information needed to switch to a new task and the evaluation of reward prospect for maintaining or adjusting behaviour. How are these functional characteristics instantiated in the brain?

Current evidence indicates that switching between tasks depends on a network of frontoparietal brain regions. A recent meta-analysis pooled together sixty functional neuroimaging studies, fifty of which contrasted task switches with task repeats, four of which contrasted task switches with rest, and six of which contrasted task switches with a sensorimotor control task (Worringer et al., 2019). The analysis identified nine frontoparietal brain regions that showed increased activation in response to switching

across studies. These included the middle and inferior frontal gyri, inferior frontal junction, anterior insula, dorsal premotor cortex, pre-supplementary motor area, intraparietal sulcus and precuneus.

Within this network, lateral PFC regions, such as the middle frontal gyrus, may be especially critical for maintaining context information in preparation for upcoming demands. Neurons in lateral PFC are sensitive to information about task rules (Asaad et al., 2000; Miller & Cohen, 2001) and lateral PFC lesions have been shown to impair actions that depend on rule cues (Milner, 1982). In addition, mid-dorsolateral PFC is recruited when processing task rules that show one level of context-dependence, with more complex rule dependency extending more rostrally along the PFC (Badre, 2008; Badre & Nee, 2018). Multi-area recordings from monkeys also highlight the importance of lateral PFC in using task rules to determine appropriate behaviour, indicating that lateral PFC and the lateral intraparietal area receive task information from sensory areas and transform it into choice signals that are propagated to motor areas (Siegel et al., 2015). When lateral PFC is lesioned in monkeys, the ability to maintain current rule information over delays is diminished (Buckley et al., 2009). Lesion studies in humans have indicated that lateral PFC is also critical for flexibly switching between rules. Lesions to the left lateral PFC increase RT switch costs (Aron et al., 2004; Mayr et al., 2006), as well as the number of perseverative errors on the Wisconsin Card Sorting Task (Barceló & Knight, 2002). Together, these studies suggest that to switch tasks, lateral PFC is a critical region where rule information should be updated. This updating process may depend on midbrain dopamine neurons, which are thought to provide a gating signal that allows new rule information to pass from sensory areas into PFC (Braver & Cohen, 2000; Ott & Nieder, 2019; Yee & Braver, 2018).

From a theoretical perspective, the increases in RT switch costs and perseveration shown in lesion studies (Aron et al., 2004; Barceló & Knight, 2002; Mayr et al., 2006), as well as the meta-analysis by Worringer et al. (2019) provide evidence towards a task reconfiguration model of rule switching, where top-down control from frontoparietal brain regions reconfigures the neural system for current task demands (Meiran, 1996). However, this active reconfiguration process is not the only possibility for the apparent behavioural cost of rule switching. Another major theoretical account is that the previous rule interferes with new rule information (Allport et al., 1994). Neural evidence for this interference account has shown that neural activity related to the irrelevant rule can persist many trials after switching to a new task and the extent of this persistence shows a positive association with switch costs (Yeung et al., 2006). A final possibility is that rule switching costs arise from bottom-up processes involved in processing new rule cues (Logan & Bundesen, 2003). However, neural evidence suggests this view is less likely, indicating that frontoparietal brain regions activated when preparing to switch to a new task are distinct from regions involved in cue encoding (de Baene & Brass, 2011). To summarise, even after decades of research, it is still unclear whether behavioural switch costs reflect an active reconfiguration process or interference between competing rules (or both).

This might explain why contemporary studies of rule-dependent processing have turned towards a new set of questions, which consider how rule information is *coded*. The main difference between the coding approach and traditional analysis approaches is its focus on the information contained within patterns of neural activity as opposed to average differences in activation (Kriegeskorte et al., 2006). To probe neural coding of specific task variables, common forms of analysis include representational similarity analysis (RSA), which measures how similar neural patterns are (Kriegeskorte et al., 2008) and pattern

classification, which tests how well condition labels can be predicted from neural activity patterns (Haxby et al., 2014). Such methods have allowed more fine-grain measurement of how rule-specific activity changes under different conditions, which might not be detected as global activation changes. For example, studies using pattern decoding have shown that rule presentation triggers dynamic trajectories of PFC activity (Stokes et al., 2013) and that high reward conditions or more complex rules result in more precise rule coding within frontoparietal regions (Etzel et al., 2016; Woolgar et al., 2015).

Despite these insights about the properties of rule coding, studies examining rule coding on switch and repeat trials are surprisingly sparse. One study found that rule coding in inferior frontal sulcus, intraparietal sulcus and occipito-temporal cortex was strongest on switch trials and decreased as a function of rule repeats (Waskom et al., 2014). These findings might be seen as consistent with the notion that rule coding is made more precise to reduce interference with the previous task rule. It could also be that an active rule representation is only needed during switch trials, with less prominent rule representations coded in other brain regions being used as responses become more automatic. In contrast, another study found weaker rule coding on switch trials across a range of frontoparietal brain regions, suggesting more stable coding of rule information when tasks are repeated (Qiao et al., 2017). The study also showed that switch trials were marked by an increase in pattern dissimilarity between current and previous trials, which was interpreted as a dynamic recoding of task rule information (Qiao et al., 2017). While the results from these two studies appear to be at odds, Qiao et al. (2017) suggest that this might be due to differences between the tasks. Waskom et al.'s (2014) task involved three consecutive trials per rule making the timing of rule switches predictable. Participants also had 10s to encode the new rule at the start of each mini-block. Qiao et al. (2017) propose that this long

preparation interval could provide the conditions for strong rule reconfiguration, with rule coding becoming weaker as the task performance becomes more automatic. In contrast, Qiao et al.'s (2017) task involved unpredictable task switches and shorter preparation times, which could lead to a mixture of rule information on switch trials and a reduction in decoding accuracy. Thus, while it might be possible to reconcile these recent results, how rule coding changes to support behavioural flexibility is still far from clear.

Having considered flexibility from the task switching perspective, let us turn to flexibility from the foraging perspective. Recall that in foraging, flexibility can be viewed as choosing to disengage from a current option (Kolling et al., 2012; Kolling, Wittmann, et al., 2016). Most foraging work in neuroscience has focused on dorsal anterior cingulate cortex (dACC) as a key region in decisions to switch away from a current option. One influential study to investigate this was performed by Hayden et al. (2011), who recorded responses from single neurons in monkey dACC during a classic stay-switch (patch-leaving) task. In this experiment, monkeys made a choice between staying or leaving on each trial. A decision to stay meant the monkey received a juice reward, which was decreased by 0.02ml for each successive stay decision. The experiment also manipulated the 'travel time' between leaving an option and arriving at a new one, introducing a cost of leaving that could be factored into the reward prospect for staying or leaving. Hayden et al. (2011) found that the firing rate of dACC neurons increased with more time exploiting a current option and reached a similar threshold when deciding to leave, regardless of when the leave decision was made. When subdividing firing rates as a function of travel time, it was found that firing rates on leave trials scaled with the time cost for leaving that option. These results led the authors to propose that dACC firing rates might provide a threshold for when to switch away from an option and that this threshold can be adjusted depending on the cost of leaving.

Soon after these results, another influential paper on dACC's role in disengaging from an option emerged. Using an accept-reject design, Kolling and colleagues (2012) devised a task in which human participants were presented with two items and decided whether to engage further with the items or reject them. When deciding to engage further, the participants had a binary choice between the items, which involved evaluating the items' reward magnitudes (learned during training) and reward probabilities (shown on screen). When deciding to reject the items, two new items were drawn from a set of six alternatives shown on screen. Kolling et al. (2012) found that the blood oxygen level-dependent (BOLD) signal in dACC showed positive modulation in response to higher average values across the set of alternatives (search values). This might suggest that dACC activity is not setting a threshold for abandoning an option per se (see Hayden et al. 2011), but signalling the prospective reward for a leave or reject decision. On trials where the current items were rejected, dACC activity showed negative modulation in response to the size of a reward cost that needed to be paid to make the reject decision (Kolling et al., 2012), consistent with previous results showing cost sensitive dACC activity (Hayden et al., 2011). In contrast to dACC, the BOLD signal in ventromedial PFC (vmPFC) did not track the search value during choices about whether engage, but did track the expected value (reward magnitude x probability) of items selected during binary choice (Kolling et al., 2012). The main interpretation for the results in this study was that dACC helps to regulate decisions about whether to reject an option, through signalling the average reward prospect for switching to a new option. This interpretation has been the subject of intense debate (Ebitz & Hayden, 2016), with competing proposals that higher dACC signals might actually reflect choice difficulty (Shenhav et al., 2014) and that dACC and vmPFC might both encode information about the value of searching over engaging with current options (Shenhav,

Straccia, et al., 2016). This thesis is not about resolving that debate. Nonetheless, some of its additional details will be mentioned in the section on theoretical models in this introduction and in chapter 4.

Contemporary foraging studies have continued to provide insight about the neural mechanisms of stay-switch decisions. For instance, when people decide whether to keep exploiting an option or switch to another with a constant reward for a fixed number of trials, activity in dACC, frontal operculum (FO) and ventral striatum (VS) tracks reward information at multiple timescales. Specifically, activity in these regions was found to be modulated by the instantaneous reward on the previous trial and the average reward rate across trials, pieces of information that could be used to infer whether an option was getting better or worse over time (Wittmann et al., 2016). As this study demonstrates, dACC is not the only region that contributes to switch decisions. This was also made clear in a study by Barack et al. (2017), which used the same task as Hayden et al. (2011) but recorded from neurons in monkey posterior cingulate cortex (PCC). Doing so revealed that firing rates of PCC neurons increased in the seconds leading up to a switch response, especially for options that provided lower reward rates during exploitation. The authors interpreted these findings with caution, suggesting these could indicate PCC helps to motivate switching, that PCC represents the value of alternative courses of action or that it tracks choice difficulty. Contemporary foraging studies have also used causal manipulations during switch decisions, highlighting the contributions of subcortical areas. In particular, slow acting pharmacogenetic activation of the locus coeruleus (LC) in rats resulted in significantly earlier patch-leaving decisions (Kane et al., 2017), whereas fast acting optogenetic activation of neurons in the dorsal raphae nucleus (DRN) in mice resulted in more nose pokes towards a current option before switching (Lottem et al., 2018). The LC and DRN are dense sites for

norepinephrine and serotonin-secreting neurons respectively, suggesting that neuromodulatory systems are causally involved in decisions to switch away from an option.

To summarise, task switching studies have shown that a network of frontoparietal brain areas involved in cognitive control gets recruited when switching from one task rule to another. Within this network, lateral PFC might be a key region where task rules are represented and updated to guide upcoming behaviour. In contrast, foraging studies have implicated more medial brain regions such as dACC and PCC, as well as neuromodulatory nuclei in decisions to switch away from an option.

How does motivation for good outcomes influence behavioural flexibility?

This thesis will focus on reward motivation, a form of motivation that has been widely studied across species and can be manipulated experimentally with relative ease. Reward motivation can be defined as the influence of prospective reward on orienting towards goal-relevant information and invigorating goal-directed action (Botvinick & Braver, 2015). It is often manipulated in experimental settings by changing the incentives for fast and accurate performance. In Etzel et al. (2016), for example, a reward prospect cue was presented at the beginning of each trial to signal whether reward could be gained for a fast and accurate response. The difference in performance as a function of reward prospect is then typically used as a window into how motivation influences neural processing (Etzel et al., 2016; Hall-McMaster et al., 2017; Padmala & Pessoa, 2011).

Theoretical models of reward-guided behavioural flexibility

Dual Mechanisms of Control

Reward prospect tends to promote greater task preparation, an effect formalised in the Dual Mechanisms of Control (DMC) model (Braver, 2012). In this model, cognitive control can be parsed into proactive and reactive control processes. Proactive processes involve the representation of goal-relevant information in preparation for upcoming cognitive demands. Such proactive control can be deployed in many classic cognitive control tasks, where the current task rule can be actively maintained prior to a target stimulus (Etzel et al., 2016). In contrast, reactive control processes respond to cognitive demands when they arise. For example, reactive control would be triggered when participants cannot prepare for a target and must respond to incongruent information when it is presented (Chiew & Braver, 2016) or when preparation is invalid and participants must react to unexpected stimulus demands (Engelmann et al., 2009; Poh et al., 2019). Under the DMC model, proactive and reactive control can be distinguished by different neural response profiles in lateral PFC. Proactive control is associated with sustained activity in lateral PFC in preparation for a target response (Burgess & Braver, 2010). In contrast, reactive control is associated with transient activity in lateral PFC following target presentation, needed to resolve stimulus interference (Burgess & Braver, 2010; D'Esposito et al., 1999). Alongside the time course of proactive and reactive control, the DMC model posits that dopamine is preferentially important for proactive control processes. Specifcally, phasic dopamine is proposed to provide a necessary gating signal that enables contextual rule information to be encoded and maintained in PFC in preparation for the upcoming task (Braver & Cohen, 2000). Whether a proactive or reactive control mode is deployed has been shown to depend on the prospect of reward for fast and accurate performance (Etzel et al., 2016). When

participants stand to gain more reward for their performance, proactive neural activity is enhanced in lateral PFC (Etzel et al., 2016), presumably helping to streamline sensorimotor processes that make target responses more precise and efficient.

Dopamine Gating

As alluded to when describing the Dual Mechanisms of Control framework (Braver, 2012), a central neural proposal about the influence of reward prospect on cognitive control is that dopamine can modulate the encoding of goal-relevant information in PFC. We will refer to this as the Dopamine Gating Model (Braver & Cohen, 2000). The model proposes that PFC maintains information about task context in an attractor state, characterised by a stable pattern of recurrent excitation, which is resistant against distracting input. Reward predicting cues trigger the release of phasic dopamine, allowing new context information to pass from sensory areas to PFC neurons, through dopamine's potentiation of afferent connections to PFC and inhibitory connections between PFC units with that are selective to different contexts. The Dopamine Gating Model therefore provides a simple neural mechanism for how reward prospect could help update task information needed for flexible behaviour.

In the two decades since publication of the Dopamine Gating Model (Braver & Cohen, 2000), substantial work has deepened our understanding about dopamine's contribution to reward-guided behavioural control. Current theory proposes that dopamine has three important roles (Ott & Nieder, 2019), which include gating the passage of sensory and context information into PFC (D'Ardenne et al., 2012), maintaining and manipulating information based on current goals (Ott et al., 2014), and sending information to pre-motor circuits (Vijayraghavan et al., 2016). The gating provided by phasic dopamine release to PFC is still proposed to occur in response to reward predicting cues, from neurons that project

primarily from the ventral tegmental area to medial frontal regions, such as anterior cingulate cortex (Ott & Nieder, 2019). However, a second phasic dopamine pathway is now recognised, which runs from dorsolateral and ventrolateral substantia nigra to lateral PFC regions (Mark Williams & Goldman-Rakic, 1998) and shows sensitivity to information that is relevant for upcoming behaviour (Matsumoto & Takada, 2013). The gating mechanism itself is thought to occur through dopamine's neuromodulatory action on PFC neurons, which has the effect of increasing the signal-to-noise ratio during information coding (Jacob et al., 2013).

A final development since the early model of dopamine gating (Braver & Cohen, 2000) has been the idea that different dopamine receptors may have distinct effects on flexible and stable cognitive processing. For example, it has been proposed that prefrontal dopamine primarily stabilises encoded information through D1 receptor activation, whereas striatal dopamine primarily supports more flexible control processes, such as rule updating, through D2 receptor activation (Cools & D'Esposito, 2011). This increase in striatal DA could serve a gating role through altering functional coupling between PFC and posterior cortical areas (van Schouwenburg et al., 2010). While this proposal (Cools & D'Esposito, 2011) has been formulated - at least in part - on the basis of more tonic dopamine manipulations, it still provides a relevant model that could help to understand how phasic dopamine signals influence flexible control. Specifically, it leads to the prediction that dopamine gating of new rule information into PFC might occur through intermediary actions on striatal D2 receptors, as opposed to more direct actions on PFC neurons, which could help to stabilise information once updated. To summarise, our understanding about dopamine's role in reward-guided behavioural control has been greatly advanced in the past two decades. Nonetheless, contemporary theoretical accounts still hold a key role for dopamine in updating context

information, in response to reward predicting cues and task-relevant information (Ott & Nieder, 2019).

Dual Competition Model

Like the Dual Mechanisms of Control model (Braver, 2012), the Dual Competition Model (Pessoa, 2017) aims to provide a theoretical framework that accounts for the effect of reward prospect on cognitive control. The model draws inspiration from Guided-Activation Theory (Miller & Cohen, 2001), an influential theory of PFC function which proposes that PFC biases sensorimotor processing in accordance with behavioural goals. The Dual Competition Model brings this idea to the area of reward-motivated cognition, proposing that reward motivation helps to bias processing at perceptual and executive levels, towards behaviours that garner reward. The first manner in which reward is proposed to achieve this is through recruiting neurons from a common pool of resources. Neurons involved in attention and cognitive control are recruited to process task-relevant information, leaving less of the common resource available to process task-irrelevant information. The second manner in which reward is proposed to influence cognitive control is through increasing the strength of neural signals that carry task-relevant information. In other words, the model proposes that high reward motivation helps to direct what information is encoded in common resource neuronal populations and how strongly that information is encoded. Both the Dual Competition model (Pessoa, 2017) and the Dual Mechanisms of Control model (Braver, 2012) predict stronger task preparation under high reward conditions. However, the Dual Competition Model (Pessoa, 2017) appears to be a more general framework that could be applied to executive demands other than task preparation. In addition, the Dual Competition Model emphasises interactions between

networks of brain regions involved in attention, perception and evaluation, as opposed to making predictions about specific brain regions such as lateral PFC (Braver, 2012).

Expected Value of Control

The Dual Mechanisms of Control and Dual Competition models focus predominantly on cognitive control functions, such as selective attention, working memory and task switching, with little focus on decision making. One cognitive framework known as the Expected Value of Control (EVC) model seeks to bridge across both domains (Shenhav et al., 2013). The EVC model has its roots in an early model of dACC function, which posits that dACC monitors task conflict, helping to signal when cognitive control is needed to resolve task demands (Botvinick et al., 2001). Building on this, the EVC model attempts to provide an optimal solution to the question of how an agent should decide what task to perform or what decision to make and at what level of behavioural performance, given the rewards at stake and amount of cognitive processing required. This framework posits that dACC computes the EVC as the value of performing a given task at a given intensity minus the cognitive cost needed to perform it. The candidate EVC signal with the highest value is then selected by dACC to determine which task or action will be performed and with what level of cognitive control. Once the most rewarding control plan is specified, this information is passed from the dACC to other control regions, such as lateral PFC, to implement control over lower-level target information.

Behavioural Flexibility Model

In contrast to the EVC model, it has been proposed that dACC encodes critical information for deciding whether to maintain or adjust behaviour, above and beyond control allocation (Kolling, Wittmann, et al., 2016). We will refer to this as the Behavioural Flexibility Model. The model is based on four main lines of evidence. First, dACC has been

shown to encode the expected value of switching away from a current option, to search for alternative options (Kolling et al., 2012). This 'search value' is usually considered to be the average value over the alternative options. Chapter 4 will discuss situations where decisions to adjust behaviour might not be guided by average value. Second, dACC has been shown to encode information about average reward rate, which can be used to determine whether the reward gained from maintaining behaviour is increasing or decreasing over time (Wittmann et al., 2016). Third, dACC activity is stronger in response to errors and negative feedback than to positive outcomes (Holroyd et al., 2004), which could help to signal the need for behavioural adjustment. Fourth, ACC more broadly has been shown to respond when participants need to update their internal model about likely task events, such as the distribution of saccade locations on a screen, to guide new behavioural responses (O'Reilly et al., 2013). This study did not observe an increase in ACC activity in response to unexpected saccade locations (O'Reilly et al., 2013), suggesting that ACC activity does not always have a clear mapping with increased cognitive demands. The idea that dACC might not signal control allocation as its primary function has been further substantiated with a biophysical model of dACC (Kolling, Wittmann, et al., 2016). This biophysical model contains three populations of neurons, which encode information about the current option, the search value and the cost of switching (such as delays to locate a new option). The populations have auto-excitatory connections, as well as mutual inhibitory connections from the current option and cost populations to the search value population. Critically, despite having no explicit information about cognitive control demands, the model predicts increasing dACC activity with increasing search values across the trial, but with increasing choice difficulty late in the trial. This arises due to the fact that more difficult choices, which involve current values and search values that are more similar, lengthen the comparison

process; it takes longer for one population to 'win out' by entering a stable activity state that is higher than the other populations. This provides initial evidence that associations between cognitive control demands and dACC activity could arise as a by-product of decision making computations about whether to maintain or adjust behaviour (Kolling, Wittmann, et al., 2016).

Adaptive Gain Theory

Another prominent neural proposal that links prospective reward and behavioural control is Adaptive-Gain Theory (Aston-Jones & Cohen, 2005). This theory focuses on the locus coeruleus (LC), which contains norepinephrine-secreting neurons that project widely across cortex and serve an important role in regulating arousal. The theory proposes that LC can shift behavioural control between two distinct modes based on phasic or tonic modes of activity. The exploitative mode, associated with phasic LC activity, optimises information processing for the current task, selectively increasing neural gain in response to task relevant information. The explorative mode, associated with tonic LC activity, does not show such selectivity, resulting in increased distractibility and poorer task performance. Aston-Jones & Cohen (2005) propose that this may provide a mechanism through which agents can explore (i.e. switch to) new behavioural alternatives. The LC Adaptive-Gain Theory differs from dopamine gating models (Braver & Cohen, 2000; Ott & Nieder, 2019) in that its major predictions are about how arousal and attention are regulated over longer timescales. Whereas phasic dopamine gating applies on a rapid trial-by-trial basis, phasic LC responses are proposed to depend on both immediate reward (on the scale of seconds during a trial) and on longer-term reward (on the scale of minutes during a block). When longer-term reward is low, LC will be tonic even when immediate reward is high. When longer-term reward prospect is high, LC will show phasic activity when immediate reward is

high. However, the largest phasic LC responses are predicted when longer-term reward is high but immediate reward is low, helping to restore performance on the current task when it dips. This differs from an early model of dopamine gating (Braver & Cohen, 2000), which predicts highest phasic dopamine responses and thus best task performance when immediate reward is highest. The Adaptive Gain Theory also differs from dopamine gating models in that LC activity is not regulated through reward information directly but through reward sensitive regions of frontal cortex. OFC and ACC are proposed to provide top-down regulation of LC based on reward prospect and external reward outcomes. If a task carries high longer-term reward, OFC and ACC signals induce different magnitudes of phasic LC activity depending on immediate reward, to help prioritise task-relevant information. If a task carries low longer-term reward, OFC and ACC signals will induce tonic LC activity, reducing task performance and promoting more exploratory behaviour.

Empirical insights into reward-guided behavioural flexibility

Reward prospect increases task preparation

One consistent finding is that reward prospect tends to enhance task preparation, in situations where it is possible to encode relevant contextual information prior to target presentation. Recall that task preparation constitutes proactive cognitive control in the Dual Mechanisms of Control framework (Braver, 2012). The relationship between reward-driven task preparation and behavioural flexibility is nuanced and requires clarification up front. Theoretical accounts, such as the Dopamine Gating Model (Braver & Cohen, 2000), propose that high reward prospect will promote stronger instantiation of current context information. This would be expected to benefit behavioural flexibility when contextual information allows for valid preparation. For example, when switching to a new task,

increased proactive encoding of the new task rule would be expected to improve switching performance. In contrast, increased preparation would be expected to impair reactive control when preparation is unexpectedly invalid. For example, increased processing of a context cue would be expected to generate more incorrect responses when a low probability target is presented with a different response mapping to a more probable target (Servan-Schreiber et al., 1996). While this division makes sense in theory, the empirical picture is not so neat. Some studies show that reward-guided increases in task preparation impair reactive responses to unexpected cues (Hefer & Dreisbach, 2017; Locke & Braver, 2008). However, other studies show that reward prospect can increase reactive control even after preparing task cues that were invalid (Engelmann et al., 2009; Poh et al., 2019). The empirical review that follows will focus on evidence regarding reward-guided increases in proactive task preparation, with less focus on how reward prospect influences reactive control.

The relationship between reward prospect and enhanced task preparation has been investigated using three main experimental tasks. Flanker tasks involve judging a central target stimulus, such as the direction of a central arrow, with surrounding distractors that are congruent with the target response (e.g. distractor arrows facing the same direction as the target) or incongruent (e.g. distractor arrows facing the opposite direction from the target). In a study by Chiew and Braver (2016), a flanker task was used to investigate how reward would influence the use of congruency information to prepare for a target. When a context cue indicated whether an upcoming target array would contain incongruent or congruent flankers, reward prospect reduced the RT cost of incongruent distractor information. Reward did not have this effect when the context cue was neutral, giving no information about the target (Chiew & Braver, 2016). This suggests that reward prospect

can help to prepare for upcoming cognitive demands using informative context information. Flanker tasks have highlighted the importance of temporal order in reward-context interactions. When reward and context information were presented separately, reward prospect was only beneficial for performance when presented in advance of the context information (Chiew & Braver, 2016) or simultaneously with context information (Vassena et al., 2019). These results are consistent with the idea that reward prospect might help to strengthen initial encoding of context information that is relevant to upcoming task demands.

The second main task used to investigate the effect of reward prospect on task preparation is the AX-continuous performance task (AX-CPT) (Servan-Schreiber et al., 1996). The AX-CPT involves the sequential presentation of letter stimuli, with four conditions based on the order of letters presented (AX, AY, BX, and BY). The first letter denotes the context cue and the second letter denotes the target. If an A context cue is followed by an X target, then a target response button must be pressed. A non-target button must be pressed for the remaining three conditions. Importantly, most trials come from the AX condition (typically 70%), creating a strong expectation that a target response will be required following the A context cue. This means that increased cue processing should benefit performance on AX trials, but impair performance on AY trials, where the second letter does not lead to a target response. Research using this paradigm has shown that reward prospect speeds responding across all conditions but selectively reduces accuracy for AY trials, suggesting that reward increases processing of contextual information (Locke & Braver, 2008). The effect of reward on reducing the accuracy of AY trials has also been shown to carry over into subsequent blocks where reward is removed (Hefer & Dreisbach, 2017). In addition, reward can delay behavioural adjustment when increased cue processing becomes

less adaptive. This has been shown in a study that made X targets less frequent in a later block by reducing the proportion of AX trials from 70% to 40%. Even in the absence of reward, individuals that performed an incentivised version of AX-CPT to begin with continued to have significantly higher error rates for AY trials when transitioning from 70% to 40% AX trials, than individuals who initially performed a non-incentivised version of the AX-CPT (Hefer & Dreisbach, 2017). These results suggest that reward prospect can trigger an increase in cue processing and that these effects can be sustained even when they are no longer advantageous to the task at hand.

The third main framework to examine reward-driven task preparation is task switching. Research has shown that longer preparation intervals, between presentation of task rule information and target stimuli, tend to reduce switch costs (Monsell, 2003). This suggests that increased preparation can support behavioural flexibility needed to switch to a new task. Piecing this together with the evidence that reward prospect can enhance task preparation, high reward prospect would be expected to reduce switch costs for a given preparation interval. Despite this prediction, evidence that reward prospect modulates the switch cost is mixed; reward prospect reduces switch costs in some studies (Hippmann et al., 2019; Kleinsorge & Rinkenauer, 2012; Shen & Chun, 2011) but not in others (Cubillo et al., 2019; Etzel et al., 2016; Kouneiher et al., 2009). This is does not mean that reward does not improve performance overall during task switching. Indeed, many reward-guided task switching experiments report that high reward produces faster reaction times (Kleinsorge & Rinkenauer, 2012; Shen & Chun, 2011) or greater accuracy (Etzel et al., 2016). Nevertheless, the inconsistent influence of reward prospect on switch costs is puzzling. What might give rise to these mixed results? One possibility could be that the influence of reward prospect on flexible control depends on immediate reward history. Increases in reward prospect

relative to the previous trial might support flexible task switching, whereas, consistently high reward prospect might reduce flexibility, due to the stabilisation of previous context information. This might explain why some studies did not observe an interaction between reward prospect and the switch cost; not accounting for reward prospect on the previous trial reward could lead opposing switch cost modulations to average out. Chapter 2 investigates this possibility.

A small corpus of studies provide evidence for the view that different reward prospect sequences might have different effects on task switching. Over six behavioural experiments, Shen and Chun (2011) found that RT switch costs were significantly attenuated for low-to-high reward prospect sequences (i.e. an increasing reward regime). This attenuation primarily occurred relative to low-to-low reward sequences, though attenuations relative to high-to-high reward sequences (i.e. a consistently high reward regime) were also observed in three of the six experiments. Kleinsorge and Rinkenauer (2012) found further evidence for the idea that increases in reward prospect support flexible task preparation. Their first behavioural experiment showed that a no reward trial followed by reward trial resulted in a significant reduction in RT switch costs. Their second behavioural experiment, which allowed for two reward trials in a row, showed that increases in reward reduced the RT switch cost relative to two reward trials in a row. This latter experiment provided some initial evidence that low-to-high reward and high-to-high reward sequences might subserve flexible rule updating and the stabilisation of rule information respectively. Fröber and Dreisbach (2016) leveraged this idea to test whether different reward sequences would also influence voluntary task switching. In contrast to the cued task switching studies discussed so far, voluntary task switching designs allow participants to decide for themselves which task they will perform on a given trial (Arrington

& Logan, 2004). In three behavioural experiments, Fröber and Dreisbach (2016) found that increases in reward prospect significantly increased the rate of voluntary task switches, compared with trial couplets where reward prospect remained high. A control experiment suggested this was not due to a simple change in reward prospect, as the rate of voluntary switching did not differ between low-low and high-low transitions. A second control experiment found this boost in voluntary switching occurred even when participants were instructed to switch randomly and do each task about equally often. More recent studies have replicated these results, showing that increases in reward boost voluntary switching at the level of task selection, prior to execution (Fröber et al., 2019) and that increases in reward can boost the rate of voluntary switching towards more difficult tasks (Jurczyk et al., 2018). Chapter 2 aimed to develop a task that could index the influence of different reward sequences on task switching performance.

Reward prospect enhances neurophysiological signatures of proactive control

So far, we have focused on behavioural evidence indicating that reward prospect can boost proactive processing of context information, in preparation for an upcoming task. Physiological and neural data are consistent with these behavioural findings. For instance, pupil diameter becomes significantly wider prior to target presentation on high compared with low reward trials, suggesting that reward might increase attentional preparation (Poh et al., 2019). Such increases in pupil dilation could also reflect more general increases in arousal, given the positive relationship between pupil diameter and phasic LC activity (Joshi et al., 2016). Pupil diameter tends to be wider in extended temporal contexts where rewards are possible, such as blocks with a mixture of reward and no reward trials. However, it also shows dynamic reward modulation, where diameter is widest on reward trials within blocks where reward is possible (Kostandyan et al., 2019).

EEG analyses have also indicated preparatory neural changes in response to reward prospect, resulting in an increased voltage over frontocentral sensors prior to target presentation in a reward-guided flanker task (Kostandyan et al., 2019). In addition, multivariate fMRI analyses have indicated that patterns of neural activity encoding task rule information are stronger prior to target preparation, on trials where participants stand to gain a reward for their performance (Etzel et al., 2016). In this study, Etzel and colleagues (2016) asked people to switch between judging the sex of human faces and the syllable composition words (two syllable or not). In an initial baseline session, participants performed the task without reward. In a follow up session, participants performed the task with a mixture of reward prospect and no-reward prospect trials. The authors then trained a multivariate classification algorithm to decode the current task rule (face judgments vs. syllable judgments) from neural activity, using the baseline session as the training data. These decoding analyses focused on the preparatory interval between the presentation of rule information and target presentation. The critical test was whether the classification algorithm would show better decoding accuracy for rule information on reward prospect trials. This was indeed the case, especially in supplementary motor area and lateral PFC, suggesting that reward prospect strengthens proactive neural encoding of context information. The strength of rule decoding also statistically mediated the effect of reward on task accuracy, providing evidence that reward-guided changes in proactive control have a meaningful relationship with behavioural performance. While the increases in proactive task preparation observed by Etzel et al. (2016) would be expected to improve switching performance (Monsell, 2003), decoding accuracy was not tested separately for on rule switch and rule repeat trials, due to the absence of a behavioural interaction between reward and task sequence. This leaves open the question of whether reward-driven

increases in rule coding are especially prominent on switch trials, where encoding the current rule is a more critical determinant of behavioural performance. In short, reward-driven increases in rule coding could be due to stronger rule updating, reduced interference with the previous rule or increased attention, all of which could result in better switch trial performance. Chapter 3 aims to investigate this open question, testing whether reward-driven increases in rule encoding are selective to switch trials. To summarise, the behavioural and neural observations outlined above provide strong evidence that high reward prospect can increase proactive cognitive control, which would be expected to enhance behavioural flexibility when switching to a new task (Monsell, 2003).

Current and average reward rates can be compared to regulate switching

Having discussed how reward prospect might support task switching through proactive task preparation, let us turn to foraging. How does reward prospect influence whether an agent should keep exploiting an option or abandon it? For specific situations, this question can be answered using a normative model for stay-switch behaviour called the Marginal Value Theorem (MVT)(Charnov, 1976). MVT was devised to solve the classic patching-leaving dilemma, in which an animal exploits a monotonically depleting choice option and must decide when to abandon it. It proposes that the animal should abandon the current option when its reward rate falls to the average reward rate for the environment. Put simply, an animal should switch when, on average, it will encounter a new option that is just as good or better than its current option. Evidence indicates that animal and human choices in classic patch-leaving tasks are qualitatively consistent with MVT. While some studies show a bias to keep exploiting longer than is optimal, differences in when participants decide to switch for different reward conditions follow a pattern

predicted by MVT (Barack et al., 2017; Constantino & Daw, 2015; Hayden et al., 2011; Kane et al., 2017; Lottem et al., 2018).

Despite its general success, MVT makes predictions about a specific form of stay-switch behaviour, where a current option's reward rate depletes monotonically and where there is uncertainty about what option will be encountered after making a switch. As a result, there are many stay-switch scenarios for which MVT does not apply. For example, when current options have reward rates that do not deplete monotonically, MVT might lead animals to switch away from an option earlier than would be optimal. Instead, it has been proposed that estimating an option's future reward rate, and leaving only when all future estimates fall below the average reward rate, will produce more optimal leaving behaviour (Kolling & Akam, 2017). A second example where MVT falls short is when animals can influence the reward rate of a current option through revisiting options over time. In this situation, it has been shown that show that a more optimal solution uses information about options' depletion rates, their renewal rates and how much time is taken before revisiting an option (Possingham & Houston, 1990). This second example highlights the importance of using *option specific information,* rather than average reward information, when making switch decisions.

To summarise, many forms of stay-switch behaviour fall outside the scope of MVT, which has dominated the design of patch-leaving studies in neuroscience. In focusing on designs where there is uncertainty about what option will be encountered after switching, such designs have neglected the influence that reward information about specific alternatives have on switch decisions. Chapters 4 and 5 aim to redress this and to understand how choices about when to switch are influenced by properties of specific options being switched to. Chapter 4 will provide a conceptual review and proof-of-concept

simulations to show that when an agent can choose which option it will encounter after switching, it might be most optimal to compare the current option's reward rate against the reward offered from specific alternatives. Chapter 5 will investigate how reward-related properties of the option being switched to can be used to infer when that option reaches maximum reward and thus when it should be exploited.

Representing latent reward information and goal-distance to regulate switching

What neural mechanisms might be important for deciding when to switch, in situations that fall outside the scope of MVT? Recent studies point to neural tracking of *latent* reward variables to help regulate switching. Recall that Wittmann et al. (2016) showed dACC, FO and VS encoded information needed to infer latent trends in whether an option was becoming more or less rewarding over time. This information also appeared to be important for switching, with a significant association between neural activity in dACC and the influence of reward trends on the probability of a switch. In another study, Stoll et al. (2016) recorded from neurons in monkey mid-cingulate cortex during decisions to engage with an on-going task or check an alternative option, to see whether the alternative had grown to a threshold that provided a reward bonus. The results showed a transient increase in mid-cingulate activity during such decisions to check, which reflected how close the alternative option was to reaching the reward threshold.

In addition to tracking latent reward information, goal-directed adjustments in behaviour might be achieved by monitoring distance to critical *goal states*. This mechanism has been well illustrated in studies of navigation behaviour. These studies have shown that CA1 firing rates in rat hippocampus track the distance to a goal location and that the strength of this tracking is associated with better task performance (Spiers et al., 2018). In humans, goal distance is encoded in posterior hippocampus and retrosplenial cortex when

navigating familiar environments without explicit navigational cues (Patai et al., 2019). It has also been shown that humans encode distance to goal states in the hippocampus and dorsomedial PFC (dmPFC), with dmPFC modulation tracking the extent to which goal distance influences reaction times (Balaguer et al., 2016). While these studies focus on navigation, recent interest in the non-spatial functions of brain systems associated with navigation (Behrens et al., 2018) raise the possibility that such results are compatible with a broader mechanism, where the distance to important behavioural states is monitored to regulate goal-directed choices. For decisions about when to switch options, this could involve monitoring the temporal distance until an option reaches its maximum value, to decide when it should be exploited. In summary, the studies outlined above provide two candidate neural mechanisms for regulating switch decisions outside the scope of MVT. These include tracking latent reward information and monitoring distance to goal states that inform goal-relevant actions.

What don't we know about reward-guided behavioural flexibility?

Based on the literature reviewed above, there are a number of open questions about how reward prospect contributes to behavioural flexibility. Contemporary studies using decoding approaches have revealed differences in the neural coding of rule information on switch and repeat trials, albeit in opposite directions (Qiao et al., 2017; Waskom et al., 2014). In addition, reward prospect has been shown to boost neural coding of rule information (Etzel et al., 2016). However, one critical gap in this literature is an understanding of whether reward prospect selectively improves coding of rule information on switch trials, where better rule updating, overcoming interference with the previous rule, or increased attention are all more critical determinants of behavioural performance than

on repeat trials. For decisions about when to switch away from an option, the majority of neuroscience studies have focused on switch behaviour that involves a comparison between a current option's reward rate and the average environmental reward rate (Barack et al., 2017; Constantino & Daw, 2015; Hayden et al., 2011; Kane et al., 2017). However, the computational assumptions needed to model these choices have neglected the influence of reward information tied to specific alternatives and how this information is tracked over time to guide switch decisions. Thus, one critical gap in this literature is to understand how the reward properties of specific alternatives influence decisions about when to switch.

Thesis Scope

This thesis examines how reward prospect influences flexible behaviour. Based on cognitive control and decision making studies, it focuses on two prerequisites for flexible control. The first is the ability to update context information (rules) to switch to a new task. The second is sensitivity to reward information, which can be used to make decisions about when to adjust behaviour. Chapter 2 presents three behavioural experiments that investigated how sequential changes in reward modulate the task switch cost. These studies were based on the hypothesis that sequential increases in reward prospect improve rule updating (Fröber & Dreisbach, 2016; Shen & Chun, 2011) but provided inconclusive results. Chapter 3 therefore focused on how current reward prospect influences task switching. Multivariate pattern analyses of human electroencephalograms replicated previous fMRI observations showing that reward prospect strengthens rule encoding (Etzel et al., 2016) and extended this work by showing the reward effect is driven by stronger rule encoding on switch trials. Chapter 4 presents a theoretical review and proof-of-concept simulations which show that using reward information about specific alternatives tends to produce earlier switch decisions and higher overall reward during patch-leaving. Chapter 5 examines how human decision makers decide when to switch towards a specific alternative, as its reward prospect grows towards a maximum in the absence of direct sensory changes. Preliminary results indicate evidence for neural encoding of the option's 'hidden' reward prospect, rather than goal distance. Finally, Chapter 6 discusses the implications of these findings and provides integration across cognitive control and decision making domains.

Chapter 2. Sequential effects of reward on flexible control

Abstract

Motivation has a critical role in tuning cognitive processing to optimise goal-directed behaviour. One theoretical dilemma is that motivation is able to boost antithetical processing modes, improving cognitive flexibility in some situations and stability in others. This raises an important theoretical question about the circumstances in which motivation supports flexible or stable processing respectively. One proposal is that sudden increases in reward prospect promotes flexibility, whereas stable reward prospect promotes stability. In this chapter, we report three behavioural studies that aimed to develop a task suitable for indexing the influence of different reward sequences on flexible cognitive control. All experiments required participants to flexibly switch between two tasks, while the prospect of reward for correct performance was varied from trial-to-trial. We predicted that a transient increase in reward (i.e. a relative increase in reward between the current and previous trial) would promote flexibility in the form of a reduced switch cost. This switch cost reduction would be expected in comparison to a sequence of two high reward trials, which should instead promote stability of the previous task rule, and correspondingly, increase the switch cost. Experiment 1.1 aimed to examine the influence of reward probability sequences on cued and voluntary task switching, using a design in which participants switched between judging the shape and location of target stimuli. Experiment 1.1 did not detect a sequential reward effect on switch costs. Experiments 2.2 and 2.3 aimed to replicate previously reported effects of reward magnitude sequence on cued task switching, wherein switch costs are lowest for low-to-high reward transitions. Experiment 2.2 differed from 1.1 in using a design in which participants switched between judging the shape and colour of target stimuli. In addition, it introduced a temporal separation between

reward prospect and rule information, which had previously been presented simultaneously, as well as an explicit incentive for RT performance. Experiment 2.2 did not detect a sequential reward effect on switch costs. Experiment 2.3 differed from 2.2 in using a design with longer ITI durations. Planned analyses for Experiment 2.3 did not detect a sequential reward effect on switch costs. However, exploratory analyses showed a strong trend towards a sequential effect. Collectively, these studies provide inconclusive results regarding the hypothesis that increases in reward prospect support more flexible behavioural control.

Introduction

Motivation and cognitive control are essential ingredients in daily life. Motivation helps orient us to goal-relevant information and invigorate action to bring about desired outcomes (Botvinick & Braver, 2015), whereas cognitive control directs neural processing based on internally represented goals (Amer et al., 2016; Braver, 2012). Both motivation and cognitive control interact to optimise goal-directed behaviour across a striking number of domains (Kool, Gershman, et al., 2017; Padmala & Pessoa, 2011), including our ability to flexibly switch between tasks (Hippmann et al., 2019; Kleinsorge & Rinkenauer, 2012). For instance, Kleinsorge and Rinkenauer (2012) showed that the RT cost of switching between tasks was reduced when performance was motivated by the prospect of a reward for fast and accurate performance.

These results raise an important theoretical question in motivation and cognitive control research. How does motivation modulate the appropriate balance between stable and flexible cognitive processing (Goschke & Bolte, 2014)? This question is especially relevant in light of results showing that motivation can improve flexibility on the one hand (Kleinsorge & Rinkenauer, 2012) and stabilise information processing on the other (Etzel et al., 2016; Hefer & Dreisbach, 2017). One solution to this theoretical dilemma has been to consider sequential motivational effects, wherein certain reward sequences might support cognitive stability, while others might support flexibility. The central proposal in this respect has been that consistently high reward (e.g. two high reward trials back-to-back) promotes the stable maintenance of task-relevant information. In contrast, transitions from low to high reward situations promote flexible updating of task-relevant information (Fröber & Dreisbach, 2016; Shen & Chun, 2011).

This distinction has gained empirical support in recent years. In a series of behavioural experiments, Shen and Chun (2011) showed that RT switch costs for externally cued task switches were significantly attenuated on low-to-high reward trials. Consistent with these findings, Kleinsorge and Rinkenauer (2012) observed significant switch cost attenuation on reward trials that followed no reward compared with reward trials that followed a reward. Moreover, Fröber and Dreisbach (2016) showed similar effects in the domain of voluntary task switching, where participants can choose which task they perform and when to switch between them (Arrington & Logan, 2004). In particular, Fröber and Dreisbach (2016) observed that the rate of voluntary task switches was significantly higher on low-to-high reward trials than on high-to-high reward trials. Increases in reward have also been shown to increase voluntary switching to more difficult tasks (Jurczyk et al., 2018), suggesting these sequential effects do not simply reflect a strategy in which people prefer to perform the easier task when reward stakes are high. Rather, the results seem to support the hypothesis that relative increases in reward promote greater flexibility, whereas consistently high rewards promote stability.

While demonstrated in the behavioural domain, it is not yet clear what neural mechanisms could give rise to such sequential reward effects on cognitive control. One theoretical possibility is that phasic release of midbrain dopamine could support flexible control by allowing relevant information into PFC (Braver & Cohen, 2000; Goschke & Bolte, 2014; Westbrook & Braver, 2016). Under this model, information encoded in working memory, such as the current task rule, is maintained in a stable, distractor-resistant format in the absence of salient events. When salient events occur, such as the presentation of high reward cues, phasic dopamine opens the 'gate' to working memory, allowing new information to be encoded within PFC. While this dopamine gating account has not been

considered in light of sequential reward effects, it offers a potential account to explain them. Specifically, if the subjective value of a high reward trial is greater following a low reward trial than following a high trial, phasic dopamine release could be higher for low-to-high reward transitions. This in turn could allow new task information to be more strongly encoded in PFC, supporting flexible behavioural control in favour of the new task.

Given that this theoretical account is still highly speculative in the context of sequential reward effects, the behavioural experiments in this chapter aimed to develop a new behavioural task, sensitive to sequential effects of reward on flexible control. Achieving this aim would allow us to subsequently probe how different reward sequences influenced the neural encoding of task rule information in the service of flexible behavioural control.

Experiment 2.1: Influence of sequential reward probability on cued and voluntary task switching

Previous research has suggested that transitions from low to high reward contexts promote cognitive flexibility more than other sequential changes in reward (Fröber & Dreisbach, 2016; Kleinsorge & Rinkenauer, 2012; Shen & Chun, 2011). This effect has been shown for both cued task switching, where the relevant task is indicated by an external cue (Shen & Chun, 2011), and voluntary task switching, where the participant has a choice over which task they perform (Arrington & Logan, 2004; Fröber & Dreisbach, 2016). The findings described above examined sequential changes in the magnitude of reward and it is unknown whether sequential increases in reward probability also result in reduced switch costs. The present experiment therefore aimed to test how sequential increases in reward probability impact cued and voluntary task switching performance. Based on previous research (Shen & Chun, 2011), our main prediction was that sequential increases in reward probability would reduce the reaction time (RT) switch cost for forced choice trials, where the relevant task rule is cued. To examine voluntary task switching, we used a measure called the voluntary switch rate (VSR), defined as the proportion of voluntary switches made on free choice trials (Fröber & Dreisbach, 2016). Based on this previous work, we predicted that sequential increases in reward probability would increase the VSR.

Method

Participants

Participants were 24 individuals between 18-28 years of age (mean age = 23, 14 female). Participants reported no history of neurological or psychiatric illness and normal or corrected-to-normal vision. Sample size was selected based on previous studies (Fröber & Dreisbach, 2016; Shen & Chun, 2011). Participants received course credit or £5 for participation and could earn up to £5 for performance. The study was approved by the Central University Research Ethics Committee at the University of Oxford and all participants signed informed consent before taking part.

Materials

The task was presented on a computer screen with a spatial resolution of 1280 x 1024. The task was run using Psychophysics Toolbox-3 (RRID: SCR_002881) in MATLAB (RRID: SCR_001622, version R2015b). Statistical analyses were performed in JASP 0.8.3.1 (RRID: SCR_015823, version 0.8.1.3). Coloured fixation crosses (purple, yellow, green) were used as task cues, “ ” and * * symbols were used as reward cues, and abstract shapes (square, circle, triangle) were used as target stimuli.

Procedure

In the present task (Figure 1), participants evaluated stimuli based on shape (square, circle, triangle) or location (left, middle, right) (Poljac & Yeung, 2014). On forced choice trials, participants were informed about whether a location or shape judgement was required to obtain points. On free choice trials, participants could choose whether to respond to stimuli based on shape or location; both tasks were equally valid to obtain points. Forced and free choice trials were combined in a single paradigm because it has been shown that forced choice trials increase the rate of voluntary task switching (Fröber &

Dreisbach, 2016), which could potentially help to boost power in our analyse of free choice trials. Each trial had a low or high reward probability for making a correct response, and this information was signalled to participants on a trial-by-trial basis.

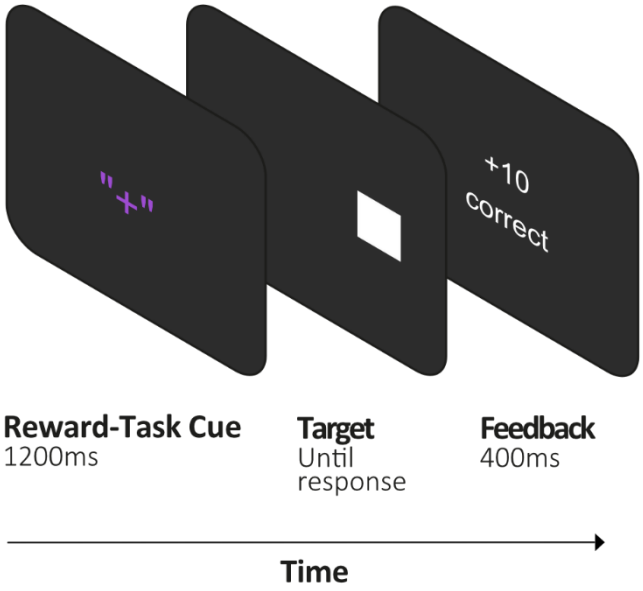


Figure 1. Task design for experiment 2.1. A composite reward-task cue was first presented. The cue colour indicated whether an upcoming target should be judged based on its shape, its location or whether the participant had a free choice between the two tasks. The probability of reward was indicated as being high or low by quotation marks or asterisks around the fixation cross. Following a response, participants saw feedback about: a) whether their response was correct and b) whether they were awarded points, as determined by the reward probability for a correct response on that trial.

Each trial consisted of a cue (1.2s), a target stimulus (until response) and feedback (1s). There were no delays between the reward-task cue phase and the target phase, nor between the target and feedback phases. The colour of the cue signalled which cognitive judgment participants should make in the response to the stimulus (shape or location) or that they had a choice between these two task rules. The correspondence between colours and task rule was counterbalanced between participants. During the experiment, colour

presentation was pseudorandomized so that approximately 50% of trials were forced choice trials and 50% were free choice trials within each block. In addition to the task rule, the cue also signalled the reward probability for a correct response. In low probability trials, correct responses were rewarded with a probability of 0.25, whereas in high probability trials correct responses were rewarded with a probability of 0.75 (Cohen et al., 2007; Smith et al., 2009). Following the cue, a target stimulus was presented on screen until a response. The stimulus was randomly selected on each trial and could be a square, circle or triangle located in the left, middle or right sections of the screen. These two features made the target bi-dimensional, capable of being categorized in two different ways. On forced choice trials, only one feature was valid to obtain reward, whereas, on free choice trials both features were equally valid. Following the participant's response, a feedback screen was used to indicate whether a correct judgment was made. It also indicated whether the participant was awarded points for their response, which depended on whether the response was accurate and the reward probability on that trial. Incorrect responses earned '+0' points, whereas, correct responses with a probabilistically determined reward earned '+10' points.

It should be noted that the cue duration selected was guided by previous behavioural (Fröber & Dreisbach, 2016; Shen & Chun, 2011), electroencephalographic (Poljac & Yeung, 2014; Richter & Yeung, 2016) and neuroimaging studies (Wisniewski et al., 2016; Zhang et al., 2013). The decision to present the stimulus until a response, as opposed to constraining the response window, is in keeping with several relevant studies (Fröber & Dreisbach, 2016; Poljac & Yeung, 2014; Shen & Chun, 2011). In addition, feedback duration was also set as the average duration used by previous behavioural studies (Fröber & Dreisbach, 2016; Shen & Chun, 2011). Participants completed six blocks of 120 trials.

However, prior to the onset of experimental blocks, participants completed the training sequence outlined below.

Training. Participants began the session by learning response keys associated with location and shape judgements. First, participants were presented with randomly selected targets (shapes in different locations) and responded based on their locations. Targets remained on the screen until a response was given, after which they received feedback about the accuracy of their response for 1s. Unless stated otherwise, participants needed to make 10 correct responses to progress to the next stage. Following location judgements, participants completed the same key learning exercise for shape judgements. Whether location and shape judgements were mapped to left hand keys (‘s’, ‘d’, ‘f’) or the right hand keys (‘j’, ‘k’, ‘l’) was counterbalanced between participants.

In the next training phase, participants were presented with coloured '+' cues before each target. The colour of the cue signalled whether the participant was required to respond to the stimulus based on location, shape or whether they had a choice between these task rules. First, participants completed training trials in which the location cue was presented for 1.2s, followed by a randomly selected target. This process was repeated for shape cues and finally for free choice cues. For free choice trials, participants were instructed to make shape or location judgements equally often and in a random fashion. This is a common instruction in voluntary switching paradigms to counteract repetition bias, in which participants consistently respond with the same type of judgement (e.g. location) under free choice conditions (Arrington & Logan, 2004).

In the final practice stage, participants completed two 20-trial practice blocks that were identical to the main experimental task. These practice blocks built on the last training step by adding reward probability symbols (“ “ and * *) to the trial cues. The

correspondence between the symbols and low (0.25) or high (0.75) reward probability was counterbalanced within subjects, across the training and experimental blocks. This meant that the reward-cue-to-probability-assignment changed for different blocks in the session. Participants were instructed about the correspondence between symbols and reward probability at the beginning of each block. Following training, participants completed 6 blocks of 120 experimental trials.

Statistical Analysis. Our main dependent measures were reaction time (RT), the proportion of correct responses (accuracy) and the voluntary switch rate (VSR). We calculated median RTs for the respective design cells to account for the skewness of RT distributions (see Ratcliff, 1993). For all RT analyses, we only included trials in which the current and previous trials were correct to mitigate potential effects of post-error slowing, in which participants tend to respond more slowly after making an error (Dutilh et al., 2012; Notebaert et al., 2009). For forced choice trials, in which a specific task was cued, data were analysed using 2x2x2 repeated measures ANOVAs with factors of previous reward probability (low x high), current reward probability (low x high) and task sequence (task switch x task repeat). These analyses included forced choice trials regardless of whether the previous trial involved forced or free choice. For free choice trials, VSRs were calculated for each previous reward-current reward condition. This calculation involved dividing the number of task switches on free choice trials (relative to previous trials) in a condition by the total number of free choice trials in that condition. Eligible trials for VSR calculation were determined based on the RT and accuracy criteria outlined above. In addition, these analyses used free choice trials regardless of whether the previous trial involved a free or forced choice. VSRs were analysed using a 2x2 repeated measures ANOVA with factors of previous reward probability (low x high) and current reward probability (low x high).

Results

Reward probability sequence did not impact cued task switching

We first conducted a 2x2x2 (previous reward x current reward x task sequence) repeated measures ANOVA on RT for forced choice trials. This showed a significant main effect of task sequence ($F(1, 23) = 10.55$, $p = 0.004$, $\eta_p^2 = 0.314$), reflecting significantly slower RTs on switch trials (M = 653ms, SD = 194.31) compared with repeat trials (M = 603ms, SD = 139.22). The analysis did not detect significant main effects of reward probability from the previous trial ($F(1, 23) = 0.318$, $p = 0.578$, $\eta_p^2 = 0.014$) or the current trial ($F(1, 23) = 1.935e^{-4}$, $p = 0.989$, $\eta_p^2 = 0.000$). Similarly, the analysis did not detect significant interactions between task sequence and reward probability on the previous trial ($F(1,23) = 1.569$, $p = 0.223$, $\eta_p^2 = 0.064$), task sequence and reward probability on the current trial ($F(1,23) = 0.062$, $p = 0.805$, $\eta_p^2 = 0.003$), reward probability on the current and previous trials ($F(1,23) = 0.764$, $p = 0.391$, $\eta_p^2 = 0.032$) or task sequence, previous and current reward probability ($F(1,23) = 0.185$, $p = 0.671$, $\eta_p^2 = 0.008$).

The equivalent 2x2x2 (previous reward x current reward x task sequence) repeated measures ANOVA for accuracy showed the same pattern of results. There was a significant main effect of task sequence on ($F(1, 23) = 9.402$, $p = 0.005$, $\eta_p^2 = 0.290$), reflecting significantly lower accuracy on switch trials (M=0.94, SD=0.08) compared with repeat trials (M=0.96, SD=0.06). The analysis did not detect significant main effects of reward probability from the previous trial ($F(1, 23) = 0.850$, $p = 0.366$, $\eta_p^2 = 0.036$) or the current trial ($F(1, 23) = 0.121$, $p = 0.731$, $\eta_p^2 = 0.005$). Similarly, the analysis did not detect significant interactions between task sequence and reward probability on the previous trial ($F(1,23) = 2.488$, $p = 0.131$, $\eta_p^2 = 0.096$), task sequence and reward probability on the current trial ($F(1,23) = 1.345$, $p = 0.258$, $\eta_p^2 = 0.055$), reward probability on the current and previous trials ($F(1,23) =$

1.151, $p = 0.294$, $\eta_p^2 = 0.048$) or task sequence, previous and current reward probability ($F(1,23) = 2.334$, $p = 0.140$, $\eta_p^2 = 0.092$).

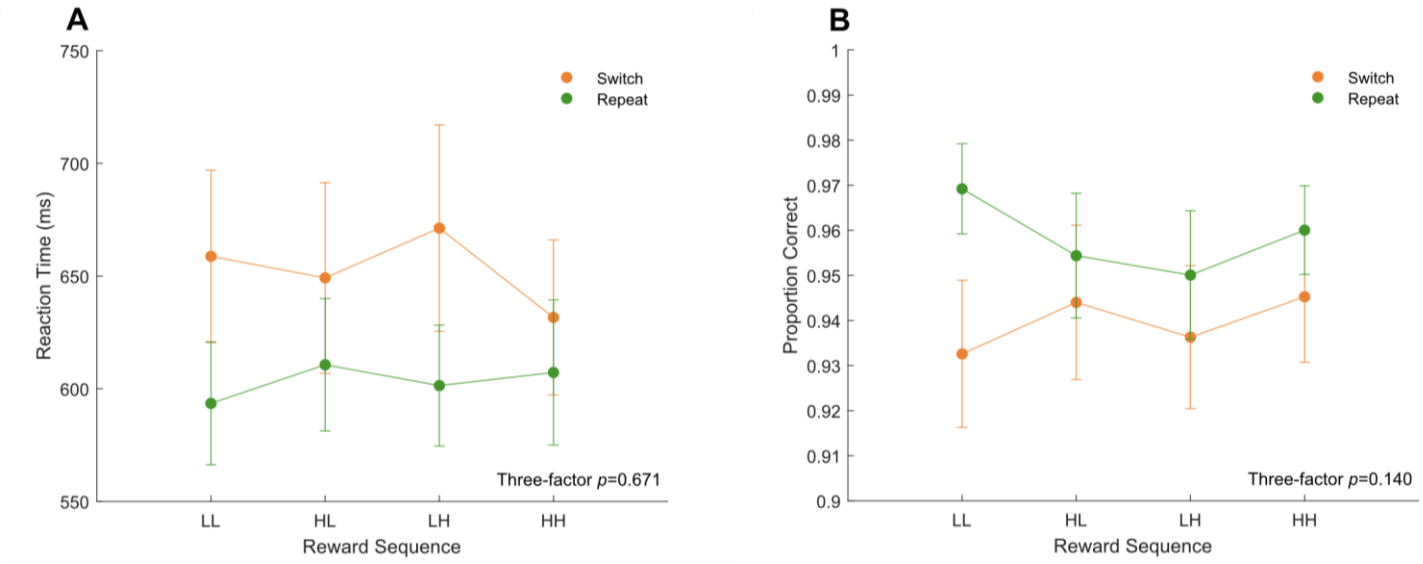


Figure 2. Behavioural performance on forced choice trials in Experiment 2.1. A: Reaction time for reward probability sequences, separated by switch and repeat trials. B: Proportion correct for reward probability sequences, separated by switch and repeat trials. A-B: LL, HL, LH and HH indicate low-to-low, high-to-low, low-to-high and high-to-high reward sequences. Three-factor p-values indicate the interaction between previous reward, current reward and task sequence. Error bars show standard error of mean.

Reward probability sequence did not impact the rate of voluntary switches

Next we performed a 2x2 (previous reward x current reward) on the VSR for free choice trials. This analysis did not detect a significant main effect of previous reward probability ($F(1, 23) = 1.868$, $p = 0.185$, $\eta_p^2 = 0.075$) or current reward probability ($F(1, 23) = 1.844$, $p = 0.188$, $\eta_p^2 = 0.074$). In addition, the analysis did not detect a significant interaction between reward probability on the previous and current trials ($F(1, 23) = 0.181$, $p = 0.675$, $\eta_p^2 = 0.008$). The VSR was determined a priori as the primary dependent measure for free

choice trials in this study. However, 2x2x2 (previous reward x current reward x task sequence) repeated measures ANOVAs for RT and accuracy on free choice trials can be found in the appendix for this chapter.

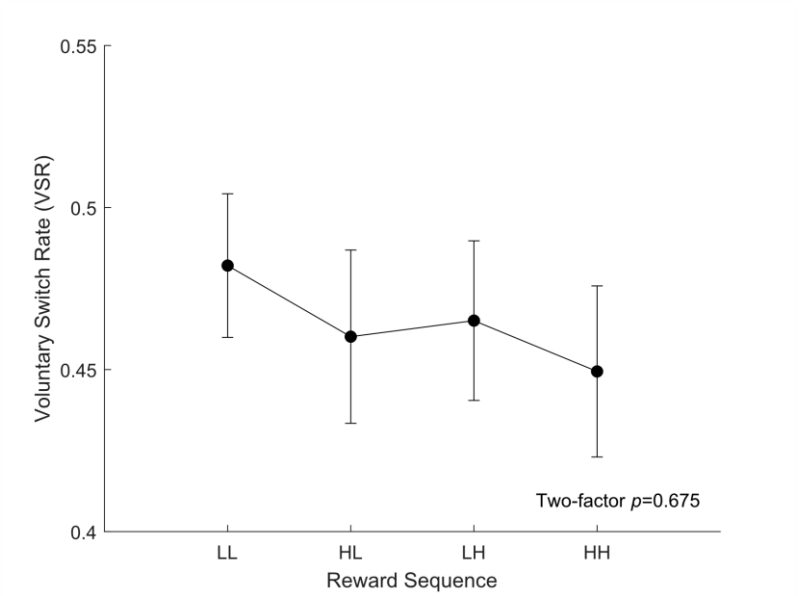


Figure 3. Voluntary switch rates on free choice trials for different reward probability sequences. LL, HL, LH and HH indicate low-to-low, high-to-low, low-to-high and high-to-high reward sequences. The two p-value indicates the interaction between previous reward and current reward on voluntary switch rate. Error bars show standard error of mean.

Discussion

This study aimed to assess the effect of sequential changes in reward probability on cued and voluntary task switching. Contrary to our initial prediction, we did not detect an effect of reward probability sequence on the speed with which participants switched tasks when cued to do so. This was also the case for accuracy. Contrary to our second prediction, we did not detect an influence of reward probability sequence on the voluntary switch rate (VSR). These null results differ from previous studies, showing sequential effects of reward magnitude on cued and voluntary task switching (Fröber & Dreisbach, 2016; Kleinsorge & Rinkenauer, 2012; Shen & Chun, 2011). One possibility would be that sequences of reward probability do not influence flexible control in the same manner as sequences of reward magnitude. However, we should also consider the simpler possibility: practical aspects of our task meant it was not a sensitive index of reward-driven changes in flexible control. First, reward-cue mappings were flipped on 50% of the experimental blocks in our task. This instability in the meaning of the reward cues may have made it difficult for participants to accurately encode the current reward condition and for information processing to be modulated accordingly. Second, the lack of reward incentive for RT performance could have increased RT variance and obscured condition differences on forced choice trials. Indeed, sequential reward effects on cued rule switching has been shown in the RT domain, using a strong RT incentive manipulation (Shen & Chun, 2011). These practical aspects make it challenging to draw substantive theoretical conclusions from the present results. We therefore opt for a cautious approach, concluding that this task was not a sensitive index of potential reward probability sequence effects on flexible control.

Experiment 2.2: Influence of sequential reward magnitude on cued task switching

The null effects observed in Experiment 2.1 could have been due to unstable mappings between cues and reward information, as well as the lack of incentive for RT performance. In Experiment 2.2, we adapted these task features to test whether we could replicate the sequential effects of reward magnitude on flexible rule switching (Kleinsorge & Rinkenauer, 2012; Shen & Chun, 2011). First, we introduced stable reward-cue mappings that were consistent throughout the session. We also included a temporal separation between reward cues and rule cues, to ensure participants had sufficient time to encode current reward information. Second, we sought to make RT a more informative measure of flexible control by introducing an RT incentive structure, in which more points are awarded for faster responses (Shen & Chun, 2011; see methods). Third, we excluded uncued (free-choice) trials to increase power for cued trial analyses. Based on previous research (Kleinsorge & Rinkenauer, 2012; Shen & Chun, 2011), we made three central predictions. First, we predicted that reward magnitude sequence would modulate the speed of cued task switches, as indicated by a significant three-way interaction between reward on the previous and current trials, and whether the task was repeated or switched. Second, we predicted that the low-to-high reward sequence would provide the greatest benefit to switching, as indicated by a significantly lower switch cost in this condition. Thirdly, we predicted that these effects would be present for RT but not accuracy, as our experimental task contained a strong RT incentive.

Method

Participants

Participants were 26 individuals with normal or corrected-to-normal vision, who reported no neurological or psychiatric illness. Sample size was selected based on previous studies (Fröber & Dreisbach, 2016; Shen & Chun, 2011) and our task design. Specifically, we aimed for a target sample of 24 participants to fully counterbalance stimulus-task rule mappings and response mappings between participants. Two participants from our recruited sample were excluded from analyses, one due to an average accuracy below 70% and another due to a failure to follow task instructions. The final sample included in our analyses were 18-35 years of age (mean = 26, 14 = female). Participants received £5 for taking part, with the opportunity to earn up to £5 extra for performance. This study was approved by the Central University Research Ethics Committee at the University of Oxford and all participants signed informed consent before taking part.

Materials

Materials were the same as Experiment 2.1, with the exceptions that auditory tones were used as task cues and upward/downward facing arrows were used as reward cues.

Procedure

Like the previous design, participants evaluated bi-dimensional target stimuli to earn points. In the current study, participants made colour (yellow, blue) or shape (square, circle) judgements about target stimuli based on an auditory rule cue (high pitch, low pitch). The location judgement task from the previous experiment was changed to a colour judgement task in the current experiment. This was because location judgements have a natural mapping with response keys from left to right, which could make them easier than arbitrarily mapped shape judgements (Poljac & Yeung, 2014). The new design therefore

contained two tasks with arbitrary response mappings. The interval between rule cue presentation and the target stimulus was also shortened from 1200ms in the previous experiment to 600ms in the current experiment. This change aimed to increase switch costs and thereby increase sensitivity to detect differences in this measure as a function of reward sequence. The reward magnitude for making a correct response was indicated at the beginning of each trial via an upward (high reward) or downward (low reward) facing arrow. Participants completed the training and experimental sequences outlined below.

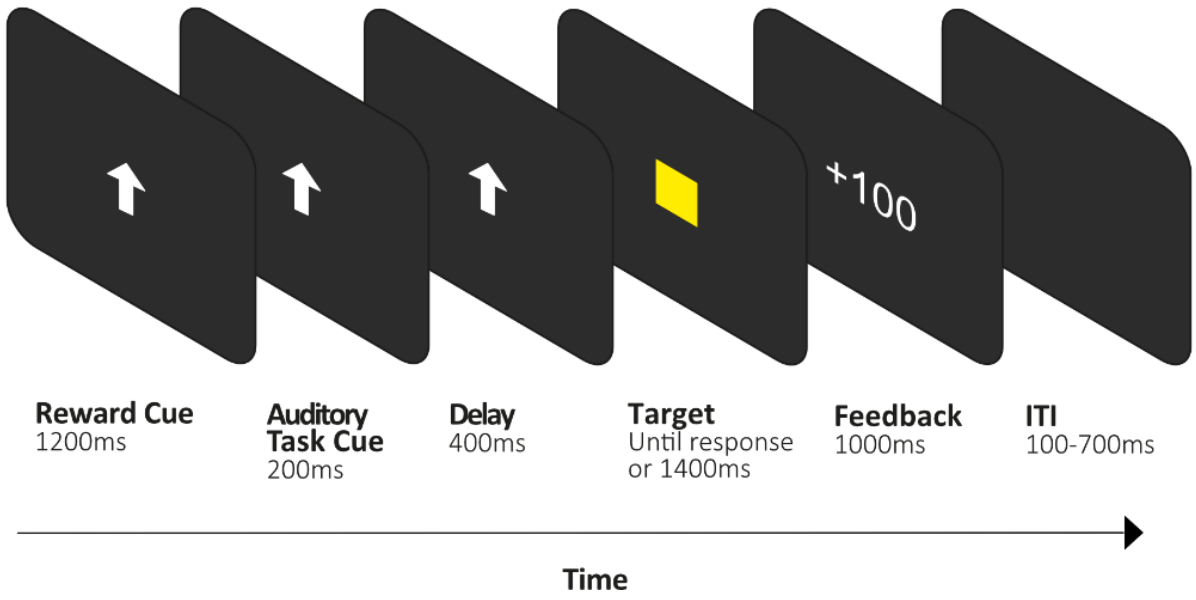


Figure 4. Task sequence for Experiment 2.2. A high or low reward cue was presented for 1.2s before the onset of a task rule cue (tone), which signalled whether participants needed to respond to an upcoming target based on shape or colour. Following the response or an omission, feedback was presented, followed by an inter-trial-interval.

Training. Training followed a similar protocol to the previous experiment. Participants began the session by learning response keys associated with the two tasks. The number of response options was reduced from three in the previous experiment to two in the present experiment, to reflect the reduction in the number of target features. Left hand response keys were 'd' and 'f', whereas, right hand responses keys were 'j' and 'k'. All other aspects of this first training stage were conceptually identical to the previous experiment. Participants were presented with randomly selected target stimuli (coloured shapes) and were asked to respond based on their colours. Targets remained on screen until a response was given, after which they received feedback about the accuracy of their response for 1s. Participants needed to make 10 correct responses to progress to the next training stage. Following colour judgements, participants completed the same key learning exercise for shape judgements. Whether colour or shape judgements were mapped to left hand or the right hand keys was counterbalanced between participants.

In the next training phase, participants were presented with an auditory tone, which had an onset of 600ms before each target. Tones were sine functions with a sample rate of 8000, a duration of 200ms and frequencies of 318Hz or 440Hz. The tone frequency determined the task rule, with one frequency signalling that a colour judgement was needed to gain points on the trial and the other frequency signalling that a shape judgement was required. The relationship between tone and task rule was counterbalanced between participants. In this training section, participants first completed trials in which the colour tone was presented for 600ms, followed by a randomly selected target. Following 10 correct responses, this process was repeated for shape judgements.

Final training stage and experiment blocks. In the final practice stage, participants completed a block of 72 trials, which was identical to the main experimental blocks. This

stage added reward cues to the start of each trial. Upward facing arrows indicated participants could earn between 50-100 points for a correct response. By contrast, downward facing arrows indicated participants could earn between 5-10 points. The amount participants earned within these ranges depended on how quickly they responded to the target once it was presented. Initially, the reaction time (RT) criteria for different points was pre-set so that responses faster than 400ms, 600ms, 800ms, 1000ms, 1200ms and 1400ms earnt 100/10, 90/9, 80/8, 70/7, 60/6 and 50/5 points on high and low reward trials respectively. For trials on which a correct response was performed, the current trial RT was added to an array for its condition (e.g. low-to-low reward, switch trial). Once there were more than six values in an array, individualised points criteria were calculated for that condition and were calculated again every time a new value entered the array. The individualised points criteria followed criteria outlined by Shen and Chun (2011), in which the most points to least points are rewarded for correct responses faster than 95%, 80%, 65%, 50% and 35% of the median RT. It should be noted that Shen and Chun's (2011) approach evaluated RTs based on all previous reaction times, which does not prevent participants from deliberately going slower on low reward trials, so that it is easier to gain more points on high reward trials. Our design circumvented this bias by individualising points criteria separately for each condition.

From this stage of training through to the end of the experiment, each trial consisted of a visual reward cue, an auditory tone, a target, feedback and a randomly selected inter-trial interval (ITI) with a value of 0.1, 0.2, 0.3, 0.5, 0.6 or 0.7s. The visual reward cue was presented for 1.8s at the beginning of each trial. Auditory tone onset occurred 1.2s after reward cue presentation and the tone duration lasted 200ms. Following the cue phase, the target (i.e. a coloured shape) was presented until a response was made or for a maximum of

1.4s without response. Participants completed one practice block of 72 trials, followed by 8 experimental blocks of 72 trials. Stimulus presentation was pseudorandomised, to ensure equal numbers of reward cues, task cues, targets and ITI values were presented within each block.

Statistical Analysis. Our main dependent measures were RT and proportion correct (accuracy). The analysis procedure was near identical to the previous experiment. For RT analyses, we used condition medians to account for the skewness of RT distributions (see (Ratcliff, 1993) and only included trials in which the current and previous trials were correct to mitigate potential effects of post-error slowing (Dutilh et al., 2012; Notebaert et al., 2009). For accuracy analyses, we included trials in which a response was given before the 1400ms cut off window. Data were analysed using 2x2x2 repeated measures ANOVAs with factors of previous reward magnitude (low x high), current reward magnitude (low x high) and task sequence (task switch x task repeat).

Results

Current reward, task sequence, and the previous reward-task sequence combination modulated RT

A 2x2x2 (previous reward x current reward x task sequence) repeated measures ANOVA on RT showed a significant main effect of current reward ($F(1, 23) = 6.162$, $p = 0.021$, $\eta_p^2 = 0.211$), reflecting slower RTs when the current trial offered a low reward (M = 507ms, SD = 124.51) compared with a high reward (M = 484ms, SD = 119.51). The analysis also showed a main effect of task sequence ($F(1, 23) = 29.49$, $p < 0.001$, $\eta_p^2 = 0.562$), reflecting slower RTs on switch trials (M = 526ms, SD = 140.68) compared with repeat trials (M = 466ms, SD = 91.97). While the analysis did not detect a main effect of reward on the previous trial ($F(1, 23) = 3.145*10^{-4}$, $p = 0.986$, $\eta_p^2 = 0.000$), there was a significant interaction between reward on the previous trial and task sequence ($F(1, 23) = 9.473$, $p = 0.005$, $\eta_p^2 = 0.292$). This interaction was driven by a reduction in the switch cost when a previous trial carried high reward (M = 49ms, SD = 52.75) than when it carried low reward (M = 72ms, SD = 59.45). The analysis did not detect a significant interaction between reward on the current trial and task sequence ($F(1, 23) = 1.239$, $p = 0.277$, $\eta_p^2 = 0.051$) or an interaction between reward on the previous and current trials ($F(1, 23) = 0.548$, $p = 0.466$, $\eta_p^2 = 0.023$). Importantly, the three-factor interaction between previous reward, current reward and trial type, which would be indicative of a sequential reward effect, was not significant ($F(1, 23) = 0.058$, $p = 0.811$, $\eta_p^2 = 0.003$).

Reward did not modulate task accuracy

A 2x2x2 (previous reward x current reward x task sequence) repeated measures ANOVA on accuracy showed a main effect of task sequence ($F(1, 23) = 11.434$, $p = 0.003$, $\eta_p^2 = 0.332$), reflecting lower accuracy on switch trials (M = 0.86, SD = 0.09), compared with

repeat trials (M = 0.89, SD = 0.08). The analysis did not detect main effects of reward on the previous trial ($F(1, 23) = 1.923$, $p = 0.179$, $\eta_p^2 = 0.077$) or reward on the current trial ($F(1, 23) = 3.073$, $p = 0.093$, $\eta_p^2 = 0.118$). Similarly, the analysis did not detect significant interactions between reward on the previous trial and task sequence ($F(1, 23) = 0.173$, $p = 0.681$, $\eta_p^2 = 0.007$), reward on the current trial and task sequence ($F(1, 23) = 0.221$, $p = 0.643$, $\eta_p^2 = 0.010$), reward on the previous and current trials ($F(1, 23) = 0.693$, $p = 0.414$, $\eta_p^2 = 0.029$) or a three-factor interaction between reward on the previous trial, reward on the current trial and task sequence ($F(1, 23) = 0.077$, $p = 0.784$, $\eta_p^2 = 0.003$).

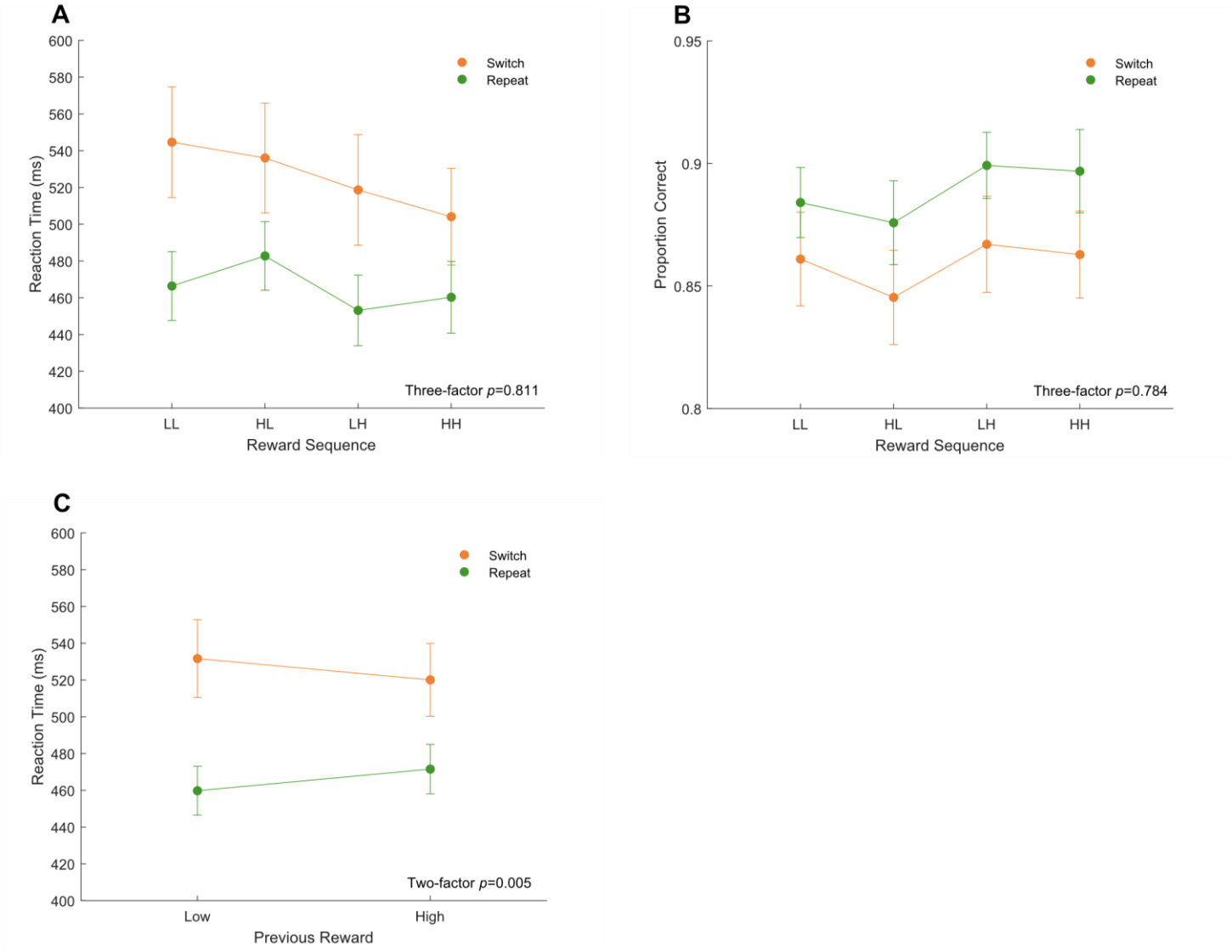


Figure 5. Behavioural performance in Experiment 2.2, in all which all trials involve a cued task rule. A: Reaction time for reward magnitude sequences, separated for switch and repeat trials. B: Proportion correct for reward magnitude sequences, separated for switch and repeat trials. A-B: LL, HL, LH and HH indicate low-to-low, high-to-low, low-to-high and high-to-high reward sequences. Three-factor p-values indicate the interaction between previous reward, current reward and task sequence. C: Reaction time for switch and repeat trials as function of reward prospect on the previous trial. The two-factor p-value indicates the interaction between previous trial reward prospect and task sequence. A-C: Error bars show standard error of mean.

Discussion

In Experiment 2.1, we did not detect any interactions between reward prospect and task sequence. In Experiment 2.2, we changed the task to use stable reward-cue mappings, removed uncued (free choice) trials and presented reward prospect information prior to the task rule. Contrary to our first prediction based on Kleinsorge and Rinkenauer (2012), as well as Shen and Chun (2011), we did not observe a three-factor interaction between reward on the previous trial, reward on the current trial and task sequence on RT, which would have indicated a sequential reward effect on flexible processing. This result meant that our second prediction, about the switch cost being significantly lower for the low-to-high reward sequence (Shen & Chun, 2011), was also unsupported. Our third prediction about reward effects being observed in RT but not accuracy gained partial support. While a three-factor interaction was not observed for either measure, we observed reduced RTs when current reward was high and reduced RT switch costs when previous reward was high. In contrast, there was a trend towards higher accuracy on high reward trials, but no evidence of reward interacting with task sequence to influence accuracy.

The unexpected RT interaction between reward on the previous trial and switch cost observed in this study might have been due to relatively short ITI durations. For instance, short temporal delays between trials could have meant that high reward feedback from the previous trial boosted attention for long enough to improve current trial performance. As repeat trials are performed with relative ease, such an attentional boost might be expected to have a greater benefit on switch trials, thereby lowering the switch cost. The fact that we did not detect an interaction between current reward and the RT switch cost could indicate that reward outcome information from the previous trial overshadowed the influence of current reward information on rule updating. While the previous reward-task sequence

interaction was intriguing, we had aimed to replicate the three-factor interaction between reward sequence and task sequence, wherein switch costs are lowest for sequential increases in reward (Kleinsorge & Rinkenauer, 2012; Shen & Chun, 2011). We reasoned that this effect may not have been detected due to previous reward outcomes overshadowing the potential influence of current reward prospect. In the next experiment, we increased the ITI duration to separate previous reward outcome and current reward prospect processing, and test whether this might unmask the sequential reward effect previously reported (Kleinsorge & Rinkenauer, 2012; Shen & Chun, 2011).

Experiment 2.3: Influence of sequential reward magnitude on cued task switching

The absence of a three-factor interaction between previous reward, current reward and task sequence in Experiment 2.2 could have been due to short temporal durations between trials. In particular, reward outcome processing from the previous trial could have overshadowed the potential influence of current reward prospect on rule updating. Experiment 2.3 aimed to test whether increasing the ITI duration would allow feedback processing to return to baseline and unmask the three-factor interaction reported in previous work (Shen & Chun, 2011). With a prospective EEG study in mind, we also modified the task to be more EEG-compatible (see method).

Method

Participants

Participants were 28 individuals with normal or corrected-to-normal vision, who reported no neurological or psychiatric illness. We aimed for a target sample of 24 participants to fully counterbalance stimulus-task rule mappings and response mappings between participants. Four participants were excluded from analyses, two due to accuracy below 70% and two due to technical errors during data collection that resulted in unusable data. The final sample included in our analyses were between 18 and 31 years of age (mean = 22, 18 = female). Subjects received a baseline of £5 or course credits for taking part and had the opportunity to earn up to £5 extra in each half of the session for performance. This study was approved by the Central University Research Ethics Committee at the University of Oxford and all participants signed informed consent before taking part.

Materials

Materials were the same as experiments 2.1 and 2.2.

Procedure

The task and analysis procedure was almost identical to experiment 2.2. The first change from the previous experiment was that one judgement from each task was mapped to the same response button. This meant that participants could not begin to prepare the upcoming response hand following presentation task rule information, which would otherwise confound the analysis of reward-driven task preparation in a subsequent EEG experiment. The left hand response key was ‘f’ and the right hand response key was ‘j’. One response button corresponded to a yellow/square judgement and the other button corresponded to a blue/circle judgement. The mapping between response keys and judgement pairs was counterbalanced across subjects. The second change from the

previous experiment was that feedback duration was shortened from 1000ms to 200ms and ITI duration was increased from 100-700ms to 900-1500ms. While this did not alter the absolute duration of trials, we assumed that the temporal duration of reward outcome processing would be reduced as a function of its on-screen duration. The final change was to balance tone-task rule mappings within each subject. This was achieved by following the structured training sequence outlined for experiment 2.2, then testing participants on four experimental blocks of 72 trials. Participants then went through the training sequence for a second time with the opposite tone task rule-mapping, followed by four test blocks of 72 trials.

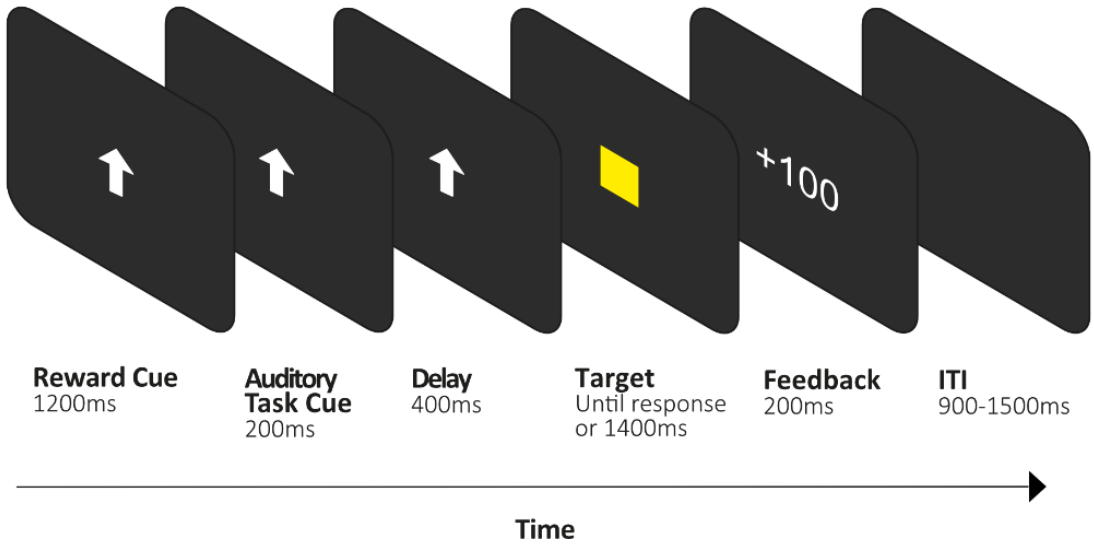


Figure 6. Task sequence for Experiment 2.3. A high or low reward cue was presented for 1.2s before the onset of a task rule cue (tone), which signalled whether participants should respond to upcoming stimulus based on shape or colour. Following the response or an omission, feedback was presented, followed by an inter-trial interval. The procedure section details differences between Experiment 2.2 and Experiment 2.3.

Results

Current reward and task sequence modulated RT

A 2x2x2 (previous reward x current reward x trial transition) repeated measures ANOVA on RT showed a significant main effect of current reward ($F(1, 23) = 16.284$, $p < 0.001$, $\eta_p^2 = 0.415$), reflecting slower RTs when current reward was low (M = 449ms, SD = 78.04) than when it was high (M = 425, SD = 74.78). The analysis also detected a significant main effect of task sequence ($F(1, 23) = 36.266$, $p < 0.001$, $\eta_p^2 = 0.612$), reflecting slower RTs on switch trials (M = 462, SD = 85.24) compared with repeat trials (M = 412, SD = 58.73). In addition to these main effects, we observed two trends. First, there was a trend towards an interaction between current reward and task sequence (current reward and task sequence ($F(1, 23) = 3.468$, $p = 0.075$, $\eta_p^2 = 0.131$), driven by a numerical reduction in the switch cost when current reward was high (M = 44ms , SD = 48.40) compared with when it was low (M = 55, SD = 47.72). Second, there was a trend towards an interaction between reward on the previous and current trials ($F(1, 23) = 4.182$, $p = 0.052$, $\eta_p^2 = 0.154$), driven by a greater RT modulation by reward on the current trial (i.e. RT low current reward – RT high current reward), when the previous trial had a high reward (M = 28ms, SD = 44.59) than when it had a low reward (M = 18ms, SD = 30.01) . The remaining effects did not reach significance. Specifically, the analysis did not detect a significant main effect of reward on the previous trial ($F(1, 23) = 0.003$, $p = 0.953$, $\eta_p^2 = 0.000$) or a significant interaction between previous reward and task sequence ($F(1, 23) = 0.073$, $p = 0.790$, $\eta_p^2 = 0.003$). In addition, the analysis did not detect a significant three-factor interaction between previous reward, current reward and task sequence ($F(1, 23) = 2.741$, $p = 0.111$, $\eta_p^2 = 0.106$). However, we note that this could be due to a lack of power. When including RTs where a correct response was made on the current and previous trials from the two 72 trial practice blocks, the interaction

between current reward and task sequence reached significance ($F(1, 23) = 4.475$, $p = 0.045$, $\eta_p^2 = 0.163$) and there was a strong trend towards a three-factor interaction ($F(1, 23) = 4.179$, $p = 0.053$, $\eta_p^2 = 0.154$), wherein switch costs were numerically lowest for the low-to-high reward sequence. These exploratory analyses are presented in the appendix for this chapter.

Current reward and task sequence showed independent accuracy modulation

The equivalent analysis for accuracy showed the same pattern of results. Specifically, the analysis showed a significant main effect of current reward ($F(1, 23) = 16.698$, $p < 0.001$, $\eta_p^2 = 0.421$), reflecting lower accuracy on low reward trials (M = 0.85, SD = 0.08) compared with high reward trials (M = 0.88, SD = 0.07). The analysis also showed a significant main effect of task sequence ($F(1, 23) = 33.053$, $p < 0.001$, $\eta_p^2 = 0.590$), reflecting lower accuracy on switch trials (M = 0.83, SD = 0.08) compared with repeat trials (M = 0.89, SD = 0.06). All other effects did not reach significance. Specifically, the analysis did not detect a significant main effect of reward on the previous trial ($F(1, 23) = 0.436$, $p = 0.516$, $\eta_p^2 = 0.019$), or significant interactions between previous reward and task sequence ($F(1, 23) = 0.007$, $p = 0.933$, $\eta_p^2 = 0.000$), current reward and task sequence ($F(1, 23) = 2.844$, $p = 0.105$, $\eta_p^2 = 0.110$), or reward on the previous and current trials ($F(1, 23) = 0.043$, $p = 0.838$, $\eta_p^2 = 0.002$). In addition, the analysis did not detect a significant three-factor interaction between previous reward, current reward and task sequence ($F(1, 23) = 0.156$, $p = 0.697$, $\eta_p^2 = 0.007$).

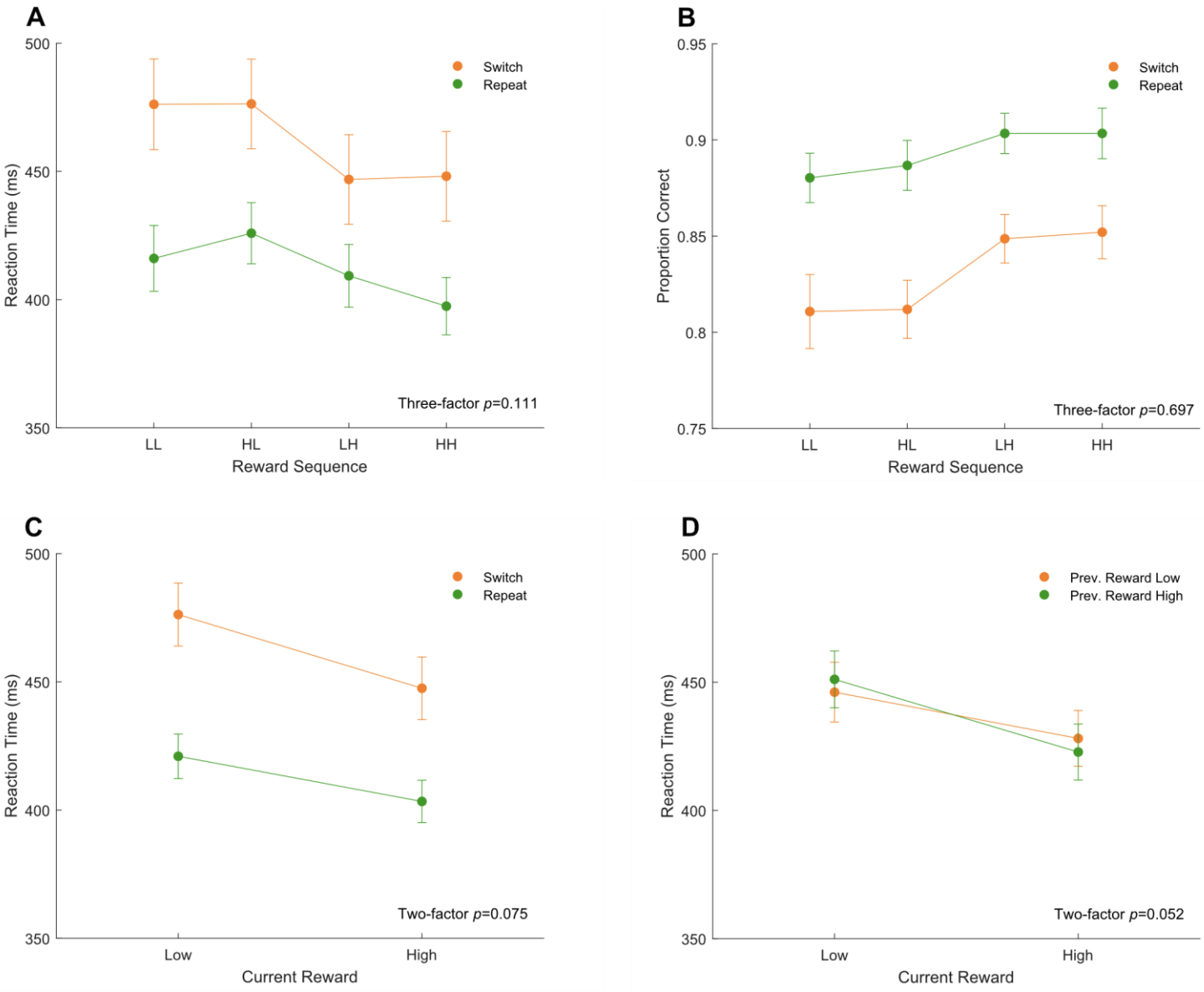


Figure 7. Behavioural performance in Experiment 2.3. A: Reaction time for reward magnitude sequences, separated for switch and repeat trials. B: Proportion correct for reward magnitude sequences, separated for switch and repeat trials. A-B: LL, HL, LH and HH indicate low-to-low, high-to-low, low-to-high and high-to-high reward sequences. Three-factor p-values indicate the interaction between previous reward, current reward and task sequence. C: Reaction time for switch and repeat trials as function of reward prospect on the current trial. The two-factor p-value indicates the interaction between current trial reward prospect and task sequence. D: Reaction time as a function of reward prospect on the current and previous trials. The two-factor p-value indicates the interaction between current and previous trial reward on RT. A-D: Error bars show standard error of mean.

Discussion

In Experiment 2.2, we observed that high reward on the previous trial reduced RT switch costs on the current trial. Experiment 2.3, we tested whether increasing the ITI would separate processing of previous reward outcome from current reward prospect, and unmask a sequential reward interaction on rule switching (Kleinsorge & Rinkenauer, 2012; Shen & Chun, 2011). Consistent with this possibility, exploratory analyses showed a strong trend toward a sequential reward effect, in which switch costs for the low-to-high reward sequence were reduced. Results also showed a trend towards lower switch costs when current reward prospect was high. This effect was significant in exploratory analyses that increased power by including trials from the practice blocks (see appendix). Regardless of task sequence, responses were significantly faster and more accurate on trials with high reward prospect. The present results provide trend-level support for the notion that sequential increases in reward magnitude might support flexible rule updating, needed to switch to new tasks (Fröber & Dreisbach, 2016; Kleinsorge & Rinkenauer, 2012; Shen & Chun, 2011).

General Discussion

This series of behavioural studies aimed to develop a task that could be used to investigate the sequential reward effects on flexible cognitive control. Experiment 2.1 focused on how reward probability sequences influence cued and voluntary task switching, but did not detect effects of reward sequence on cued switching or the voluntary switch rate. Experiments 2.2 and 2.3 therefore focused on replicating reported sequential effects of reward magnitude on cued switching (Kleinsorge & Rinkenauer, 2012; Shen & Chun, 2011). In contrast to Experiment 2.1, Experiment 2.2 separated reward from rule cue presentation and added a short ITI (100-700ms). The results from Experiment 2.2 showed faster RTs on high reward trials and reduced RT switch costs when previous reward was high. Experiment 2.3 extended the ITI duration (900-1500ms). Results showed faster RTs and accuracy on high reward trials, as well as a trend towards a reduced switch cost when current reward was high. Exploratory analyses showed a significant reduction in switch costs on high reward trials, as well as a strong trend ($p=0.053$) towards a sequential reward interaction, in which switch costs were lowest for the low-to-high reward sequence.

The exploratory analyses from Experiment 2.3 are partially consistent with previous studies motivating this work (Fröber & Dreisbach, 2016; Kleinsorge & Rinkenauer, 2012; Shen & Chun, 2011). One central theoretical question these previous studies aimed to address is how reward motivation is able to both shield goal-relevant information from distraction but update this information flexibly when the environment changes. Experiment 2.3 provides trend-level support for the proposal that low-to-high reward transitions aid flexible rule updating (Shen & Chun, 2011). However, the present data do not provide support for the proposal that consistently high reward prospect promotes task stability (Fröber & Dreisbach, 2016; Kleinsorge & Rinkenauer, 2012). This was evident in follow up

analyses of the three-factor trend in Experiment 2.3, which indicated that switch costs were reduced for low-to-high reward trials, in comparison to low-to-low but not high-to-high reward trials. From a neural perspective, increased flexibility for low-to-high compared with low-to-low reward transitions might be interpreted under a model of dopamine gating (Braver & Cohen, 2000; Westbrook & Braver, 2016). More specifically, higher reward values following low reward trial could induce a phasic dopamine response, allowing new task information to be encoded with greater fidelity in PFC on low-to-high reward trials compared with low-to-low reward trials.

Overall, this series of studies was partially successful in meeting its aim of developing a task to study sequential reward effects on flexible control in a subsequent EEG experiment. While Experiment 2.3 showed evidence towards a three-factor interaction, wherein switch costs were lowest on low-to-high reward trials, the interaction was not observed in Experiments 2.1 and 2.2. Given that these inter-trial effects were unreliable across experiments and involve many conditions, we focused our subsequent EEG investigation on how current reward prospect modulates task rule encoding on switch trials. This allowed us to increase statistical power by comparing two reward levels, as opposed to four reward transitions.

Chapter 2 Appendix

A1. Supplementary analyses for experiment 2.1.

The voluntary switch rate (VSR) was our main dependent measure for free choice trials in experiment 2.1. However, we also performed 2x2x2 (previous reward probability x current reward probability x task sequence) on RT and accuracy for these trials. The analysis for RT showed a significant main effect of task sequence ($F(1, 23) = 8.575$, $p = 0.008$, $\eta_p^2 = 0.272$), reflecting significantly slower RTs on switch trials (M = 755ms, SD =224.68) compared with repeat trials (M = 688ms, SD =179.12). The analysis did not detect significant main effects of reward probability from the previous trial ($F(1, 23) = 2.096$, $p = 0.161$, $\eta_p^2 = 0.084$) or the current trial ($F(1, 23) = 0.719$, $p = 0.405$, $\eta_p^2 = 0.030$). Similarly, the analysis did not detect significant interactions between task sequence and reward probability on the previous trial ($F(1,23) = 0.002$, $p = 0.967$, $\eta_p^2 = 0.000$), task sequence and reward probability on the current trial ($F(1,23) = 2.652$, $p = 0.117$, $\eta_p^2 = 0.103$), reward probability on the current and previous trials ($F(1,23) = 0.805$, $p = 0.379$, $\eta_p^2 = 0.034$), or task sequence, previous and current reward probability ($F(1,23) = 1.600$, $p = 0.219$, $\eta_p^2 = 0.065$).

The equivalent 2x2x2 (previous reward x current reward x task sequence) repeated measures ANOVA for accuracy did not detect a significant main effect of task sequence on ($F(1, 23) = 0.023$, $p = 0.881$, $\eta_p^2 = 0.001$). The analysis did not detect significant main effects of reward probability from the previous trial ($F(1, 23) = 0.767$, $p = 0.390$, $\eta_p^2 = 0.032$) or the current trial ($F(1, 23) = 0.050$, $p = 0.825$, $\eta_p^2 = 0.002$). Similarly, the analysis did not detect significant interactions between task sequence and reward probability on the previous trial ($F(1,23) = 0.075$, $p = 0.787$, $\eta_p^2 = 0.003$), task sequence and reward probability on the current trial ($F(1,23) = 0.195$, $p = 0.663$, $\eta_p^2 = 0.008$), reward probability on the current and

previous trials ($F(1,23) = 0.086$, $p = 0.772$, $\eta_p^2 = 0.004$) or task sequence, previous and current reward probability ($F(1,23) = 0.883$, $p = 0.357$, $\eta_p^2 = 0.037$).

A2. Supplementary analyses for experiment 2.3.

We explored whether trends towards reward-by-task sequence interactions were influenced by a lack of power. To test this, we included RTs where a correct response was made on the current and previous trials from the two 72-trial practice blocks in our 2x2x2 (previous reward x current reward x trial transition) repeated measures ANOVAs.

The 2x2x2 (previous reward x current reward x trial transition) repeated measures ANOVA on RT showed a significant main effect of current reward ($F(1, 23) = 14.173$, $p = 0.001$, $\eta_p^2 = 0.381$), reflecting slower RTs when current reward was low (M = 455ms, SD = 77.27) than when it was high (M = 433, SD = 76.56). The analysis detected a significant main effect of task sequence ($F(1, 23) = 36.687$, $p < 0.001$, $\eta_p^2 = 0.615$), reflecting slower RTs on switch trials (M = 469, SD = 86.30) compared with repeat trials (M = 419ms, SD = 57.89). There was also a significant interaction between current reward and task sequence ($F(1, 23) = 4.475$, $p = 0.045$, $\eta_p^2 = 0.163$), driven by a reduction in the switch cost on high (M = 44ms, SD = 51.18) compared with low reward trials (M = 57ms, SD = 44.59). In addition to these effects, we detected two trends. First, we found a trend towards a main effect of previous trial reward ($F(1, 23) = 3.584$, $p = 0.071$, $\eta_p^2 = 0.135$), reflecting a numerical decrease in RT following low reward trials (M = 442ms, SD = 76) compared with high reward trials (M = 446ms, SD = 79.29). Second, we found a trend towards a three-factor interaction between previous reward, current reward and task sequence ($F(1, 23) = 4.179$, $p = 0.053$, $\eta_p^2 = 0.154$). Follow up t-tests comparing switch costs between different reward sequences showed that this was driven by a significant reduction in the switch cost for low-to-high reward

transitions (M = 35.28ms, SD = 44.73), compared with low-to-low reward transitions (M = 62.31ms, SD = 49.01) (t(23)=3.555, Bonferroni corrected p=0.0120). The five remaining pairwise comparisons did not reach significance (all corrected p-values > 0.2), including the comparison between switch costs on low-to-high reward trials versus high-to-high reward trials (M = 53.07, SD = 57.44) (t(23)=-2.223, uncorrected p = 0.036, Bonferroni corrected p = 0.2160). The 2x2x2 ANOVA on RT did not detect significant interactions between previous reward and task sequence (F(1, 23) = 0.358, p = 0.556, η_p^2 = 0.015) or an interaction between previous and current reward (F(1, 23) = 1.040, p = 0.319, η_p^2 = 0.043).

The equivalent analysis for accuracy showed a significant main effect of current reward (F(1, 23) = 10.838, p = 0.003, η_p^2 = 0.320), reflecting lower accuracy on low reward trials (M = 0.84, SD = 0.08) compared with high reward trials (M = 0.86, SD = 0.07). The analysis also showed a significant main effect of task sequence (F(1, 23) = 46.377, p < 0.001, η_p^2 = 0.668), reflecting lower accuracy on switch trials (M = 0.82, SD = 0.07) compared with repeat trials (M = 0.86, SD = 0.06). All other effects did not reach significance. Specifically, the analysis did not detect a significant main effect of reward on the previous trial (F(1, 23) = 0.273, p = 0.606, η_p^2 = 0.012), or significant interactions between previous reward and task sequence (F(1, 23) = 0.119, p = 0.734, η_p^2 = 0.005), current reward and task sequence (F(1, 23) = 1.823, p = 0.190, η_p^2 = 0.073), reward on the previous and current trials (F(1, 23) = 0.067, p = 0.797, η_p^2 = 0.003). In addition, the analysis did not detect a significant three-factor interaction between previous reward, current reward and task sequence (F(1, 23) = 0.163, p = 0.691, η_p^2 = 0.007).

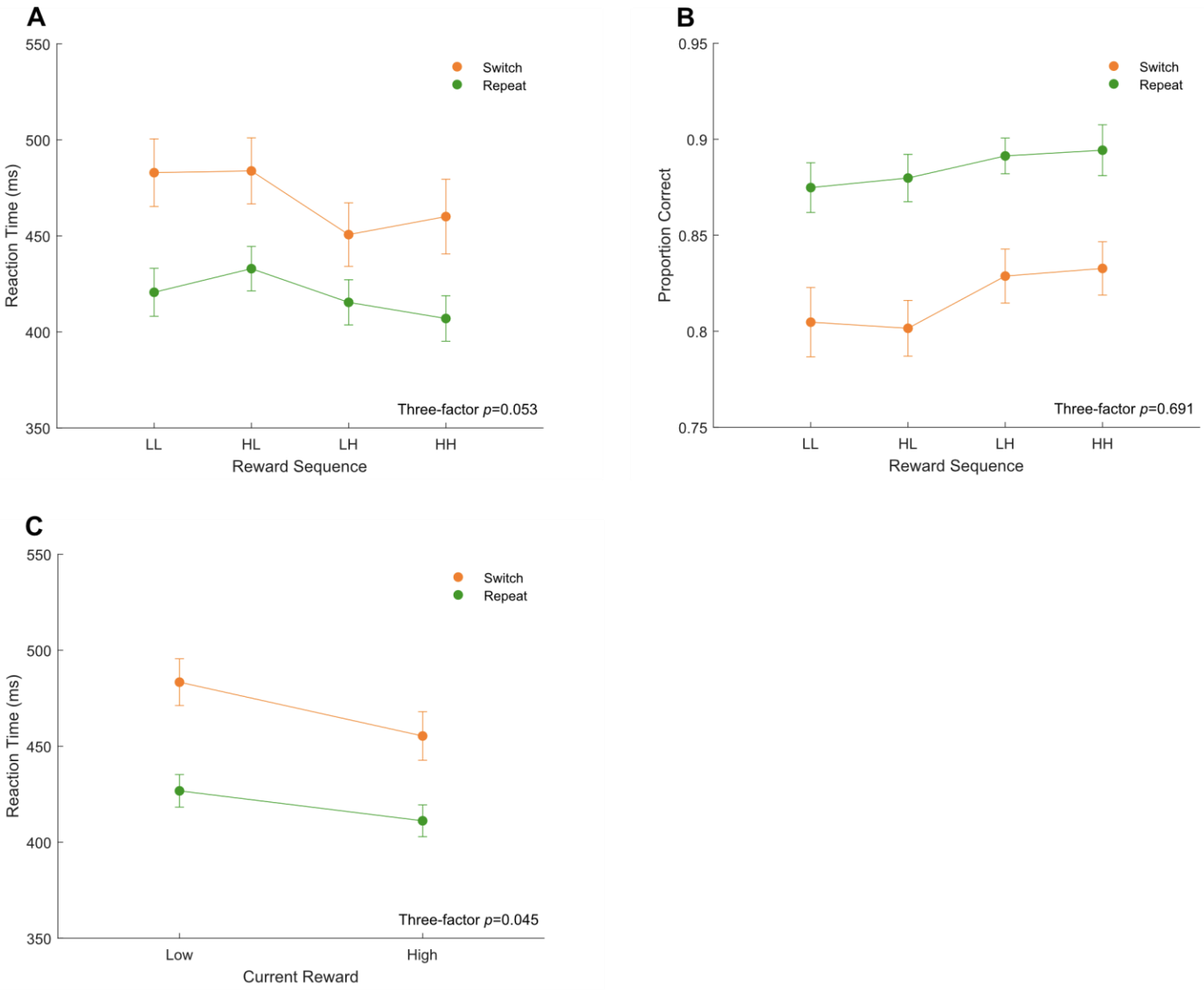


Figure A1. Exploratory analysis of Experiment 2.3, in which practice trials were included to boost statistical power. A: Reaction time for reward magnitude sequences, separated for switch and repeat trials. B: Proportion correct for reward magnitude sequences, separated for switch and repeat trials. A-B: LL, HL, LH and HH indicate low-to-low, high-to-low, low-to-high and high-to-high reward sequences. Three-factor p-values indicate the interaction between previous reward, current reward and task sequence. C: Reaction time for switch and repeat trials as function of reward prospect on the current trial. The two-factor p-value indicates the interaction between current trial reward prospect and task sequence. A-C: Error bars show standard error of mean.

Chapter 3. Reward boosts neural coding of task rules to optimise flexible control[1]

Abstract

Cognitive flexibility is critical for intelligent behaviour. However, its execution is effortful and often suboptimal. Recent work indicates that flexible behaviour can be improved by the prospect of reward, which suggests that rewards optimise flexible control processes. In experiment 3.1, we investigated how different reward prospects influence neural encoding of task rule information to optimise cognitive flexibility. We applied representational similarity analysis (RSA) to human electroencephalograms, recorded while participants performed a rule-guided decision-making task. During the task, the prospect of reward varied from trial to trial. Participants made faster, more accurate judgements on high reward trials. Critically, high reward boosted neural coding of the active task rule and the extent of this increase was associated with improvements in task performance. Additionally, the effect of high reward on task rule coding was most pronounced on switch trials, where rules were updated relative to the previous trial. These results suggest that reward prospect can promote cognitive performance by strengthening neural coding of task rule information, helping to improve cognitive flexibility during complex behaviour. While experiment 3.1 demonstrated that high reward can strengthen task representations over short time scales (i.e. within a trial), we speculated that high reward could also lead to more persistent encoding of the task rule over longer time scales (i.e. between trials). While beneficial for current trial performance, such persistence could impair performance on subsequent trials involving a rule switch. We tested this prediction in experiment 3.2, where we investigated whether high reward feedback would reduce performance on subsequent

[1] Most of this chapter is published in: Hall-McMaster, S., Muhle-Karbe, P. S., Myers, N. E., & Stokes, M. G. (2019). Reward boosts neural coding of task rules to optimise cognitive flexibility. *Journal of Neuroscience.* The exception to this is experiment 3.2, which aims to follow up the findings from 3.1.

switch trials, due to greater interference between current and previous task representations. The results from this behavioural follow up showed no effect of previous trial reward feedback on current switching performance. We interpret this result as an indication that our design was not sensitive to the potential effect of high reward outcomes on strengthening active task representations and thereby creating interference with new rule information during task switches.

Experiment 3.1

Introduction

Flexible cognitive control is critical to human intelligence. When vying to win a card game, we can use arbitrary rules to play the best hand. When navigating a new city, we can apply navigation rules to sensory input from the world around us to arrive at the next tourist attraction. Controlled processes requiring flexibility result in slower performance, compared to behaviours that require less flexible processing (Kleinsorge & Rinkenauer, 2012; Shen & Chun, 2011). However, this can be improved by motivational factors, such as the prospect of reward for speed and accuracy (Braem & Egner, 2018). When the prospect of reward is high, performance improves on flexible rule-based tasks (Etzel et al., 2016; Kleinsorge & Rinkenauer, 2012), suggesting that reward might optimise cognitive flexibility by fine-tuning rule-based neural coding patterns.

The existing neuroimaging literature provides support for this perspective, indicating that reward prospect leads to stronger recruitment of frontoparietal brain regions implicated in cognitive control (Parro et al., 2018). When reward cues are presented at the beginning of each trial, activity in these regions is typically enhanced prior to the onset of a target stimulus, suggesting that proactive control mechanisms contribute to reward-induced performance benefits (Engelmann et al., 2009; Krebs et al., 2012; Padmala & Pessoa, 2011). Critically, a recent study combining fMRI with pattern classification methods found that frontoparietal activity prior to target onset encoded abstract task rules with greater fidelity on reward trials than on no-reward trials (Etzel et al., 2016). The extent to which reward prospect enhanced the decodability of task rules also mediated behavioural benefits, suggesting that performance improvements may result from reward-driven tuning in cognitive control processes that prioritise task-relevant processing.

This proposal has considerable theoretical appeal. Yet, it is unknown whether reward prospect differentially influences rule coding in situations that require more or less flexible processing. The present study therefore aimed to characterise how changes in reward influence neural coding of task rule information when rules must be updated. To do so, we developed a behavioural task that involved switching between task rules, as reward prospect was manipulated. We then applied representational similarity analysis (RSA) (Kriegeskorte et al., 2008) to human electroencephalograms to examine changes in neural coding as a function of reward prospect. Based on Etzel et al. (2016), we predicted that task rule coding would be greater with high reward prospect and that the extent of this increase would track improvements in behavioural performance. However, the critical contribution of this work was to consider how reward prospect would impact rule coding when flexible processing was needed. Based on prior work showing selective reward enhancements on rule switching (Kleinsorge & Rinkenauer, 2012; Shen & Chun, 2011), we reasoned that task rule coding should be greatest on high reward switch trials, where task rules must be updated. Switch trials involve the most interference between rule codes and thus increased neural separation between them should support flexible rule updating (Waskom et al., 2014). Reward prospect could therefore optimise flexible processing by helping to increase neural dissimilarity between rules, when they come into conflict. RSA is well suited to test this prediction because it provides a sensitive index of reward-driven changes in neural dissimilarity, while allowing clean separation of overlapping task variables in the lead up to adaptive behavioural responses.

To summarise the main results, we found that high reward prospect produced significant performance improvements in accuracy and reaction time (RT). Consistent with the view that reward prospect increases proactive cognitive control, we found a significant

increase in neural coding for task rules under high reward conditions prior to the onset of a target stimulus. The average difference in rule coding between reward conditions during this period also correlated with RT improvements. Especially striking, reward's effect on rule coding was most pronounced in situations requiring the most flexible processing, when task rules were updated relative to the previous trial.

Method

Participants

We set a target sample size of 30 participants. During recruitment three participants were excluded, one due to a corrupt EEG recording and two due to excessive artefacts that led to the rejection of more than 120/650 trials. We therefore collected three more participants to reach the 30 participant target. The final sample were between 18 and 35 years of age (mean age = 23, 19 female), with normal or corrected-to-normal vision, who reported no history of neurological or psychiatric illness. Participants received £8 per hour or course credit for taking part and could earn up to £10 extra for their performance. This study was approved by the Central University Research Ethics Committee at the University of Oxford and all participants signed informed consent before taking part.

Materials

Stimuli were presented on a 22-inch screen with a spatial resolution of 1280 x 1024 and refresh rate of 60Hz. Stimulus presentation was controlled using Psychophysics Toolbox-3 (RRID: SCR_002881) in MATLAB (RRID: SCR_001622, version R2015b). Reward cues and feedback shown during the task were presented in size 30 Arial font. Task cues and target stimuli had approximate visual angles of 2.52° (100x100 pixels) and 1.26° (50x50 pixels) respectively, with visual angles calculated based on an approximate viewing distance of 60cm. F and J keys on a standard QWERTY keyboard were used to record left and right hand responses. EEG data were recorded with 61 Ag/AgCl sintered electrodes (EasyCap, Herrsching, Germany), a NeuroScan SynAmps RT amplifier, and Curry 7 acquisition software (RRID: SCR_009546). EEG data were pre-processed in EEGLAB (RRID: SCR_007292, version 14.1.1b) (Delorme & Makeig, 2004), behavioural analyses were performed using JASP (RRID:

SCR_015823, version 0.8.1.3) and EEG analyses were performed in MATLAB (version 2018b), using the FieldTrip toolbox (RRID: SCR_004849) as well as custom scripts.

Code Accessibility

Task and analysis code, as well as raw and pre-processed data can be accessed at: https://osf.io/kuzye/

Experimental Design and Statistical Analysis

In this task (Figure 1), participants' overarching goal was to gain as many points as possible. To do so, participants categorized bi-dimensional target stimuli based on their colour (yellow vs blue) or shape (square vs circle). On each trial, only one feature dimension of the target (colour or shape) was relevant to gaining points, while the other feature served as an irrelevant distractor. The relevant feature dimension was signalled through a visual task cue prior to target onset. In addition to a single relevant feature dimension, each trial offered a high or low reward magnitude for making a correct response. This was signalled to participants at the beginning of each trial by a single pound sign (low reward: 5-10 points) or three pound signs (high reward: 50-100 points).

The experimental sequence consisted of a reward cue, task cue, target, feedback screen and an inter-trial interval (ITI). The reward cue (£ or £££) was first presented for 800ms, followed by a 400ms delay. The task cue (one of four possible abstract shapes) was then presented for 200ms. The mapping of cues to tasks was counterbalanced between participants. Task cue offset was followed by a 400ms delay. The target (a yellow square, blue square, yellow circle or blue circle) was then presented and remained on screen until a response was given or for a maximum of 1400ms. If the active task rule was colour, the correct response mapping was ‘f’ or ‘j’ for yellow or blue targets respectively. If the active rule was shape, the correct response mapping was ‘f’ or ‘j’ for square or circle targets

respectively. The response phase was followed by feedback lasting 200ms. An incorrect response or omission resulted in feedback showing '+0' points. A correct response resulted in feedback showing '+X', where X was a value within the high or low reward point ranges, the precise value of which was determined by RT. More specifically, RT criteria for different points were initialised so that responses faster than 400ms, 600ms, 800ms, 1000ms, 1200ms and 1400ms earned 100/10, 90/9, 80/8, 70/7, 60/6 and 50/5 points on high and low reward trials respectively. For correct trials, the current trial RT was added to an array for its reward condition. When each array contained more than six values, individualised points criteria were calculated for that condition and were calculated again every time a new value entered the array. The individualised points criteria followed criteria outlined by Shen and Chun (2011), in which the most (to least) points are rewarded for correct responses faster than 95%, 80%, %65, %50 and 35% of the median RT. The trial concluded with a randomly selected ITI duration, drawn from a uniform distribution with values of 1000, 1100, 1200, 1300 or 1400ms. Participants were trained to reach a criterion of 70% accuracy before completing 10 experimental blocks of 65 trials. Excluding the first trial in each block, equal numbers of reward cues, task cues, and stimuli were presented within each block. Equal numbers of ITI durations were presented across blocks. Presentation was pseudo-randomised to ensure trials were balanced based on task, target-congruency and task sequence for each reward condition. Target congruency refers to whether task-relevant and irrelevant target features are mapped to the same (congruent) or different (incongruent) response hands. Task sequence refers to whether the task rule was different from the previous trial (switch trial) or the same as the previous trial (repeat trial).

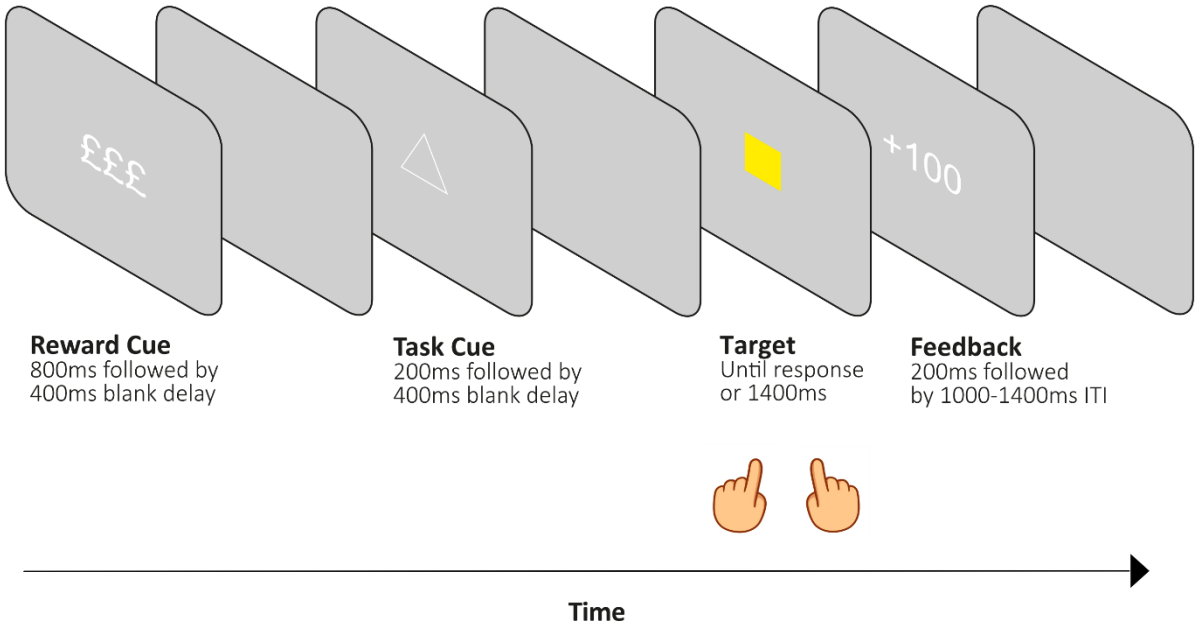


Figure 1. Task design. On each trial, a high or low reward cue was presented followed by a blank delay. A task rule cue was then presented, indicating whether participants should respond to the upcoming target based its colour or shape. Following a second blank delay, a bi-dimensional target (a coloured shape) appeared until a response was given or for a maximum duration of 1400ms. This was followed by feedback (based on accuracy and reaction time) and a variable inter-trial-interval.

Behavioural measures. Our main dependent measures were reaction time (RT) and the proportion of correct responses (accuracy). We calculated median RTs for the respective design cells to account for the skewness of RT distributions (Ratcliff, 1993). For all RT analyses, we only included trials in which the current and previous trials were correct to mitigate potential effects of post-error slowing, in which participants tend respond more slowly after making an error (Dutilh et al., 2012; Notebaert et al., 2009). For accuracy analyses, we included all trials in which a response was made within the 1.4s response window. Behavioural data were analysed using 2x2 repeated measures ANOVAs with factors of reward and task sequence. The task sequence factor consisted of switch trials, where the task rule was different from the previous trial, and repeat trials, where the task

rule was the same as the previous trial. In addition, we used a paired sample t-test for normally distributed data and a Wilcoxon signed-rank test for data showing significant deviation from normality, to compare performance on task-switch trials and cue-switch trials. This analysis served as a behavioural control. It has been argued that switch costs could arise simply from processes related to processing the cue stimulus, rather than changes in task rule per se (Logan & Bundesen, 2003). Comparing trials where both task cue and task rule changed, with trials where the cue changed but the rule stayed the same allowed us to delineate the extent to which switch costs were driven by changes in task rules and task cues. More specifically, if the effect of rule switching on behavioural performance was primarily driven by rule changes, we should see greater performance reductions on trials where the rule is changed relative to trials where the rule is repeated but the cue is changed.

EEG pre-processing. EEG data were down-sampled from 1000 to 250Hz and filtered using 40Hz low-pass and 0.01Hz high-pass filters. For each participant, channels with excessive noise were identified by visual inspection and replaced via interpolation, using a weighted average of the surrounding electrodes. Data were then re-referenced by subtracting the mean activation across all electrodes from each individual electrode at each time point. Data were divided into epochs from -1 to +5 seconds from the onset of the reward cues. Epochs containing artefacts (such as muscle activity) were rejected based on visual inspection. In the final stage of pre-processing, data were subjected to an Independent Component Analysis (ICA). Structured noise components, such as eye blinks, were removed, resulting in the data set used for subsequent analyses. Prior to each analysis, data were z-scored over the trial dimension and baseline-corrected using a time window of 200ms to 50ms prior to the trial event of interest (e.g. cue or target presentation).

EEG analyses. We used RSA (Kriegeskorte et al., 2008) to investigate how reward prospect influenced neural coding for different kinds of task-relevant information. There were two main advantages of using RSA to address this question. First, multivariate approaches leverage pattern information that would normally be averaged out in univariate analyses. This makes multivariate methods more sensitive to effects based on distributed patterns of neural activity (Kriegeskorte et al., 2006). Second, RSA allowed us to examine multiple, overlapping neural codes. In particular, the combination of RSA with models for different task variables provided a powerful method for separating neural coding of task rules, relevant and irrelevant target features, as well as motor responses. This ability to separate overlapping neural activity would be difficult to achieve with more traditional EEG methods, such as the analysis of event-related potentials.

In addition, we selected EEG for its high temporal resolution. Constraints on the temporal resolution of fMRI can make it challenging to isolate task rule coding from subsequent perceptual processing because the slow hemodynamic response can make it difficult to pinpoint effects in time and distinguish sustained anticipatory activity from transient stimulus-evoked responses. In contrast, high temporal resolution methods such as electroencephalography are needed because their ability to distinguish rapid stimulus-evoked dynamics makes them ideal for isolating the effects of reward on task rule coding from subsequent neural coding patterns.

The logic of our approach was to characterise neural coding patterns elicited by different trial conditions and test whether reward prospect led to more distinct task representations for the two tasks being performed (colour vs shape judgements)(Etzel et al., 2016; Westbrook & Braver, 2016). We were especially interested in the effect of reward on proactive control mechanisms, wherein goal-relevant information is encoded in preparation

for upcoming task demands. This stands in contrast to reactive control processes, which serve to resolve task demands after their detection (Braver, 2012). Evidence of reward-modulated proactive control in our design would be seen as a difference in rule coding between high and low reward conditions, prior to target onset. Moreover, the difference in rule coding during this period should also be associated with improvements in behavioural performance (Etzel et al., 2016). Building on the results of Etzel et al. (2016) a central goal in the present work was to examine the effect reward prospect on rule coding during switch trials, when rules required flexible updating. Here we reasoned that switch trials involve the greatest interference between rule representations and thus reward prospect could improve flexible cognitive control (Kleinsorge & Rinkenauer, 2012; Shen & Chun, 2011) by increasing neural dissimilarity between rule codes. As a secondary aim, we performed theory-driven analyses that tested whether high reward lead to stronger neural coding for task-relevant compared to irrelevant target features, and whether relevant feature prioritisation was associated with improved performance (Pessoa, 2017). Finally, we performed exploratory analyses that tested whether high reward modulated motor response coding and whether reward-induced changes in task coding were associated with downstream changes in sensorimotor processing. All analyses used data from all 61 EEG channels. While this limits inference about regional sources of neural activity, it has the advantage of including all available data without any assumptions about source localization. Incorrect and omission trials were excluded from EEG analyses.

Neural coding across the trial. For each participant, trials were divided into conditions based on reward condition (low, high), as well as task-relevant and irrelevant target features. Dividing the trials this way also implicitly divided trials by task. If the task-relevant target feature was yellow, for example, then the task must have been to judge the target

colour. We then averaged trials in each condition to get an array with channels x time points x conditions. To measure neural dissimilarity between conditions, we used Mahalanobis distances (MDs). This distance metric shows similar reliability to correlation distance measures (Walther et al., 2016) but explicitly takes covariance into account, making it well suited to EEG data where channel values tend to be highly correlated. For EEG data, the MD between two conditions is computed at each time point, using two key pieces of information. The first is the difference in topographies between two conditions. A topography is a vector of 61 channel values that has been averaged across all trials in a condition, for a specific time point (Figure 2A). The second key piece of information is the channel covariance matrix. This is computed using a matrix of trials x channels. For this study, we used within-condition error to compute the channel covariance matrix (Walther et al., 2016). This meant that prior to computing the covariance matrix, trials in the trial x channel matrix were mean-centered by subtracting the mean topography for each condition from all trials within that condition. The covariance calculation also used shrinkage estimator (Ledoit & Wolf, 2004), which has the effect of down-weighting noisy covariance estimates. Using this information, the MD is formally computed as: $MD_{AB} = \sqrt{(Pattern\ A - Pattern\ B)^T \times Cov^{-1}(Pattern\ A - Pattern\ B)}$, where Pattern A-Pattern B is the difference between topographies, T is the transpose and Cov-1 is the inverse of the channel covariance matrix. This calculation is also presented in Figure 2. MDs were calculated between all topographies for the sixteen conditions, separately at each time point. This procedure yielded a 16 x 16 representational dissimilarity matrix (RDM) of multivariate condition distances for each time point and participant. We then constructed a set of five 16 x 16 model RDMs to capture neural coding patterns related to different task variables. These variables included reward coding, task coding, task-relevant feature coding,

task-irrelevant feature coding, and motor coding. The logic of all models was to place zeroes in cells of a 16x16 matrix where conditions matched on the variable of interest and ones in all remaining cells. For example, the task coding model was a 16x16 matrix containing zeros in cells where two conditions involved the same task (e.g. both colour judgements) and ones in cells where two conditions involved different tasks (e.g. shape vs colour judgements). The task-relevant feature model was a 16x16 matrix containing zeros in cells where two conditions had the same task-relevant target feature (e.g. both yellow on colour trials) and ones in cells where two conditions had different task-relevant features (e.g. yellow vs blue or yellow vs square). Data RDMs (not z-scored and averaged for illustration) and model RDMs for all task variables are presented in Figure 4. For analysis, data and model RDMs were z-scored. As RDMs are symmetric, the upper triangular portion of each matrix was transformed into a vector. The resulting data and model distance vectors were then entered into a multiple regression analysis that was conducted at each time point (4ms apart after down-sampling). The data-derived distance vector was the dependent variable and the model-derived distance vectors were independent variables. The regression also included a constant to model the intercept of the linear regression equation. This led to the following general linear model (where DV stands for distance vector): $Data\ DV = \beta 0(intercept) + (\beta 1\ x\ reward\ DV) + (\beta 2\ x\ task\ DV) + (\beta 3\ x\ task\ relevant\ feature\ DV) + (\beta 4$ x task irrelevant feature DV$) + (\beta 5\ x\ motor\ response\ DV) + \varepsilon(residual\ error)$. The regression procedure was performed three times, once with a baseline window prior to reward cue onset, once prior to task cue onset and once prior to the target onset. Running the regression procedure at each time point produced a time course of regression coefficient estimates, one for each coding model. We refer to coefficient estimates as 'model fit to neural dissimilarity matrix' in subsequent figures. We interpret the magnitude

of these coefficient estimates to reflect the magnitude of neural dissimilarity/neural coding for each of the task variables across time.

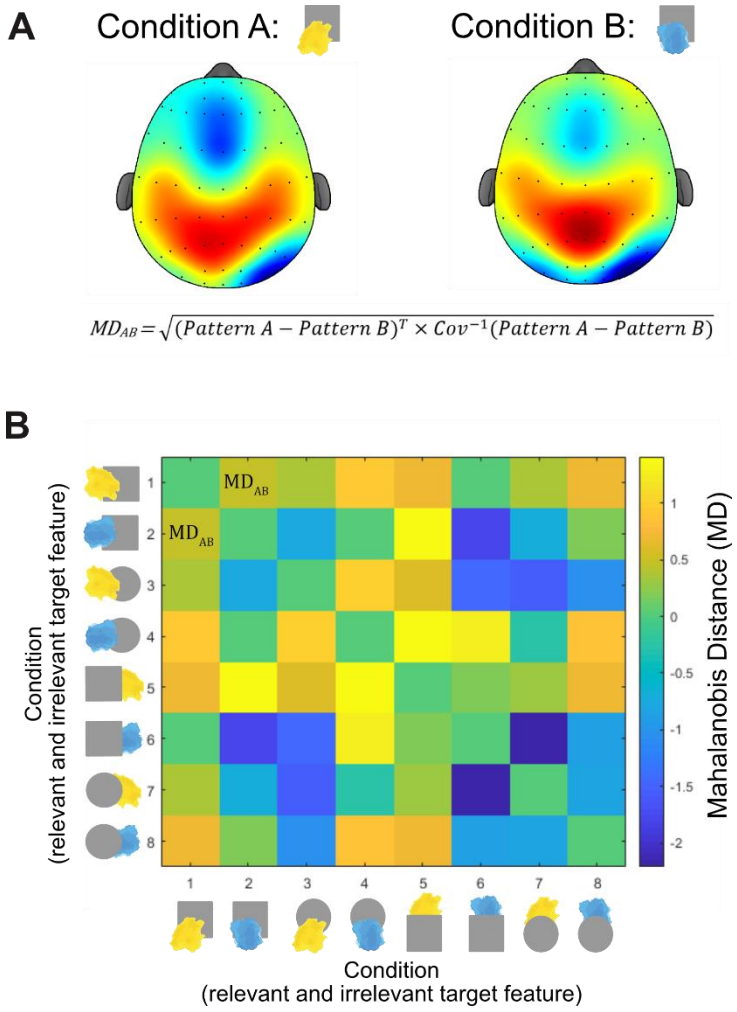


Figure 2. Logic of Representational Similarity Analysis (RSA). A: Trials were divided into conditions based on the relevant and irrelevant features of target stimuli. The dissimilarity between activity patterns was computed as the Mahalanobis distance (MD), which captures the multivariate distance between topographies. B: MDs between conditions are entered into the relevant cells of a representational dissimilarity matrix (RDM). The process outlined in A is repeated until the MDs between each pair of conditions has been computed. The process is then repeated at the next time point, and for subsequent time points of interest. The data RDM produced in B is then regressed against model RDMs that reflect predicted differences in dissimilarity structure for different task variables. These model RDMs are shown in Figure 4.

Neural coding as a function of reward prospect. We repeated the analyses above separately for high and low reward trials. For each participant, trials were divided based on the task-relevant and irrelevant features of the target. Conditions were then averaged over the trial dimension. Mahalanobis distances were calculated between all scalp topographies for the eight conditions, separately at each time point, generating a set of 8x8 RDMs. For these analyses, model RDMs were generated to reflect condition differences based on task coding, the task-relevant feature of the target, the task-irrelevant feature of the target and the motor response. These models followed the same logic as those described in the previous section: zeroes were placed in cells of an 8x8 matrix where two conditions did not differ in the variable of interest (e.g. the task-relevant target feature) and ones were placed in cells where conditions differed in the variable of interest. Model data RDMs were z-scored and entered into a single regression at each time point (including a constant). The regression procedure was performed twice, once using only low reward trials and once using only high reward trials. For this set of analyses, it was particularly important to match the number of high and low reward trials, so that regression coefficients were not higher for high reward conditions because more data was being included in the regression. To address this, these analyses subsampled trials to match the number of high and low reward trials. 100 iterations were performed per participant and final regression coefficients resulted from averaging the estimated regression coefficients over iterations. For task coding, we performed a theory-driven one tailed t-test to assess whether tasking coding was significantly greater under high reward conditions during the pre-target period (averaged from 1400-1800ms). This test was based on Etzel et al. (2016) and remained significant when non-directional. For reward-modulated coding of the relevant target feature, we tested whether the observed results could be driven by correlations with other coding

models. To do so, we ran a control analysis in which we first regressed task coding, task-irrelevant feature coding and motor coding against participants' neural dissimilarity matrices. We then regressed the task-relevant feature coding model against the residual variance, which had not been accounted for by the other models.

Neural coding as a function of rule updating. We examined reward-modulated neural coding separately on switch and repeat trials. This was motivated by previous work showing reward prospect benefits switching performance (Shen & Chun, 2011) and can exert a greater performance benefit on switch trials (Kleinsorge & Rinkenauer, 2012), where there is highest interference between task sets and thus where neural separation between task rules could be particularly important in determining behavioural performance. These analyses followed the procedure outlined in the previous section but only used switch or repeat trials, instead of all trials.

***Neural coding and cognitive performance*.** To test relationships between neural coding and cognitive performance, we averaged regression coefficients over time windows of interest and correlated them with behavioural scores using non-parametric Spearman correlations. For task coding, we correlated the difference between regression coefficients averaged over the pre-target period (1400-1800) and the difference in switch costs between reward conditions. We undertook the same procedure for reward coding averaged from 0-5000ms from the reward cue. To control for the influence of reward coding when correlating reward-modulated task coding with the changes in performance, we regressed out variance that could be explained by reward coding (averaged from 0-5000ms) from the reward-task coding effect (high-low reward) and the reward-behavioural effect (e.g. RT low reward-RT high reward). We then correlated the residual variance from the reward-task coding effect (averaged over pre-target period: 1400-1800ms or post-target period: 1800-

2000ms), with the residual variance in the reward-behavioural effect. We also tested whether the effect of reward on task coding as function of rule updating, was correlated with the difference in switch cost between reward conditions. For this analysis, the interaction term from Figure 6C was averaged over the pre-target period (1400-1800ms).

In addition to time averaged-correlations, we performed time resolved brain-behaviour correlations. To do this, we took participants' regression coefficients for different task variables at each time point along the trial and correlated this value with their difference in RT (low-high reward) and accuracy (high-low reward). Specifically, these analyses correlated a vector of 30 regression coefficient scores at a single time point with a vector of 30 behavioural scores. The same correlation approach was then performed at the next time point until a correlation had been performed at each time point along the window of interest. The false positive rate for this time-resolved correlation approach was controlled using a cluster-based permutation procedure (detailed in *statistical testing for neural analysis*), in which subjects' behavioural scores were randomly shuffled over many permutations. With the exception of reward coding (Figure 4C) which used the regression coefficient estimate directly, the neural data for these correlations were differences in regression coefficients between reward conditions. The behavioural data, such as the RT difference between reward conditions, were computed based on all trials used for the behavioural results section. In other words, these behavioural scores did not necessarily use identical trials used in EEG analyses, but instead made use of all behavioural trials available. The correlation analyses performed included neural data reflecting the difference in task coding (Figure 5B), relevant feature coding (Figure 8A), the interaction between reward and relevant feature prioritisation (Figure 8C), as well as motor coding locked to the response (Figure 9B). Time-resolved correlations involving reward-differences in neural regression

coefficients used the same test windows as their corresponding figure, from which differences were computed. All analyses applied non-parametric Spearman correlations to reduce the influence of outliers in behavioural difference measures, which were more than three scaled median absolute deviations away from the median (RT difference distribution: 2 outliers, lower threshold=-42ms, upper threshold=68ms, median=13ms; accuracy difference distribution: no outliers, lower threshold=-0.06, upper threshold=0.10, median=0.02).

Relationships between neural coding of task, sensory and motor information. To test whether reward-induced sensorimotor modulations arose from upstream changes in task coding, we performed a series of correlations between average task and sensorimotor coding regression coefficients. To do so, we selected time windows of interest based on results from the previous analysis section. For task coding, we averaged regression coefficients for each participant from 1400-1800ms. We did not include 1200-1400ms in this average due to a stimulus-evoked signal artefact shown in Figure 4B. For feature coding and motor coding, we averaged regression coefficients from 2000-2400ms, where we observed the peak difference between reward conditions (Figures 8A & 9A). For motor coding locked to the response, we averaged regression coefficients -200-0ms from the response, within which we found the peak difference in motor coding between reward conditions. As initial tests, we correlated average task coding with average feature coding and average motor coding, independent of reward. Significant correlations from this first step were followed up by correlating the mean difference in task coding (high-low reward) with the mean difference in the relevant sensorimotor variable (high-low reward). All correlations were non-parametric Spearman correlations. The Bonferroni-Holm correction was used to correct for multiple comparisons (Holm, 1979).

Statistical testing for neural analyses. Data were smoothed with a 12ms Gaussian kernel immediately prior to non-parametric cluster-based permutation testing, which was used to correct for multiple comparisons (Maris & Oostenveld, 2007; Sassenhagen & Draschkow, 2019; Spaak et al., 2017). This non-parametric approach is preferable to parametric statistical alternatives because it does not assume a particular distribution of the data. Cluster-based permutation testing computes a t-statistic at each point in the observed data and many times in permuted data. Each permutation involves shuffling labels for the data, creating a permutated data set that preserves temporal correlations within the EEG signal. Resulting t-statistics are thresholded at an alpha-level of 0.05. Considering only t-values that remain, the cluster of time points with the largest absolute sum of t-statistics (the largest cluster mass) from each permutation is placed into the null distribution. This null distribution shows the likelihood of obtaining a particular cluster mass due to chance. As a final step, candidate clusters from the observed data can then be compared against the null distribution. If the candidate cluster is larger than the 95^{th} percentile of the null distribution, then the effect is considered significant at an alpha level of 0.05. 10,000 permutations were performed to generate null distributions for each analysis in the present study.

Results

High reward decreased reaction time and improved accuracy

To assess the impact of reward prospect on behavioural performance, we performed a 2x2 repeated measures ANOVA on RT, with factors of reward (low x high) and task sequence (switch x repeat). This revealed a significant main effect of reward, $F(1,29)=18.676$, $p<0.001$, $\eta 2p=0.392$, which reflected lower RTs on high reward trials (M=415ms, SD=55.45) compared with low reward trials (M=438ms, SD=62.28). There was also a significant effect of task sequence, $F(1,29)=35.546$, $p<0.001$, $\eta 2p=0.551$, driven by lower RTs on repeat trials (M=412ms, SD=57.64) compared with switch trials (M=440ms, SD=59.36). While switch costs were numerically reduced on high compared with low reward trials, the interaction between reward and task sequence did not reach significance, $F(1,29)=2.662$, $p=0.114$, $\eta 2p=0.084$. A control analysis comparing switch trials with repeat trials involving different task cues showed significantly higher RTs on switch trials, confirming the main effect of task sequence was driven by genuine changes in task-set and not simple changes in visual cues (Shapiro-Wilk test for normality: W=0.822, $p=0.003$; Wilcoxon signed-rank test: $T=438.0$, p<0.001, matched rank biserial correlation=0.884).

The equivalent 2x2 ANOVA for proportion correct showed the same pattern of results. In particular, the analysis showed a significant main effect of reward, $F(1,29)=14.268$, $p<0.001$, $\eta 2p=0.330$, driven by greater accuracy on high reward trials (M=0.87, SD=0.06) compared with low reward trials (M=0.85, SD=0.07). It also showed a significant main effect of task sequence, $F(1,29) = 51.894$, $p<0.001$, $\eta 2p=0.642$, reflecting greater accuracy for repeat (M=0.89, SD=0.06) compared with switch trials (M=0.84, SD=0.07). We did not detect a significant interaction between reward and task sequence, $F(1,29)=0.317$, $p=0.578$, $\eta 2p=0.011$. Like RT, a control analysis comparing switch trials with

repeat trials involving different task cues showed significantly lower accuracy on switch trials, confirming the effect of task sequence on accuracy was driven by genuine changes in task-set (Shapiro-Wilk test for normality: W=0.977, p=0.744; t(29)=-6.110, p<0.001, d=-1.116). To summarise, our behavioural analyses show that high reward prospect improved both RT and accuracy performance.

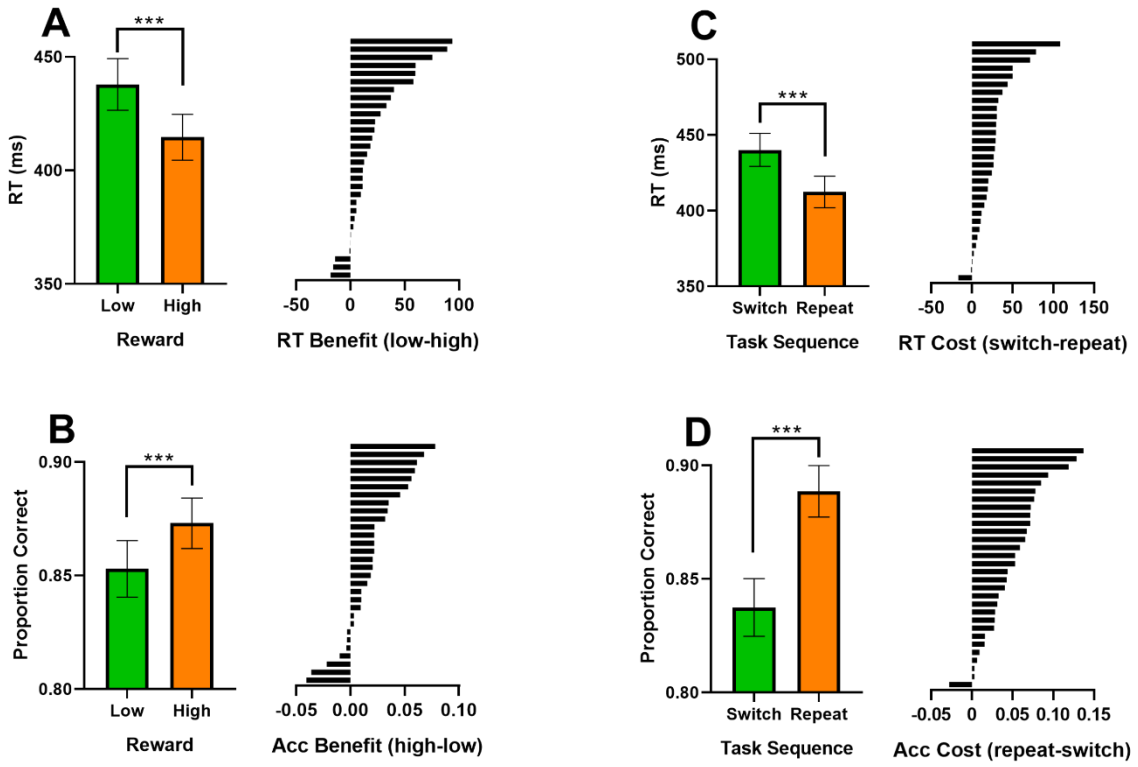


Figure 3. Behavioural performance as a function of reward and task sequence. A: Reaction time for high and low reward trials (left) and the difference between reward conditions for each participant (right). B: Proportion correct (accuracy) for high and low reward trials (left) and the difference between reward conditions per participant (right). C: Reaction time as a function of task sequence (left) and the difference between switch and repeat trials for each participant (right). D: Proportion correct (accuracy) as a function of task sequence (left) and the difference between switch and repeat trials (right) per participant. Error bars show standard error of mean.

Neural coding across the trial

Having established the beneficial impact of reward prospect on cognitive performance, we tested neural coding of task variables across the trial. Reward coding emerged shortly after reward cue onset and was sustained throughout the trial (Figure 4: window tested=0-3500ms from reward cue onset, cluster window=68-2940ms, p=0.0002). Task coding peaked shortly after the task rule cue was presented and continued into the response phase (window tested=1200-3500ms from reward cue onset, first cluster=1260-2220ms, first cluster p=0.0002, second cluster=2420-2948ms, second cluster p=0.0094). Finally, coding of task-relevant and irrelevant target features, as well as motor response coding rose shortly after target presentation (relevant feature coding: window tested = 1800-4500ms, cluster window=1900-3016ms, p=0.0002; irrelevant feature coding: window tested= 1800-4500ms from reward cue onset, cluster window=1892-2284ms, cluster p=0.0036; motor coding: window tested=1800-4500ms, cluster window=2004-3952ms, cluster p=0.0002). In summary, we verified our multivariate analysis approach was sensitive to dynamic temporal changes in neural coding of task variables at sensible stages within the trial.

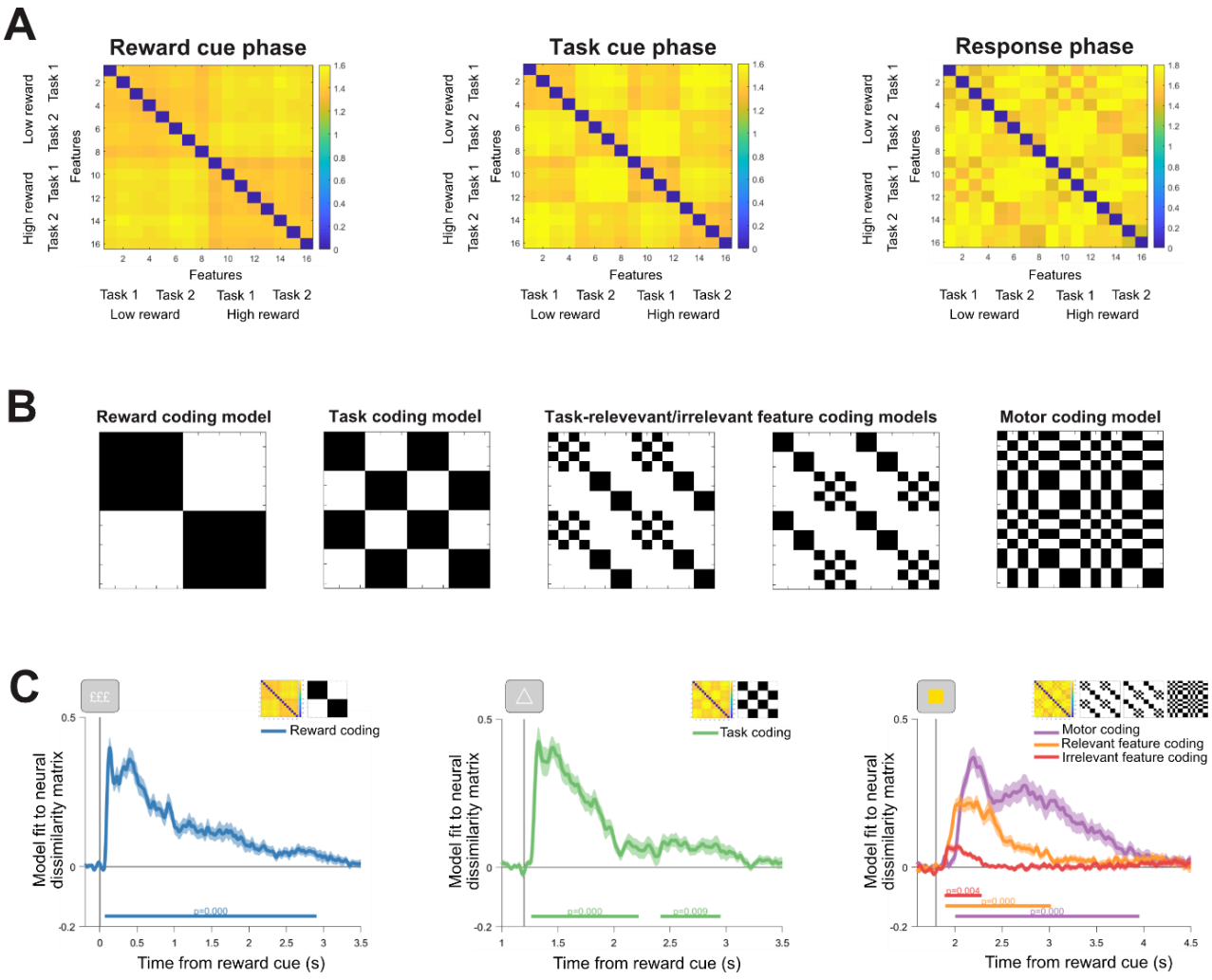


Figure 4. A: Representational dissimilarity matrices averaged across participants and trial periods following: reward cue onset (0-800ms), task cue onset (1200-1800ms) and target onset (1800-2600ms). B: Model representational dissimilarity matrices for: reward coding, task coding, coding of the task-relevant target feature, task-irrelevant target feature and motor coding. C: General linear model regression coefficients from regressing model and neural dissimilarity matrices across time. Shading around principle lines indicates standard error of mean. Cluster-corrected p-values are shown below each time course.

Task coding as a function of reward prospect

Reward prospect increased proactive task coding.

Having verified core task variables were encoded in the EEG signal at the plausible stages within the trial, we tested which variables were influenced by the reward manipulation. We found that average task coding was significantly greater on high reward trials prior to target onset (Figure 5A: window averaged=1400-1800ms; Shapiro-Wilk test for normality: W=0.973, *p*=0.633; theory driven one-tailed test based on Etzel et al. (2016): *t*(29)=2.881, *p*=0.004, *d*=0.526). Time-resolved permutation analyses confirmed robust encoding of task rules prior to the target under both reward conditions (Figure 5B: window tested=1200-2500ms, first low reward cluster=1280-2008ms, low reward *p*=0.0002; second low reward cluster=2020-2168ms; high reward cluster=1272-2032ms, high reward *p*=0.0002; difference cluster=1876-1968ms, *p*=0.0460).

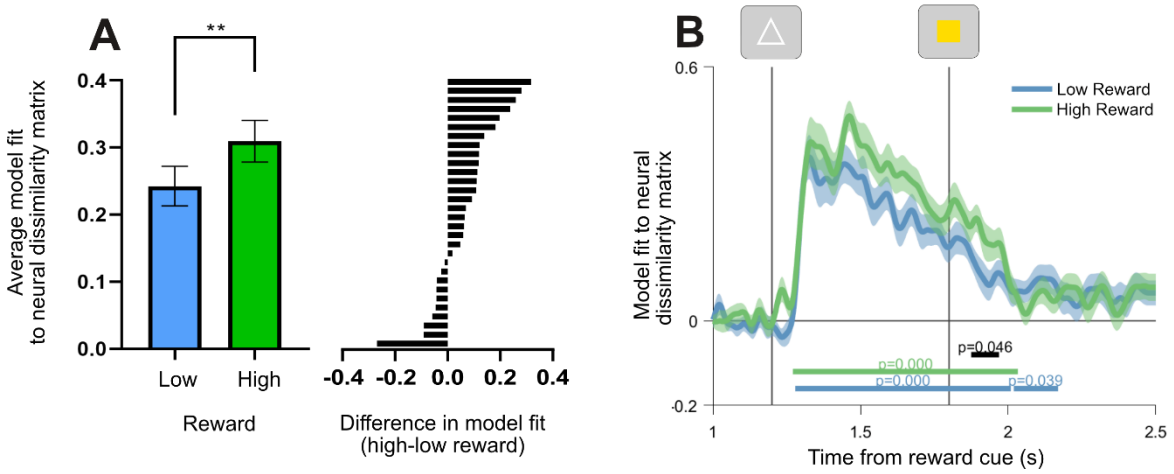


Figure 5. Generalised linear model regression coefficients for task coding as a function of reward. A: coefficients averaged over time points in the pre-target interval (1400-1800ms). The left sub-panel shows group means for average regression betas as a function of reward. The right sub-panel shows the difference between high and low reward regression coefficients for each participant. B: Time resolved regression coefficients for task coding as a function of reward. Vertical lines show onset of the task cue and target respectively. Shading around principle lines indicates standard error of mean. Cluster-corrected p-values are shown below time courses.

Task coding as a function of rule updating

Reward-induced increases in task coding were higher on switch trials.

Having shown reward prospect modulated task coding overall, we tested whether the effect of reward was amplified during rule updating, where interference between competing rules is highest and thus increased separation between rules could benefit flexible behaviour. This analysis revealed a significant increase in task coding on high reward switch trials compared with low reward switch trials (Figure 6A: window tested=1200-2500, difference cluster=1456-2024ms, p=0.0002). By contrast, we did not detect a difference in task coding as a function of reward on repeat trials (Figure 6B: window tested=1200-2500ms, longest candidate cluster=1224-1244ms, p=0.6825). The interaction between reward and task coding, as a function of rule updating, was significant when averaged over the pre-target interval (1400-1800ms, $t(29)$=2.9247, p=0.0066) and showed a trend towards significance when fully time resolved (Figure 6C: window tested=1200-2500ms, longest candidate cluster=1464-1544ms, p=0.0824).

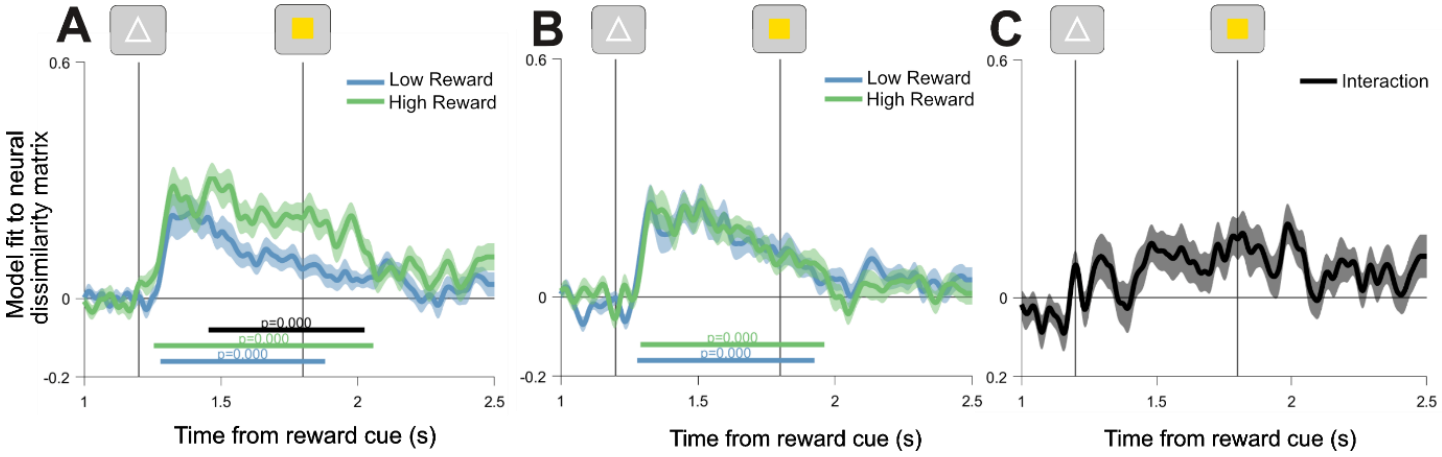


Figure 6. Time resolved regression coefficients for task coding as a function of reward on switch trials (A) and repeat trials (B). C: Interaction between reward and task sequence regression coefficients. This time course is computed by taking the difference between task coding regression coefficients on switch and repeat trials, within each reward condition. Resulting time courses are subtracted to show the extent to which high reward modulates the difference in task coding between switch and repeat trials. Vertical lines show onset of the task cue and target respectively. Shading around principle lines indicates standard error of mean. Cluster-corrected p-values are shown below time courses.

Task coding and cognitive performance

Neural encoding of task rules was associated with performance improvements.

To understand the strong performance benefits observed in behaviour, we tested whether task coding was associated with performance improvements (Etzel et al., 2016). The difference in task coding between reward conditions showed a significant correlation with reward-induced changes in RT (Figure 7A: window tested=1200-2500ms, RT difference: cluster window=1860-1968ms, mean Rho=0.4927, *p*=0.0225), but not with accuracy (longest candidate cluster=2280-2320ms, mean Rho=-0.4083, *p*=0.5382). Control analyses confirmed the relationship between reward-induced changes in task coding and RT reflected the difference in RT between reward conditions, rather than either reward condition individually (window tested=1200-1800ms, high reward RT only: longest candidate

cluster=1928-1948ms, mean Rho=-0.4004, *p*=0.3030; low reward RT only: longest candidate cluster=1636-1664ms, mean Rho=0.4602, *p*=0.3502). We did not detect significant associations between the reward-rule updating interaction (Figure 6C) and the change in switch costs between reward conditions for RT or accuracy (window averaged=1400-1800, RT switch cost difference: Spearman's Rho=-0.2725, *p*=0.1449; accuracy switch cost difference: Spearman's Rho=-0.3351, *p*=0.0703).

Reward coding itself also showed a strong relationship with changes in performance. The magnitude of reward coding was significantly correlated with participants' difference in RT on high compared with low reward trials (Figure 7A: window tested=0-5000ms, first cluster window=392-1992ms, mean Rho=0.5049, *p*=0.0008; second cluster window=2000-3456ms, mean Rho=0.6001, *p*=0.0004, third cluster window=3496-3892ms, mean Rho=0.5278, *p*=0.0078; fourth cluster window=4072-4524ms, mean Rho=0.5483, *p*=0.0100; fifth cluster=4588-4748ms, mean Rho=0.4487, *p*=0.0415; sixth cluster=4848-4996ms, mean Rho=0.4112, *p*=0.0459). Control analyses correlating reward coding with RT separately on high or low reward trials confirmed this reward coding association was specific to the difference in performance between performance (window tested=0-5000ms, high reward RT only: longest cluster=1188-1248, mean Rho=-0.4113, *p*=0.3258; low reward RT only: longest cluster=4484-4512ms, mean Rho=0.3939, *p*=0.3656). We did not detect a significant correlation between reward coding and the difference in accuracy between reward conditions (window tested=0-5000ms, longest candidate cluster=4-8ms, mean Rho=-0.3822, *p*=0.7373). Similarly, we did not detect significant relationships between reward coding and the difference in switch cost between reward conditions for RT (window averaged=0-5000ms, Spearman's Rho=0.2912, *p*=0.1185) or accuracy (window averaged=0-5000ms, Spearman's Rho=0.0472, *p*=0.8045).

To test whether the relationship between reward-driven changes in task coding and improved RT performance was driven by reward coding itself, we ran a control analysis that regressed out variance associated with reward coding (averaged from 0-5000ms) from the reward-task coding effect (high-low) and the reward-RT effect (high-low). We then tested the association between the residual variance in the reward-task coding effect and the residual variance in the reward-RT effect, using non-parametric Spearman correlations. This control indicated that reward-modulated task coding during the pre-target period (averaged from 1400-1800ms) was associated with the change in RT performance, even after removing variance that could be accounted for by reward coding (Figure 7C: Spearman's Rho=0.4554, *p*=0.0122). This was also the case for reward-modulated task coding during the 1800-2000ms post-target period (Figure 7D: Spearman's Rho=0.3971, *p*=0.0306). As a final control, we applied this procedure to cluster windows taken from Figures 6A-B (task difference cluster range=1860-1968ms, reward cluster range=392-4996ms). This confirmed the significant time resolved correlation cluster identified for reward-modulated task coding (Figure 6A) was associated with reward-driven changes in RT, independent of reward coding itself (Spearman's Rho=0.5075, *p*=0.00469).

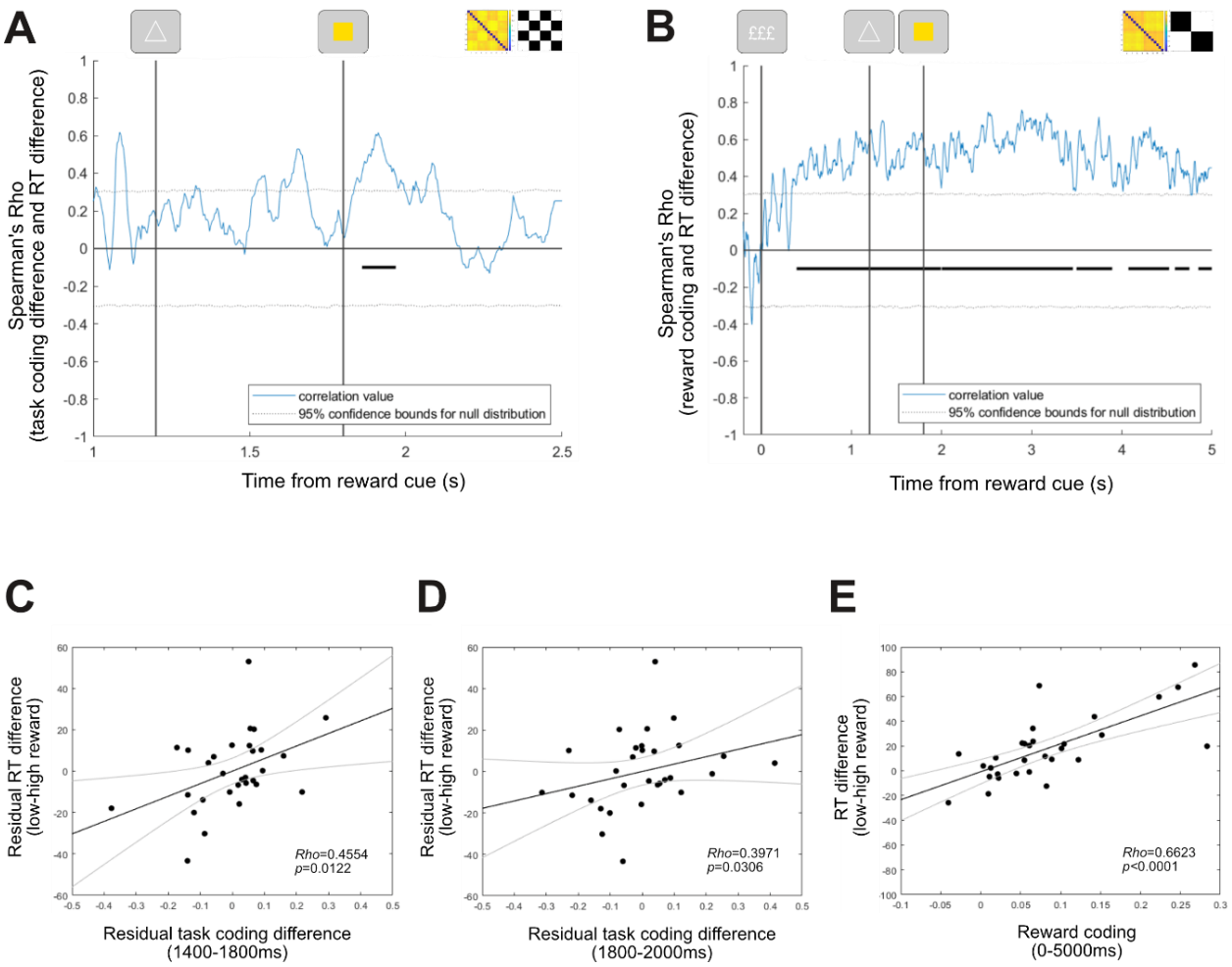


Figure 7. Relationships between reward-modulated task coding and performance. A: Spearman Rho values for correlation between the difference in task coding regression coefficients (high-low reward) and their difference in reaction time between reward conditions. B: Spearman Rho values for correlation between participant reward coding regression coefficients and their difference in reaction time between reward conditions. In both panels A-B, black lines indicate significant correlation clusters, corrected for multiple comparisons using cluster-based permutation testing. Grey dotted lines indicate 95% confidence intervals for the null distribution. C: Scatterplot showing the relationship between the RT difference (low-high reward) and pre-target task coding (high-low reward, 1400-1800ms), after removing variance associated with reward coding. D: Scatterplot showing the relationship between the RT difference (low-high reward) and post-target task coding (high-low reward, 1800-2000ms), after removing variance associated with reward coding. E: Scatterplot showing the relationship between the RT difference (low-high reward) and reward coding (0-5000ms). In panels C-E, black lines indicate linear fits to the data and grey lines indicate 95% confidence intervals of the fits.

Feature coding as a function of reward prospect

Reward prospect increased coding of task-relevant target features.

After evaluating the impact of reward prospect on task coding, we examined the effect of reward prospect on neural coding of task-relevant and irrelevant target features. Relevant feature coding was observed shortly after target presentation on both high and low reward trials (Figure 8A: window tested=1800-3000ms, low reward cluster=1920-2812ms, p=0.0002; first high reward cluster=1916-2892ms, p=0.0002). In addition, neural coding for task-relevant features was significantly higher under high reward conditions (window tested=1800-3000ms, first difference cluster=2036-2172, p=0.0100; second difference cluster=2252-2404ms, difference p=0.0052). A control analysis which regressed the task-relevant feature model against the residual variance, not explained by any of the other coding models, confirmed this result was not driven by correlated regressors (window tested=1800-3000; low reward cluster=1920-2812ms, p=0.0002; high reward cluster=1916-2892ms, p=0.0002; first difference cluster=2036-2172ms, p=0.0120; second difference cluster=2252-2404, p=0.0058). Task-irrelevant information was also represented following target onset for both reward levels (Figure 8B: window tested=1800-3000ms, low reward cluster=2008-2284ms, p=0.0002; second low reward cluster=2460-2560ms, p=0.0470; high reward cluster=1916-2208ms, p=0.0008). However, the strength of these coding patterns did not differ as a function of reward (window tested=1800-3000ms, no candidate clusters). As a consequence of these target-related effects, reward prospect showed a significant interaction with the difference in task-relevant and irrelevant feature coding. This reflected a greater difference between task-relevant and irrelevant coding under high reward conditions (Figure 8C: window tested=1800-3000ms, interaction cluster=2240-2364ms, p=0.0258). In summary, we found evidence that high reward prospect increased the

difference in neural coding for task-relevant and irrelevant target information. This difference was due to a selective increase in the neural coding of task-relevant feature information in high reward contexts.

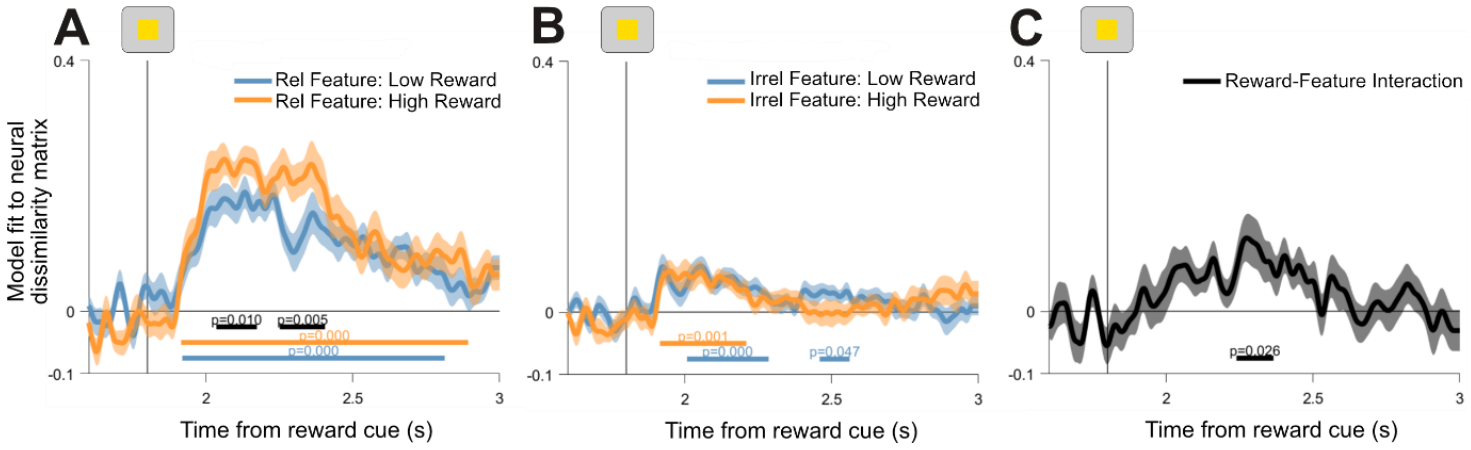


Figure 8. General linear model regression coefficients for coding of target features as a function of reward. A: Task-relevant target feature regression coefficients. B: Task-irrelevant feature regression coefficients. C: Interaction between reward and feature regression coefficients. This time course is computed by taking the difference between task-relevant and task-irrelevant regression coefficients within each reward condition. Resulting time courses are subtracted to show the extent to which high reward increases the difference between task-relevant and irrelevant target features. Vertical lines show target onset. Shading around principle lines indicates standard error of mean. Cluster-corrected p-values are shown below time courses.

Feature coding as a function of rule updating

Reward-feature coding modulations did not differ on switch trials.

While high reward prospect increased neural coding of task-relevant feature information overall (Figure 8), the difference in relevant feature coding between reward conditions was not significant when switch and repeat trials were analysed separately (window tested=1800-3000ms, switch trials: longest candidate cluster=2144-2172ms, p=0.2274; repeat trials: longest candidate cluster=2104-2148ms, p=0.1902).

Feature coding and cognitive performance

Reward-feature coding modulations were not linked to performance improvements.

In contrast to the relationships between task coding and performance improvements, we did not detect significant associations between reward-induced changes in feature coding and reward-induced changes in cognitive performance. This was the case for relevant feature coding (window tested=1800ms-3000ms, correlation with RT: longest cluster=1880-1884ms, mean Rho=-0.3809, p=0.8172; correlation with accuracy: longest cluster=2660-2676ms, mean Rho=-0.3842, p=0.6628), as well as the interaction between reward and relevant feature prioritisation (Figure 8C) (window tested=1800-3000ms, correlation with RT: longest cluster=1868-1880ms, mean Rho=-0.4230, p=0.6568; correlation with accuracy: longest cluster=2752-2764ms, mean Rho=0.3789, p=0.7073).

Motor coding as a function of reward prospect

Reward prospect increased neural encoding of task-relevant motor output.

Having established reward prospect modulated neural activity coding for task-relevant target features, we examined the effect of reward prospect on activity patterns related to the upcoming motor response. When the analysis was locked to the onset of the reward cue, motor coding appeared after target presentation during both high and low reward conditions (Figure 9A: window tested=1800-4500ms, low reward cluster=1996-3716ms, low reward p=0.0002; high reward cluster=2016-3792ms, high reward p=0.0002) and showed a trend towards higher motor coding under high reward conditions (Figure 9A: window tested=1800-4500ms, difference cluster=2040-2152ms, p=0.0900). When the analysis was locked to the onset of the motor response itself, we observed significantly

higher motor coding on high reward trials (Figure 9B: window tested =-500-2000ms from response, low reward cluster=-220-1636ms, low reward *p*=0.0002; high reward cluster=-180-1784ms, high reward *p*=0.0002; difference cluster=-36-164ms, *p*=0.0158).

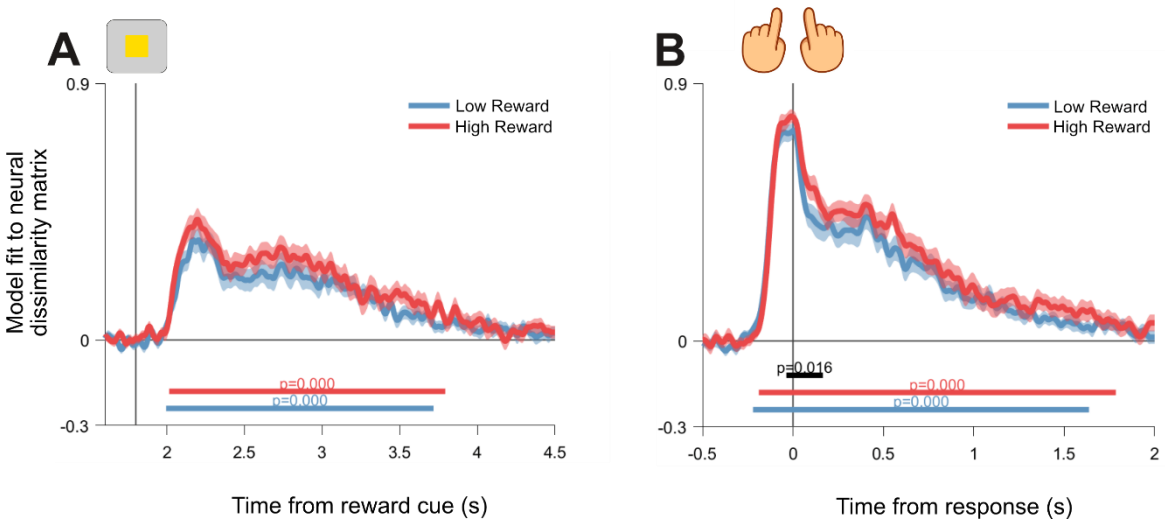


Figure 9. General linear model regression coefficients for coding of the upcoming motor response as a function of reward. A: Motor model regression coefficients for data locked to reward cue onset. The vertical line indicates target onset. B: Motor model regression coefficients for data locked to the response. The vertical line indicates the point responses were made. Shading around principle lines indicates standard error of mean. Cluster-corrected p-values are shown below time courses.

Motor coding as a function of rule updating

Reward prospect modulated motor coding on switch and repeat trials.

Based on the reward-motor coding effect observed in our response locked analysis (Figure 9B), we tested whether this effect differed on switch and repeat trials. Motor coding was significantly greater on high compared with low reward switch trials (window tested=-500-2000ms from response, longest difference cluster=-72-180ms, *p*=0.0040), as well as high compared with low reward repeat trials (window tested=-500-2000ms from response,

first difference cluster=-84-120ms, *p*=0.0200; second difference cluster=1512-1748ms, *p*=0.0240).

Motor coding and cognitive performance

Reward-motor coding modulations were not linked to performance improvements.

When testing whether the reward boost in motor coding (Figure 9B) was associated with changes in performance, we did not detect a significant association between the increase in motor coding (high-low reward) and the change in RT between reward conditions (window tested=-500-2000ms, longest difference cluster=-192 to -152ms, mean Rho=-0.4415, *p*=0.3505)**.** Similarly, we did not detect a significant association between the reward boost in motor coding and the change in accuracy between reward conditions (window tested=-500-2000ms, longest difference cluster=468-524ms, mean Rho=-0.4618, *p*=0.1985).

Relationships between neural coding of task rule, feature and motor information

Reward-induced changes in task coding were not associated with sensorimotor effects.

Having established that reward prospect modulated proactive coding of task information, we investigated whether such proactive changes could account for modulations in post-target processing. To do so, we first tested whether average task coding during the pre-target interval (1400-1800ms) was associated with average feature coding (2000-2400ms), as well as average motor coding locked to the reward cue (2000-2400ms) and the response (-200-0ms). Task coding was significantly correlated with coding of the relevant target feature (Spearman's Rho=0.5355, Bonferroni-Holm corrected *p*=0.0156,) and

the difference between task-relevant and irrelevant features (Spearman's Rho=0.4719, Bonferroni-Holm corrected p=0.0455), indicating that participants with greater task coding also tended to exhibit greater prioritisation of task-relevant target features. We did not detect significant correlations between task coding and motor coding locked to the reward cue (Spearman's Rho=-0.1511, Bonferroni-Holm corrected p>0.99) or the response (Spearman's Rho=-0.1448, Bonferroni-Holm corrected p>0.99). Having established these relationships, we tested whether reward modulations in task coding during the pre-target period (high-low reward, 1400-1800ms) were correlated with reward modulations in relevant feature coding following target onset (high-low reward, 2000-2400ms). This analysis did not detect a significant correlation between reward-modulated task and reward-modulated feature coding (Spearman's Rho=0.1097, Bonferroni-Holm corrected p=0.8870). The same pattern of results was found when testing the association between reward-modulated task coding (high-low reward, 1400-1800ms) and the extent to which high reward increased the representation of the relevant target feature over the irrelevant feature (2000-2400ms, Spearman's Rho=0.2957, Bonferroni-Holm corrected p=0.4512).

Discussion

The present study aimed to investigate how rewards modulate neural coding of task rule information to support flexible cognitive control. We employed RSA to examine changes in neural coding of tasks as the prospect of reward changed dynamically, from trial to trial. Using this method, we were able to track neural representations for reward prospect, task rules, task-relevant and irrelevant perceptual features of target stimuli and neural representations related to accurate motor output. We found that high reward prospect boosted the encoding of multiple task variables. Critically, reward increased encoding of the active task rule in preparation for the target and the extent of this increase was associated with reward-based reductions in RT. In addition, reward's modulatory effect on rule coding was amplified on switch trials, where more flexible processing of rules was needed. Following the target, we observed increased encoding of task-relevant perceptual and motor information under high reward conditions.

Consistent with the results of previous fMRI decoding studies (Etzel et al., 2016; Qiao et al., 2017; Waskom et al., 2014; Wisniewski et al., 2015; Woolgar et al., 2011), RSA successfully tracked neural representations for task rule information. In using this approach, our results replicate the fMRI findings of Etzel et al. (2016) in human electroencephalographic data, showing that high reward prospect increased average task coding prior to target onset, and that the difference in task coding between reward conditions was associated with improvements in cognitive performance. The effect of reward on task coding was most pronounced on switch trials, where task rules needed to be updated relative to the previous trial. This could suggest that dynamic increases in reward prospect primarily promote flexible rule updating, as opposed to maintenance of existing task rule representations in prefrontal regions. Such updating might be mediated by phasic

dopamine release in the striatum, which is thought to be important for driving flexible and targeted updating of contextual information in PFC (Westbrook & Braver, 2016; Yee & Braver, 2018). In light of our neural results, showing a larger reward effect on switch trials, it is surprising that there was not a significant corresponding effect in behaviour. We believe this behavioural result should be interpreted with caution, as there was a clear trend in the expected direction, and previous studies have observed robust reward effects on rule switching (Kleinsorge & Rinkenauer, 2012; Shen & Chun, 2011). Overall, the present results provide evidence that high reward prospect can improve cognitive performance by increasing proactive neural representations for task rule information. Moreover, these results seem to be consistent with the theoretical view that reward prospect could help separate the representations of competing task rules, which is especially critical during transitions between different behavioural contexts (Waskom et al., 2014). Indeed such a mechanism could account for the selective effects of reward on switching observed in previous studies (Kleinsorge & Rinkenauer, 2012; Shen & Chun, 2011).

One interesting aspect of these results was that reward-driven changes in task-coding were primarily associated with improvements in RT but not accuracy. This result is distinct from Etzel et al. (2016), who found reward-driven increases in task-coding mediated improvements in accuracy but not RT. A simple explanation for this could be our RT incentive manipulation, which made RT a more informative behavioural measure in the current study. On top of this, RT values have a wider dynamic range than accuracy, which could have made it easier to detect associations between neural measures and RT performance. From a theoretical perspective, the asymmetry between our RT and accuracy correlation results might be viewed as an optimal-control problem, in which subjects maximise expected value over time through balancing speed and accuracy (Bogacz, 2007;

Manohar et al., 2015). In our task, accuracy rates were relatively high and while reward significantly increased accuracy, the effect was much smaller than for RT. One possible explanation for this comes from Manohar et al. (2015), who propose an optimal control model to account for the effects of reward on neural noise and behavioural performance. One important prediction of this model is that when signal-to-noise in a neural system is high, rewards will have a disproportionate effect on response vigour and the reduction of RT. These signal-to-noise conditions might well have reflected our task conditions, in which accuracy rates were high and stimuli were unambiguous.

Consistent with the results from previous studies focusing on neural responses to perceptual targets (Hickey & Peelen, 2015; Padmala & Pessoa, 2011; Serences, 2008; Serences & Saproo, 2010), we found that reward enhanced the representation of task-relevant perceptual features. The present results complement this literature in two ways. First, many previous studies have focused on perceptual stimuli associated with reward over many trials (Hickey & Peelen, 2015; Serences, 2008; Serences & Saproo, 2010). Here we show that transient changes in prospective reward can also modulate task-relevant perceptual feature representations. Second, research using prospective reward cues has led to the proposal that reward motivation might benefit attentional filtering, either by enhancing task-relevant perceptual representations or supressing task-irrelevant representations (Pessoa, 2017). Hickey and Peelen (2015) found evidence for both of these mechanisms, showing that reward could either enhance or suppress perceptual representations as a function of task-relevance. In the present study, we report a more selective effect on perceptual representations, wherein high reward prospect increased the neural representation of task-relevant information without impacting the representation of irrelevant information. This might suggest that transient reward coupling with sensory

target features has a different impact on perceptual encoding than reward-associations established over many trials.

While task coding and the prioritisation of task-relevant target features were strongly correlated, we did not find evidence that reward-driven modulations in these variables were associated. We are cautious not to over-interpret these null-effects. These results do not rule out the possibility that reward-modulated task coding affects downstream perceptual representations. However, they do raise the possibility of an alternative mechanism, wherein reward prospect acts independently on multiple neural variables. Among these variables, our results suggest that neural encoding of reward prospect and task rule information are important factors associated with dynamic shifts in improving performance. The present results to do not permit conclusive interpretations about the functional role of reward-driven perceptual and motor changes; although we did not detect significant associations between perceptual representations and behavioural measures, this does not imply that these variables were functionally irrelevant to task performance.

How might reward motivation translate into performance improvements more broadly? Previous studies have pointed to the idea that reward motivation might upregulate attention (Etzel et al., 2016; Padmala & Pessoa, 2011; Pessoa & Engelmann, 2010). For instance, reward and attention have been shown to recruit overlapping frontoparietal control regions (Pessoa & Engelmann, 2010) and have analogous effects on electrophysiological signatures of task preparation (van den Berg et al., 2014). One possibility in our study is that reward prospect had additional effects that were not captured by the conditions of our task. For instance, high reward prospect could have increased alertness and temporal attention to information proximal to reward cue presentation. This

may explain why associations between behavioural measures and downstream reward effects, such as target feature and motor representations, could then be noisier and less reliable than the strong correlation between behaviour and reward itself.

To conclude, previous work has shown that reward motivation may improve cognitive performance by boosting the neural coding of task rules (Etzel et al., 2016). Here we demonstrate that high reward prospect can increase proactive coding of task rule information and that this proactive effect has a strong relationship with performance improvements. In addition, reward's effect on rule coding was heightened during switch trials, where task rules were updated relative to the previous trial. This suggests reward prospect might optimise flexible control processes by increasing the neural separation between task rules.

Experiment 3.2

Introduction

Experiment 3.1 provided evidence that rewards could boost the neural encoding of task rules and that this effect was stronger on trials with higher interference between the previous and current task rule (i.e. on switch trials). In experiment 3.2, we tested whether rewards influenced the persistence of task rules even in situations where this was counterproductive. Here we stepped away from manipulating reward prospect and focused on how reward feedback could influence the maintenance and updating of task rules. We reasoned that, if maintaining the current rule was sensitive to reinforcement, high reward feedback should boost maintenance of the current rule. In turn, this would be expected to create more interference on subsequent trials with a rule switch, thereby impairing switch trial performance. This hypothesis is supported by work showing that high reward feedback on the previous trial increases switch costs for both accuracy and RT (Aarts et al., 2010). Studies on visual search also support this perspective, providing evidence that high reward on the previous trial can impair performance by biasing processing towards information that was previously relevant. For example, high reward feedback slows response times on subsequent visual search trials, when colour features associated with targets and distractors are reversed (Hickey et al., 2010). The present study therefore aimed to test the prediction that high reward feedback on the previous trial would impair switch trial performance.

Method

Participants

Following behavioural experiments from the previous chapter, we aimed for a sample size of 24 participants. We therefore collected 25 participants across two group testing sessions. Two participants were excluded, due to accuracy less than 70%. The participants in the final sample were between 18 and 21 years of age (mean age = 19, 19 female), with normal or corrected-to-normal vision, who reported no history of neurological or psychiatric illness. Participants received £10 per hour or course credit for taking part and could earn up to £10 extra for their performance. This study was approved by the Central University Research Ethics Committee at the University of Oxford and all participants signed informed consent before taking part.

Materials

Stimuli were presented on a 23-inch screen with a spatial resolution of 1920 x 1080 and refresh rate of 60Hz. All other materials were identical to experiment 3.1.

Experimental Design and Statistical Analysis

The task design for experiment 3.2 highly similar to 3.1. The main differences were that reward prospect cues from the beginning of the trial were replaced with a fixation cross. In addition, we introduced a delay between the motor response and feedback. This was implemented with a follow up EEG experiment in mind, to allow motor coding to dissipate so that rule codes could be more easily resolved following feedback. Finally, reward feedback for correct responses were randomly determined as being high (+100 points) or low (+10 points). Note that this meant the points awarded did not depend on RT.

The main experimental sequence (Figure 10) consisted of a fixation cross, task cue, target, feedback screen and an inter-trial interval (ITI). The fixation cross was first presented

for 200ms, followed by a 400ms delay. The task cue was then presented for 200ms. In keeping with the previous experiment, the mapping of cues to tasks was counterbalanced between participants. Task cue offset was followed by a 400ms delay. The target was then presented and remained on screen until a response was given or for a maximum of 1400ms. Unlike the previous experiment, response mappings for yellow/square and blue/circle responses were counterbalanced between participants. The response phase was followed by a 400ms blank delay and feedback lasting 200ms. An incorrect response or omission resulted in feedback showing '+0' points, whereas a correct response resulted in randomly determined high or low feedback, showing '+100' or '+10' points respectively. The trial concluded with a randomly selected ITI duration, drawn from a uniform distribution with values of 1000, 1100, 1200, 1300 or 1400ms. Participants were trained to reach a criterion of 70% accuracy before completing 10 experimental blocks of 65 trials. Excluding the first trial in each block, equal numbers of task cues, reward cues, target stimuli and ITI durations were presented within each block. Across the session, there were approximately equal numbers of switch and repeat trials where the previous trial had high or low reward. In addition, there were approximately equal trial numbers for each task sequence-previous reward condition, as a function of target-congruency on the previous trial.

Behavioural measures. Behavioural analysis was similar to experiment 3.1. Our main dependent measures were the proportion of correct responses (accuracy) and RT. For accuracy analyses, we included trials in which a response was made within the 1.4s response window and a correct response was given on the previous trial. For RT analyses, we only included trials in which the current and previous trials were correct to mitigate potential effects of post-error slowing (Dutilh et al., 2012; Notebaert et al., 2009) and computed participants' median RTs to account for the skewness of RT distributions (see

Ratcliff, 1993). Accuracy and RT data were analysed using 2x2 repeated measures ANOVA with factors of task sequence (switch x repeat) and reward on the previous trial (high x low).

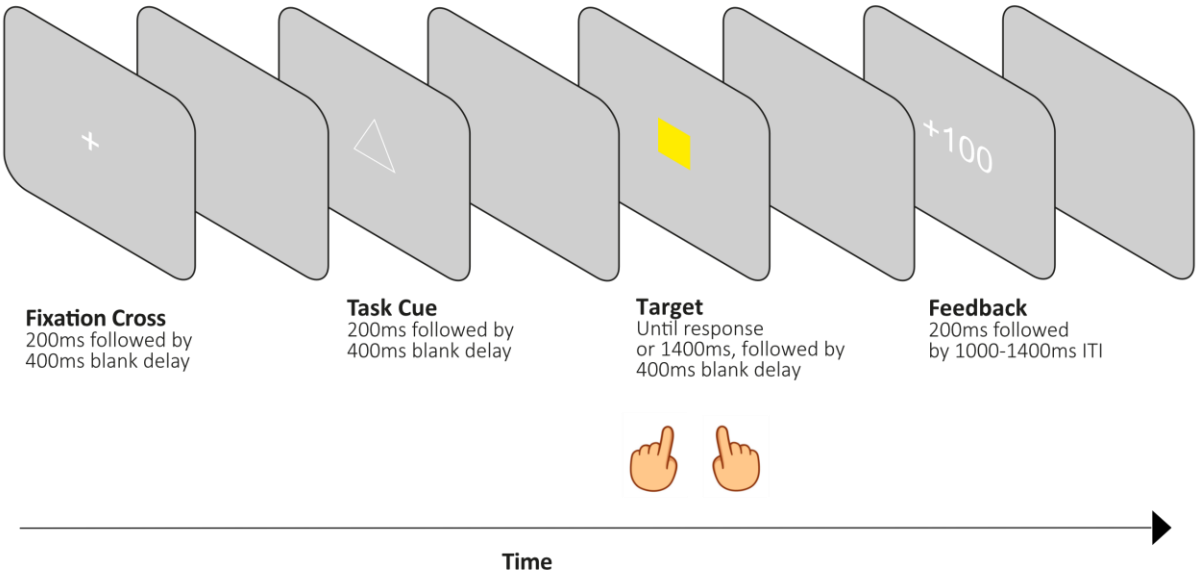


Figure 10. Task design. On each trial, a fixation cross was presented followed by a blank delay. A task rule cue was then presented, indicating whether participants should respond to the upcoming target based its colour or shape. Following a second blank delay, a bi-dimensional target (a coloured shape) appeared until a response was given or for a maximum duration of 1400ms. Following a third blank delay, random high or low reward feedback was given (for correct responses). The trial concluded with a variable inter-trial-interval.

Results

Reward on the previous trial did not modulate current trial performance

To assess the impact of previous trial reward on behavioural performance, we performed a 2x2 repeated measures ANOVA on accuracy, with factors of task sequence (switch x repeat) and reward feedback on the previous trial (low x high). This revealed a significant main effect of task, $F(1,22)=77.263$, $p<0.001$, $\eta 2p=0.778$, which reflected lower accuracy on switch trials (M=89.47%, SD=5.09) compared with repeat trials (M=95.28%, SD=2.87). This analysis did not detect a significant main effect of previous trial reward on current trial accuracy, $F(1,22)=1.368$, $p=0.255$, $\eta 2p=0.059$, or a significant interaction between task sequence and previous trial reward, $F(1,22)=0.109$, $p=0.744$, $\eta 2p=0.005$.

The equivalent 2x2 ANOVA for RT showed the same pattern of results. Specifically, the analysis showed a significant main effect of task sequence, $F(1,22)=32.714$, $p<0.001$, $\eta 2p=0.598$, which reflected slower RTs on switch trials (M=516ms, SD=109.00) compared with repeat trials (M=459ms, SD=82.86). The analysis did not detect a significant main effect of previous trial reward on current trial RT, $F(1,22)=0.322$, $p=0.576$, $\eta 2p=0.014$, or a significant interaction between task sequence and previous trial reward, $F(1,22)=0.251$, $p=0.622$, $\eta 2p=0.011$.

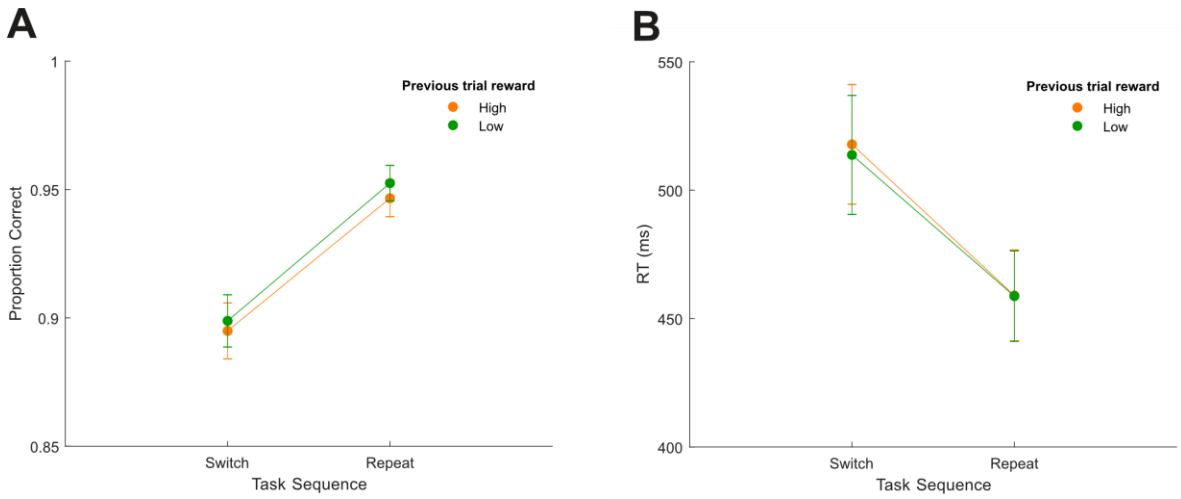


Figure 11. Behavioural performance. A: Proportion correct for task switches and repeats, as a function of reward on the previous trial. B: Reaction time for task switches and repeats, as a function of reward on the previous trial. A-B: Error bars show standard error of mean.

Discussion

This behavioural follow up aimed to test the prediction that high reward feedback would strengthen neural coding for the active task rule and thereby impair performance on subsequent switch trials. Contrary to this prediction, we did not detect an interaction between previous trial reward and task sequence for accuracy or RT. We also did not detect a main effect of previous reward on either dependent measure.

These results differ from Aarts et al. (2010), who showed that high reward on the previous trial resulted in higher switch costs for both accuracy and RT. In contrast to our design which used non-contingent rewards, Aarts et al. (2010) included reward prospect cues, making reward feedback performance contingent. This might suggest that contingent rewards are needed to reinforce current task associations. Previous studies lend support to the view that contingent rewards can have stronger effects on behaviour and neural activity than non-contingent rewards. For instance, people saccade more quickly to a target location when rewards are contingent on speed than when they are delivered at random (Manohar et al., 2017). In addition, contingent rewards are more strongly encoded in dorsomedial PFC, parietal cortex and the striatum compared with non-contingent rewards (Wisniewski et al., 2019). Nevertheless, the idea that contingent reward outcomes might reinforce current rules more strongly is complicated by results showing the absence of a reward contingency x task coding interaction in dorsomedial PFC (Wisniewski et al., 2019). We had deliberately removed prospective reward cues in our design, so that the effects of previous reward feedback on current task encoding would not be confounded by proactive reward effects in a prospective EEG study. Notwithstanding the mixed evidence above, it is conceivable that making rewards non-contingent in the current study could have dampened reward feedback effects. If this were true, we might expect an interaction between previous reward and

switching in the data from experiment 3.1. However, this was not the case for RT, $F(1,29)=2.130$, $p=0.155$, $\eta 2p=0.068$, or accuracy, $F(1,29)=0.740$, $p=0.397$, $\eta 2p=0.025$. Despite an absence of reward effects in the current study, our design (E3.2) was more appropriate to address our research question than a contingent-reward design, which used reward prospect cues. As investigated in the previous chapter, moving from low to high reward prospects might result in significantly reduced switch costs (Shen & Chun, 2011). This could bias an analysis of previous reward towards better performance after low reward trials, simply because these conditions would include low-high reward transitions. To circumvent this, a future design could use contingent rewards that are not known until the end of the trial. For example, reward feedback could be randomly determined as a high or low reward range, but the amount within the range could scale with faster RTs.

In addition to contingent rewards, the task in Aarts et al. (2010) presented reward feedback immediately after responses, whereas our task introduced a 400ms delay between responses and feedback. This might suggest that the potential effect of reward feedback on current task representations is sensitive to the temporal delay between task-guided actions and outcomes. We introduced a delay to help separate feedback processing from motor responses in a prospective EEG study. Nevertheless, this temporal separation could have prevented reward feedback from reinforcing current task representations. While we are not aware of previous work systematically investigating feedback delay on task representations to speak to this directly, previous work on rule and stimulus-response learning have failed to observe an effect of feedback delay (Foerde & Shohamy, 2011). However, evidence suggests that rule learning and rule use preferentially recruit different regions of lateral PFC, with rule learning primarily recruiting ventrolateral PFC and flexible rule use primarily recruiting dorsal PFC (Bunge, 2004). The involvement of these different PFC regions in learning and use

makes it difficult to draw direct parallels between the learning experiment above and the present experiment. It is therefore still conceivable that the temporal delay between task-guided actions and outcomes mitigated the potential effect of reward feedback on task rule coding. In a similar manner, it is possible that the ITI duration was too long for feedback-rule interactions to extend into the subsequent trial. However, we consider this less likely, as Aarts et al. (2010) observed these effects with even longer ITIs. Future designs could therefore focus on removing the action-response delay from the current design. Overall, we interpret the absence of a feedback-task sequence interaction to indicate our task was not sensitive to the potential relationship between these variables.

To conclude, the results of experiment 3.2 failed to replicate the finding that high reward feedback on the previous trial would impair switch trial performance (Aarts et al., 2010). We interpret this result as showing our design was not sensitive to feedback-task sequence interactions.

Chapter 3 Appendix

A1. Training protocol for experiments 3.1 and 3.2

The training protocol was similar to experiments reported in the previous chapter. Participants began the session by learning response keys associated with colour and shape judgements. For experiment 3.1, the left hand response key 'F' was required when the relevant target feature was yellow or square. By contrast, the right hand response key 'J' was required when the relevant target feature was blue or circle. For experiment 3.2, response buttons for yellow/square and blue/circle were counterbalanced between participants. In both experiments, this response structure meant participants could encounter target stimuli where the relevant and irrelevant features had congruent response mappings (yellow square and blue circle) or incongruent response mappings (yellow circle and blue square). Participants were instructed to always use their left and right index fingers for 'F' and 'J' responses respectively. Participants were first presented with randomly selected target stimuli (coloured shapes) and were asked to respond based on their colours. Target stimuli remained on screen until a response was given, after which participants received feedback about the accuracy of their response for a duration of 1000ms. Feedback text showing '+1' was used to signal a correct response and '+0' was used to signal an incorrect response. Participants needed to make 10 correct responses to progress to the next training stage in experiment 3.1 and 12 correct responses in experiment 3.2. Following colour judgements, participants completed the same key learning exercise for shape judgements.

In the next training phase, participants were presented with visual task cues for 200ms, with an onset of 600ms before each target. Task cues were abstract shapes (shown in Appendix Figure A2), which did not resemble target stimuli. Two task cues signalled that

participants would need to judge the upcoming target based on its colour to gain points on the trial, whereas two different cues signalled that participants would need to judge the shape of the upcoming target. On each trial, only one task cue was presented ahead of the target. The mapping of cues to tasks was counterbalanced between participants. In this training section, participants first completed trials in which the colour task cue was presented, followed by a randomly selected target. Following 10 correct responses in experiment 3.1, or 12 correct responses in experiment 3.2, this process was repeated with shape task cues.

The final practice stage built on the last training step by adding reward information to each trial. In experiment 3.1, reward prospect cues were added at the start of each trial. Three pound signs (£££) indicated participants could earn between 50-100 points for a correct response, whereas a single pound sign (£) indicated participants could earn between 5-10 points. The amount participants earned within these ranges depended on how quickly they responded to the target once it was presented. The RT incentive structure, based on Shen and Chun (2011), is described in the main text. However, it should be noted that Shen and Chun's (2011) approach evaluated RTs based on all previous reaction times, which does not prevent participants from deliberately going slower on low reward trials, so that it is easier to gain more points on high reward trials. Our design circumvented this bias by individualising points criteria separately for high and low reward conditions. In experiment 3.2, no reward prospect cues or RT incentives were used. Instead, random high (+100) or low (+10) reward feedback was presented for correct responses from this stage onwards.

From this stage of training, the procedure became identical to the main experimental task. Participants completed one practice block of 65 trials in this format. In experiment 3.1, participants with accuracy below 70% in the practice block were given

another practice block until they met this training criterion. In experiment 3.2, the same criterion was applied to progress to the main task, up to a maximum of four practice blocks. Following each block, participants were presented with their average RT and percentage of correct responses. In experiment 3.1, participants additionally rated how much attention they had paid to the reward cues on a scale from one to nine at the end of each block (1=I was completely unaware of whether each trial had high or low reward; 9=I was completely aware of whether each trial had high or low reward).

A2. Experimental stimuli for experiments 3.1 and 3.2

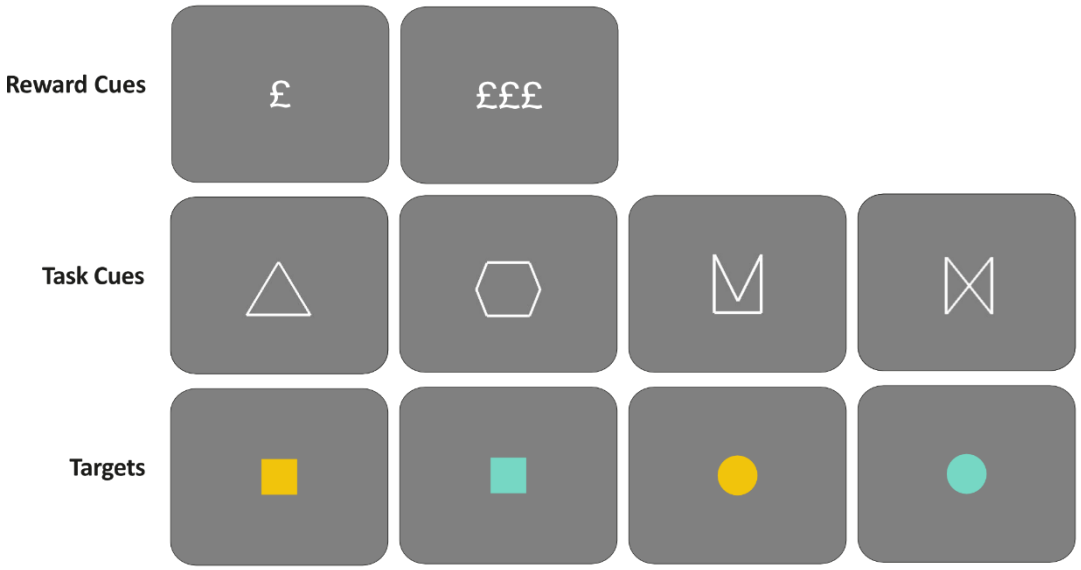


Figure A2. Experimental stimuli. Experiment 3.1 used all stimuli shown in the figure. Experiment 3.2 used all stimuli except reward prospect cues.

A3. Neural analysis for experiment 3.1, showing task coding for switch and repeat trials within reward condition.

Figure 6 gives the appearance that task coding might be equal across conditions, with the exception that it is significantly reduced on low reward switch trials. One way to

interpret this would be that interference on switch trials reduces the extent to which task rules can be distinguished in our analyses. But when reward prospect is high, the level of interference is restored to normal (i.e. repeat trial) levels. To formally test whether task coding was significantly reduced on low reward switch trials, we ran a similar analysis to Figure 6, but compared task coding for switch and repeat trials within each reward condition. Contrary to the interpretation above, the analysis did not detect a significant difference between low reward switch and low reward repeat trials (Figure A3A: window tested=1200-2500, longest difference cluster=2276-2300ms, p=0.5247). Moreover, task coding was significantly boosted on high reward switch compared with high reward repeat trials (Figure A3B: window tested=1200-2500ms, first difference cluster=1720-1836ms, p=0.0228, second difference cluster=1956-2064ms, p=0.0322). At least under high reward conditions, these results appear to be consistent with Waskom et al. (2014), who observed strongest rule coding on switch trials.

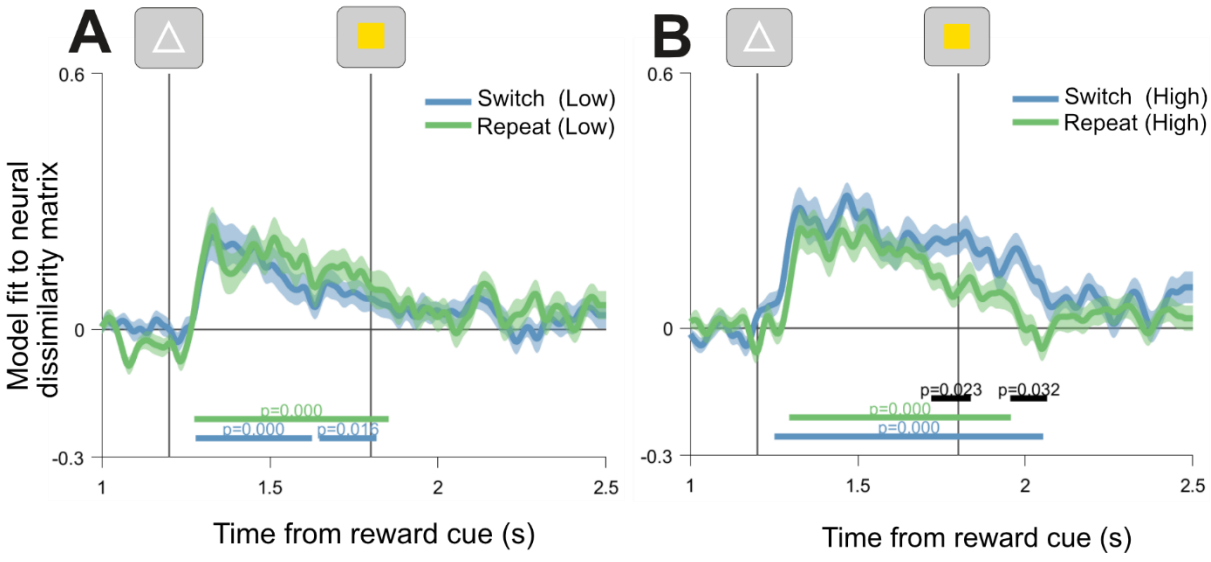


Figure A3. Time resolved regression coefficients for task coding as a function of trial sequence on low reward trials (A) and high reward trials (B). Vertical lines show onset of the task cue and target respectively. Shading around principle lines indicates standard error of mean. Cluster-corrected p-values are shown below time courses.

Chapter 4. Theoretical influence of option-specific properties on switch decisions[2]

Abstract

Many complex real-world decisions, such as deciding which house to buy or whether to switch jobs, involve trying to maximize reward across a sequence of choices. Optimal Foraging Theory is well suited to study these kinds of choices because it provides formal models for reward-maximization in sequential situations. In this article, we review recent insights from foraging neuroscience, behavioral ecology, and computational modelling. We find that a commonly used approach in foraging neuroscience, in which choice items are encountered at random, does not reflect the way animals direct their foraging efforts in certain real-world settings, nor does it reflect efficient reward-maximizing behavior. This is corroborated through proof-of-concept simulations showing that random foragers accrue significantly less reward and spend spending longer exploiting items, than foragers that direct their encounters with choice items, using knowledge about item properties. Based on this, we propose that task designs allowing subjects to encounter choice items strategically will further improve the ecological validity of foraging approaches used in neuroscience, as well as give rise to new behavioral and neural predictions that deepen our understanding of sequential, value-based choice.

[2] Most of this chapter is published in: Hall-McMaster, S., & Luyckx, F. (2019). Revisiting foraging approaches in neuroscience. *Cognitive, Affective, & Behavioral Neuroscience*, *19*(2), 225-230. The section on 'proof-of-concept simulations' did not appear in this publication, but aims to more formally test the main arguments about theoretical benefits of using item-specific knowledge to guide switch decisions.

Introduction

For many decades, neural processes governing reward maximization have been a central interest in decision neuroscience. One framework that has contributed to our understanding of these processes is reinforcement learning (RL), in which an animal learns to maximize long-run future reward by trial-and-error (Sutton & Barto, 1998) or model-based strategies (Lee et al., 2014). As one example of this contribution, studies in RL have led to the proposal that midbrain dopamine neurons convey a prediction error signal to dorsal anterior cingulate cortex (dACC), allowing it to influence motor control selection when errors are made, or reward outcomes are lower than expected (Holroyd & Coles, 2002; Silvetti et al., 2014).

In recent years, RL has been complimented by a new approach, based on Optimal Foraging Theory (Stephens & Krebs, 1986). The foraging approach has two main characteristics. The first is that it focuses on optimization problems, in which subjects must maximize a reward quantity in relation to another variable, such as time or effort expended. The second is that foraging tasks consist of a foreground option, which is currently being pursued, and background options the subject is not currently engaged with (Hayden, 2018; Hayden & Walton, 2014; Pearson et al., 2014; Stephens, 2008). The foraging approach is thus distinct from many RL and economic choice tasks, in which choice items are presented simultaneously (Mobbs et al., 2018; Pearson et al., 2014). In addition, most foraging studies in neuroscience do not allow decision-makers to direct their encounters with choice items, a feature that is distinct from many sequential RL tasks, in which choice items or bandits can be revisited through model-based planning (Daw et al., 2011; Kool, Gershman, et al., 2017; Mobbs et al., 2018).

Two long-standing theoretical problems in animal behavior research have guided the recent interest in foraging within neuroscience (Kolling et al., 2012; Pearson et al., 2014). The first deals with *stay-switch* (or patch-leaving) dilemmas, in which an individual experiences an option with diminishing returns and must decide when it is best to leave that option. The other deals with *accept-reject* (or diet-selection) dilemmas, in which an individual chooses whether to engage with an option or ignore it in search of a better one (Stephens, 2008; Stephens & Krebs, 1986).

Recent applications of foraging approaches in neuroscience have drawn on foundational models in Foraging Theory (Charnov, 1976; Stephens & Krebs, 1986), generating new avenues and insights in the study of sequential, value-based choice (Hayden et al., 2011; Kolling et al., 2012). Nevertheless, growing evidence from behavioral ecology demonstrates that these approaches do not capture situations in which animals can use their knowledge of the environment to encounter choice items strategically (Fagan et al., 2013; Janmaat et al., 2013; Menzel, 1999; Merkle et al., 2014, 2017; Sayers & Menzel, 2012). The purpose of this chapter is to a) review current approaches in foraging neuroscience, b) present evidence that these approaches are violated by animals in some settings, c) present simulations comparing the performance of traditional foraging agents (i.e. that use decision rules considered in neuroscience studies to date) against those that leverage item-specific properties and d) to provide practical recommendations that offer new scope for understanding sequential, value-based choice.

Insights from foraging neuroscience

Neural investigations using stay-switch and accept-reject tasks have primarily revealed important contributions of dACC to decision-making. Research using the stay-

switch framework has shown that firing rates of neurons in monkey dACC positively correlate with decisions to leave a patch (option disengagement). These firing rates scale with increasing travel time to the next patch, suggesting that dACC neurons may set a threshold for switching and that this threshold is sensitive to the time cost of switch decisions (Hayden et al., 2011). Research using the accept-reject framework has shown that neurons in monkey dACC encode the time cost associated with choosing an option (option engagement). The activity patterns in a subset of dACC neurons also correlate with the value of rejected offers, suggesting that dACC keeps track of post-decision variables (Blanchard & Hayden, 2014).

In addition to insights from non-human primates, accept-reject studies in humans have played an important role in functional comparisons between dACC and the ventromedial prefrontal cortex (vmPFC) (Rushworth et al., 2012). In particular, studies comparing accept-reject choices with economic ones have shown that BOLD activity in the dACC is positively correlated with the value of searching for alternative options, whereas vmPFC activity is positively correlated with the chosen option's reward magnitude during binary economic choices (Kolling et al., 2012). These results seem to suggest that dACC plays a more prominent role in decisions to engage with options or search for alternatives and that the vmPFC plays a more prominent role in classic binary economic decisions. However, more recent work has provided evidence for another interpretation of dACC function in foraging scenarios, wherein dACC BOLD activity reflects choice difficulty (Ebitz & Hayden, 2016; Shenhav et al., 2014). Additionally, BOLD signals in dACC and vmPFC have been shown to correlate with the relative value of searching for alternatives over engaging with current options, suggesting that both regions may represent similar information in foraging scenarios (Shenhav, Straccia, et al., 2016).

Alongside ongoing debate about frontal contributions to foraging, two major trends are emerging in foraging neuroscience. One is that we are beginning to expand our knowledge of foraging circuitry beyond the dACC and vmPFC (Barack et al., 2017; Kane et al., 2017; Lottem et al., 2018). For example, it has recently been shown that neurons in posterior cingulate cortex (PCC) ramp up activity several seconds before monkeys make switch decisions, suggesting PCC signals reflect either an increasing motivation to disengage from current options, the value of alternative options or choice difficulty (Barack et al., 2017). In addition, recent evidence demonstrates that pharmacogenetically increasing neuronal activity in the locus coeruleus, a major site of norepinephrine neurons, causes rats to make switch decisions significantly earlier than control animals (Kane et al., 2017). By contrast, optogenetic activation of neurons in the dorsal raphe nucleus, a major site of serotonin neurons, causes mice to persist with active exploitation decisions for longer and abandon patches later than controls (Lottem et al., 2018). Taken together, these results suggest that several regions beyond dACC, including subcortical structures with distinct neurotransmitter classes (catecholamines and monoamines), make important contributions to the neural computations performed during stay-switch dilemmas and the choices that result.

The second major trend is the increasing use of RL techniques in foraging research, to model behavior in environments where on-going learning is needed to optimize decisions (Constantino & Daw, 2015; Frankenhuis et al., 2018; Kolling & Akam, 2017; Wittmann et al., 2016). One classic model for optimal stay-switch behavior in foraging is called the marginal value-theorem (MVT), which states that, for environments with patches that have monotonically depleting reward rates, it is optimal to leave the current patch when its reward rate drops to the average reward rate of the environment

(Charnov, 1976). Constantino and Daw (2015) showed that MVT better predicts trial-by-trial choices in a stay-switch task than a temporal difference RL model. However, when patches are not monotonically depleting or people need to anticipate future values within a patch, models using RL parameters can provide more optimal solutions (Kolling & Akam, 2017). For instance, an RL model that generates value estimates for stay-switch choices by integrating reward information over multiple time scales successfully predicted whether people would switch away from current options in a dynamic patch environment, and components of this model were reflected in BOLD activity of human dACC (Wittmann et al., 2016).

The case for directed encounters

In many of the foraging studies reviewed above, subjects have limited (and often no) control over the successive items they encounter (Constantino & Daw, 2015; Hayden et al., 2011; Shenhav et al., 2014; Shenhav, Straccia, et al., 2016). In other words, new prey and patch items are encountered randomly or pseudorandomly. The use of this approach in neuroscience has undoubtedly deepened our understanding about foraging behavior and sequential decision-making more broadly (Hayden et al., 2011; Kolling et al., 2012). Nevertheless, mounting evidence from studies in behavioral ecology indicate that designs with purely random encounters do not reflect situations in which animals can direct their foraging efforts by choosing to revisit valuable patches or prey items. To illustrate, consider a fruit tree that contains a small portion of ripe fruit and a large portion of unripe fruit. The ripe fruit can be harvested immediately but the animal may also benefit from revisiting the tree to harvest the remaining fruit once it ripens. Similarly, it may be beneficial to remember the locations of trees bearing a large amount of fruit in the

summer months, so they can be exploited the following season. This kind of foraging strategy, in which high value items are revisited and low value items are avoided, could greatly reduce search time and thereby improve biological fitness (Merkle et al., 2014, 2017; Riotte-Lambert et al., 2015; Sayers & Menzel, 2012).

A substantial body of empirical work corroborates the notion that, in several situations, animals do not encounter choice items at random but use their knowledge of the environment to encounter items strategically (Fagan et al., 2013). Bison move to meadows that have previously been profitable, especially after encountering a poor quality patch (Merkle et al., 2014). When modelling bison patch selection, a model using random patch encounters is less efficient than a model incorporating memory processes, and cannot fully capture bison choice behavior (Merkle et al., 2017; Riotte-Lambert et al., 2015). In non-human primates, chimpanzees have been observed to monitor candidate fruit trees during travel but also engage in goal-directed travel to specific trees, based on the size and location of trees they have learned about from previous visits (Janmaat et al., 2013). Alongside real-world studies, more controlled laboratory experiments indicate that when animals are allowed to direct their search strategy, they are unequivocally non-random. Chimpanzees direct their search by integrating information about food quantity, handling time, and distance from previous encounters to maximize the rate of energy intake (Sayers & Menzel, 2012), and rhesus monkeys rarely revisit patches they know to be empty from previous experience, but do so when memory is impaired by lesioning the sulcus principalis (Passingham, 1985).

Situations allowing animals to direct their encounters with choice items introduces a new decision dilemma. In stay-switch situations, for example, animals must not only decide *when* to disengage from the current option, but also *where* to move next. In

natural environments, this selection dilemma may be especially relevant to central-place foragers, such as bees or owls, who undertake foraging trips from a central hive or nest. Evidence suggests that patch-selection choices of central-place foragers are not random (Rosenberg & McKelvey, 1999) and that directed selection-based knowledge of the environment can increase reward intake when patches are spatially clustered (Barraquand et al., 2009). This selection dilemma may also be pertinent to primates, who exhibit a range of foraging strategies, such as foraging from multiple central-place locations (Chapman et al., 1989) and exploiting food items within group territories (Garber, 1987). In particular, the use of item-specific knowledge by primates, to decide what to exploit next, is thought to be especially important in environments where sensory information is limited, such as dense canopies where visual cues are restricted to immediate surroundings (Garber, 1987). This is not to say that all adaptive foraging behaviour involves a decision about what to exploit next. Indeed, random sampling of choice items is important for discovering new options and learning about how options might have changed, a balance known as the exploration-exploitation trade-off in RL (Sutton & Barto, 1998). Nevertheless, the assumption underpinning most foraging neuroscience tasks, that agents always encounter items at random, does not account for a critical aspect of real world foraging choices which, as we have outlined above, often involves leveraging item-specific knowledge to guide inter-dependent choices about a) when to stop exploiting an option and b) what option to exploit next.

Practical suggestions: Implications for decision making computations

Based on the evidence outlined above, one way to extend foraging approaches in neuroscience is to allow non-random encounters in future designs. For illustrative

purposes, we can imagine a simple grid-world task as one possible assay. The grid contains a number of randomly distributed 'patch' tiles, from which rewards can be harvested and a number of intervening tiles, which represent the travel time between patches. Following a leave decision, participants select the next patch tile they wish to visit, which could include both previously visited patches and unexplored ones. To provide context to the predictions that follow, we assume patches have different maximum reward values but follow the same depletion function when being exploited. We also assume the maximum reward capacity of each patch is fixed and that subjects' patch selections are deterministic. Our final assumption is that patches replenish to their maximum faster than the travel time needed to reach them. Under this final assumption, MVT converges to the optimal solution (Possingham & Houston, 1990), making it an adequate predictive framework for our purposes.

With these changes in mind, the empirical work reviewed above implies several behavioral and neural predictions. Behaviorally, one prediction is that patch selection will become less random over time. From an RL perspective, this can be viewed as a reduced exploration relative to exploitation (Sutton & Barto, 1998). As subjects learn about the value and distribution of different patches, they can encounter items strategically by revisiting high-value items, which have high reward-travel time ratios. Once patch values have been learnt, designs allowing directed encounters should produce lower switching thresholds than random encounter designs. To explain this, we will distinguish between *global* and *local* switching thresholds. A *global* switching threshold refers to the average experienced reward rate, because it is a global measure based on all items experienced in the environment. A *local* switching threshold refers to an item-specific measure about when to switch. In this chapter, we use local threshold to mean the highest starting value-

travel time ratio of the alternative options. However, a local threshold could also be operationalized in other ways, such as the average reward rate of a specific item. Now, let us return to why switching thresholds should be lower in directed encounter designs. From the perspective of MVT, choosing only options with high reward-travel time ratios will increase the average experienced reward rate, allowing this global switching threshold to be met more quickly by the current option. From another perspective, subjects may set their leaving threshold to the next best alternative (e.g., a local threshold set to the alternative with the highest starting value-travel time ratio) in a directed encounter design. Because low value options can be avoided, this local threshold would also be met more quickly by the current option. Whether subjects use a global or local leaving threshold in directed designs is an empirical question. For our purposes, what is most important is that both perspectives predict lower leaving thresholds, compared with a random encounter design in which MVT provides the optimal solution. From a neural perspective, if the firing rate of dACC neurons reflects a threshold for switching (Hayden et al., 2011), dACC thresholds for stay-switch decisions should be lower in task designs that allow directed encounters compared with random encounter designs. Again, this is because repeated exposure to low-value alternatives can be avoided in directed encounter designs, resulting in a lower global or local threshold for switching away from current options.

Proof-of-concept simulations

To more formally test the proposal that directing encounters with choice items should lead to lower switching thresholds and higher reward accrual, we simulated agents in a simple stay-switch decision-making task. Five choice items (patches) were randomly

distributed in a 10x10 grid and assigned a random starting value between 1 and 50 points. All non-assigned grid points had a value of zero throughout simulations. On the first trial, agents began in a random patch and could exploit the current patch once per time step. Following previous work, the amount of reward gained with each step followed an exponential decay (Kane et al., 2017). This was given by the expression: $r(t) = sv * e^{-t}$, where *r* is reward, *t* is time in the current patch (i.e. the number exploit actions taken) and *sv* is the patch starting value. Because these simulations involved one exploit action per time step, r was equivalent to the instantaneous reward rate for the current item (i.e. $\frac{r}{t} = \frac{r}{1} = r$).

We simulated two classes of decision-making agent. The first class was a random agent, which exploited patches using a marginal value rule. Specifically, this agent exploited a given patch until the reward received on the current step fell below a global reward rate. Following Kane et al. (2017), the global reward rate was computed as: $g = \frac{rt}{tt}$, where *rt* is the total reward gained from the beginning of an agent simulation and *tt* is total time steps since the agent simulation began. Decisions about whether to leave the current item, using *g*, were made after each exploit action. Therefore, *g* did not need to be initialized before it could be calculated directly, using trial information. The value of *g* was re-calculated after each exploit action.

Once the current reward rate fell below the global reward rate, the random agent selected a new patch at random to travel to. For simplicity, agents travelled to randomly selected patches directly, as opposed to encountering patches through a random walk along the grid. The time to reach a new patch was the Euclidean distance between the current location and the patch location on the grid, calculated as the hypotenuse joining

the two grid points. This distance was divided by a movement speed parameter, which was set to 1 for the following simulations, meaning that agents could move one grid point per time step during travel.

While travelling, all patches replenished towards their maximum value, a feature distinct from foraging studies to date (Constantino & Daw, 2015; Hayden et al., 2011; Kolling et al., 2012; Shenhav et al., 2014; Shenhav, Straccia, et al., 2016). The growth in each patch was given by the linear function: $pv = pvl + rr * tr$, where *pvl* is the value of a patch when leaving the current patch, *rr* is the replenish rate and *tr* is the travel time to the next patch. Patch values also replenished while exploiting a current patch. However, unlike growth during travel, this only applied to unattended patches. The replenish function for unattended patches was given by: $pv(t) = pv(t-1) + rr$, where *pv(t-1)* is the value of a patch on the previous time step, *rr* is the replenish rate and *t* is the time (number of exploit actions taken) in the current patch. Importantly, a patch would stop replenishing when it had reached its maximum (i.e. its starting value). In situations where a patch replenished beyond its maximum, it would be rounded down to its maximum value.

The second class of agent was a directed forager. Unlike random agents, which used a global leaving threshold and selected patches at random, directed agents used local patch information to determine when to leave and where to go next. In particular, the leaving threshold for these agents was set to the highest reward-travel time ratio among the alternative patches. This was computed as: $lt = \frac{pvl\,(ax)}{d\,(ax)}$, where *lt* is the leaving threshold, *pvl(ax)* is the value of an alternative patch (ax) and *d* is the distance to an alternative patch (ax) from the current location. Note that this is not necessarily the optimal leaving threshold. Rather, it is a proof-of-concept example of a local threshold

agents could use to make patch leaving choices, which leverages knowledge about item-specific properties (in this case current values and distance).

Using this simulation setup, we simulated 100 random and 100 directed agents, each performing patch-leaving decisions for 5000 time steps. We then compared random and directed agents on four dependent measures. These included: a) the average time steps spent in a patch before leaving, b) the average leaving threshold, c) the average number of leave decisions performed within the 5000 available steps and d) the average reward accrued. Based on the theoretical proposal outlined in this chapter, we predicted that directed agents would a) spend significantly less time in patches on average, b) have significantly lower leaving thresholds, c) would complete more patch leaving choices within the allotted time steps and d) gain more reward overall.

The simulation results corroborated these predictions. In the first set of simulations, the replenish rate was set at 10 times the movement speed, so that patches were always full when the agent reached the next patch (Figure 1). This was done to ensure the agent did not change the value of the item it encountered based on its previous actions, which retains the potential validity of using a marginal value decision threshold (Possingham & Houston, 1990). In this case, we found that the average patch dwelling time was significantly lower in directed compared with random agents (Figure 1A: Wilcoxon rank sum Z=-16.4161.1007, p<0.0001). Consistent with such behaviour, directed agents left patches when the current reward was significantly higher than random agents, indicating a lower leaving threshold for patch leaving (Figure 1B: Wilcoxon rank sum Z=17.2967, p<0.0001). Directed agents also completed significantly more leave decisions within the 5000 time steps (Figure 1C: Wilcoxon rank sum Z=-14.229, p<0.0001)

and accrued significantly more reward than their random counterparts (Figure 1D:

Wilcoxon rank sum Z=-12.8509, p<0.0001).

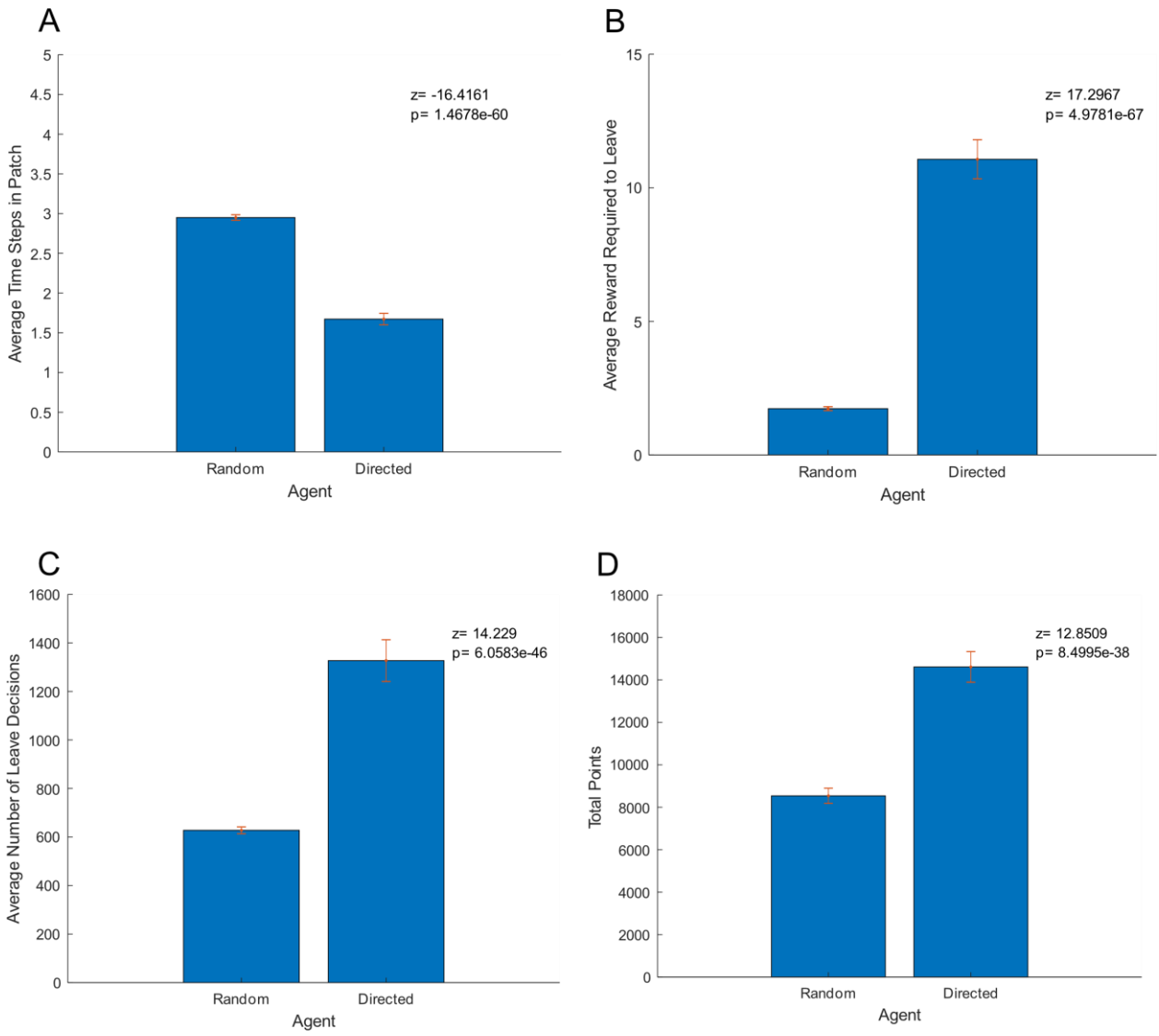


Figure 1. A-D: Simulation results when replenish rate is high (10) for random and directed agents. Random agents left a patch when the current reward rate fell below the average reward rate, computed over the whole simulation for that agent. It then selected an alternative patch at random to travel to. The directed agent used the alternative patch with the highest reward-travel time ratio, as its threshold for leaving, and travelled to that alternative when deciding to leave. A: Average number of time steps in each patch before leaving for random and directed agents. B: Average reward used as a threshold for leaving the current option for random and directed agents. Note that a higher value indicates a lower threshold for leaving because higher values will be met more quickly by reward from the current option, during its exponential decay. C: Number of leave decisions made within the 5000 allotted time steps, averaged separately over random and directed agents. D: Total points accrued during the 5000 time steps, averaged separately over random and directed agents. A-D: Error bars show 95% confidence intervals.

In a second set of simulations, the replenish rate was set to a 10^{th} of the movement speed (Figure 2A-D). In contrast to the higher replenish rate reported above, previous exploit decisions of the agent now could (and did) influence the reward amount encountered when arriving at a new patch. That is, if an agent exploited a given patch, the slow replenish rate meant the patch was unlikely to have returned to its full value when the agent returned. Under these conditions, we observed the same pattern of results. Directed agents spent less time before leaving each patch on average (Figure 2A: Wilcoxon rank sum Z= -16.3176, p<0.0001) and left patches when the current reward was higher, indicating a lower leaving threshold (Figure 2B: Wilcoxon rank sum Z= 17.2396, p<0.0001). Directed agents also completed more leave decisions within the allotted time steps (Figure 2C: Wilcoxon rank sum Z= 13.969, p<0.0001) and gained more reward overall (Figure 2D: Wilcoxon rank sum Z=3.0926, p=0.001984). Under these low replenish conditions, the reward advantage appeared to be more modest, with directed agents earning 641.8 points on average compared with 626.6 points for random agents. Follow up simulations indicated that the reward benefit for directed agents over random agents seems to scale with increasing replenish rate (Appendix, Figure A2.1).

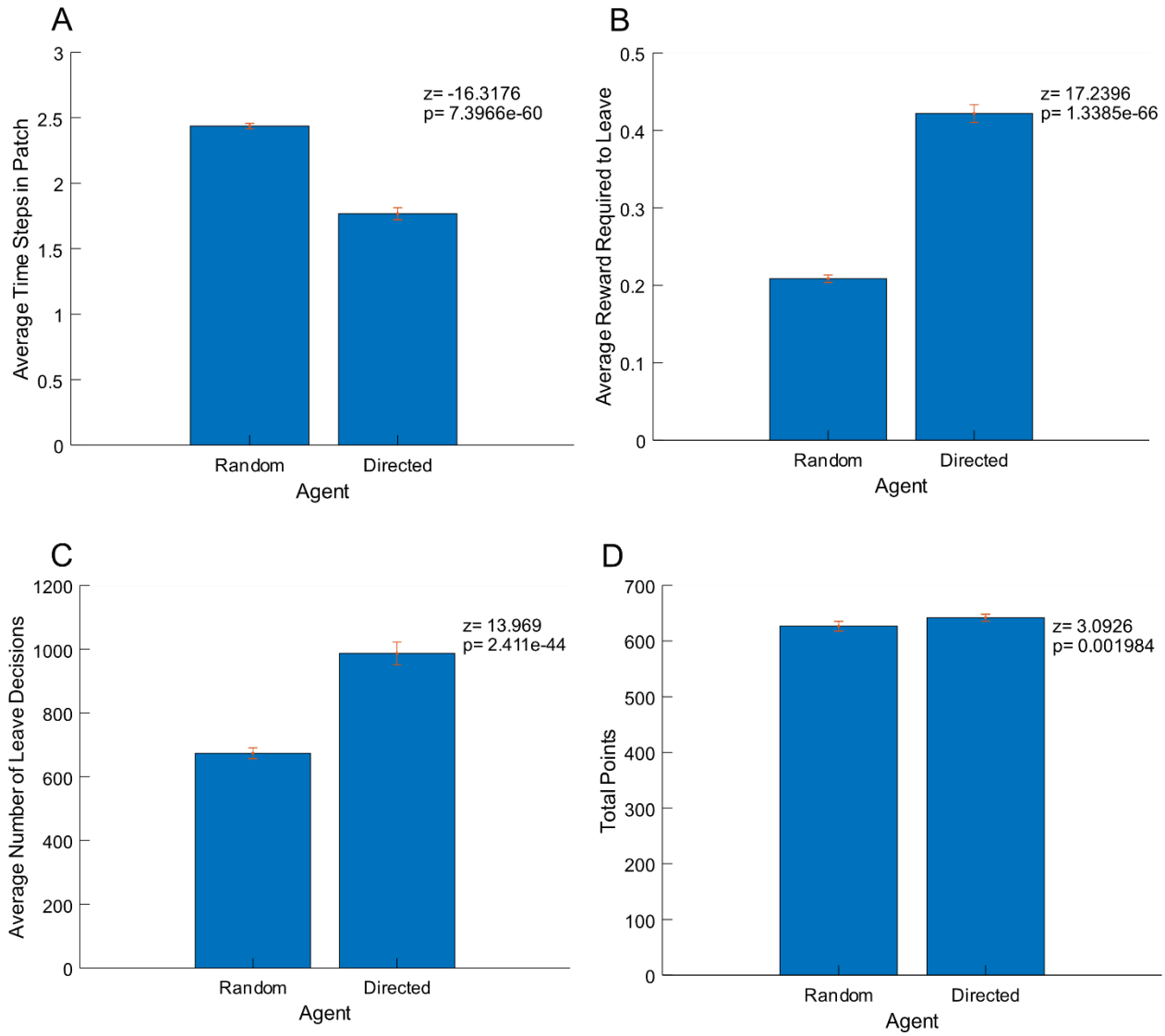


Figure 2. Simulation results when replenish rate is low (0.1) for random and directed agents. A: Average number of time steps in each patch before leaving for random and directed agents. B: Average reward used as a threshold for leaving the current option for random and directed agents. Note that a higher value indicates a lower threshold for leaving because higher values will be met more quickly by reward from the current option, during its exponential decay. C: Number of leave decisions made within the 5000 allotted time steps, averaged separately over random and directed agents. D: Total points accrued during the 5000 time steps, averaged separately over random and directed agents. A-D: Error bars show 95% confidence intervals.

To summarise, these proof-of-concept simulations show that agents directing encounters with choice items had lower leaving thresholds and accrued more reward than agents encountering choice items at random. These simulations constitute a highly simplified decision scenario, with two notable limitations. First, directed agents had full knowledge about item properties, such as starting value and location. While we used a relatively small numbers of patches in these simulations, this assumption is unrealistic for biological agents, particularly in environments with large numbers of choice items. A similar point has been made in the domain of RL, in which memory constraints are known to influence the time humans take to learn stimulus-response mappings, in a way that is not captured by memory-independent RL models (Collins & Frank, 2012). Second, these simulations assume a strict separation between directed and random encounters. While a purely directed agent performs well in the current environment, with a static distribution of options, it is unlikely to perform well in an environment where new options are introduced over time. In particular, it would fail to discover new options because it relies exclusively on knowledge about known items to decide when to leave and where it will go next. Thus, these particular simulations do not address how an important balance should be struck in naturalistic environments, between directing encounters with known options and exploring new options that could increase reward over the long run. Such a balance could be investigated in more complex simulation environments, by including a parameter that controls the balance between random and directed choice, a common feature in many RL models (Sutton & Barto, 1998). Despite these limitations, the present simulations provide initial support that – at least in principle - directing encounters with choice items could help to promote greater reward maximization than completely random encounters with choice items.

Discussion

The suggestions and simulations presented in this work represent a first step towards the development of new foraging approaches in neuroscience, which allow directed encounters with choice items. More complex designs could consider different relationships between replenish rate and travel time. The importance of these factors has been highlighted by Possingham and Houston (1990), who show that when patches replenish slower than the time it takes to revisit them, MVT fails to maximize reward. This has also been demonstrated in stochastic environments, where repeated sampling is required to determine patch quality, or environments where reward rates initially increase as the richest part of the patch is located (McNamara, 1982; Oaten, 1977). In these latter situations, Kolling and Akam (2017) propose that comparing the expected future reward rate of a patch against the global average will lead to more optimal patch-leaving choices than traditional MVT (Charnov, 1976). Thus, RL frameworks may be increasingly important for understanding choice behavior in more complex foraging designs, such as those that allow patch revisiting.

In addition to these considerations, dACC has many competing theories about its role in decision-making. As a starting point, we have framed our neural predictions under the view that dACC may encode a threshold for switching away from current options (Hayden et al., 2011). A related view proposes that dACC modulates behavioral flexibility by tracking the value of alternative choice options (search value) (Kolling, Behrens, et al., 2016; Kolling, Wittmann, et al., 2016). These accounts may not be mutually exclusive (Blanchard & Hayden, 2014); however, it remains an open question whether dACC activity will be lower when switches are made in directed compared with random encounter designs, reflecting a lower activity threshold for switching (Hayden et al., 2011), or

whether dACC activity will be higher due to higher search values (Kolling, Behrens, et al., 2016; Kolling et al., 2012; Kolling, Wittmann, et al., 2016). In contrast to accounts focusing on dACC's role in behavioral persistence and flexibility, Expected Value of Control Theory contends that dACC signals the need for cognitive control and the value of exerting control, given its computational cost. By integrating this information, the theory ascribes dACC a key role in determining where cognitive control should be directed and how much it should be deployed (Ebitz & Hayden, 2016; Shenhav, Cohen, et al., 2016; Shenhav et al., 2014). Single-cell recordings and anatomical studies provide evidence for an integrative view, wherein dACC tracks contextual variables. This view emphasises dACC's role in linking task contexts with appropriate action plans, where action plans could constitute a decision to about whether to abandon the current item or whether to increase cognitive control (Heilbronner & Hayden, 2016). To connect this idea to the framework developed in this chapter, this view predicts that dACC should encode all relevant information that define the current task state (e.g. current item value, value and distance of next best alternative) and action information (e.g. stay-switch decision variables, such whether the current item is above or below the leaving threshold). Foraging studies have played a prominent role in contemporary debate about dACC (Ebitz & Hayden, 2016) and the development of new foraging approaches, such as those involving dynamic replenish rates or allowing patch revisiting, may be useful to further distinguish competing theories about its function.

In this respect, one limitation of the stay-switch designs we have covered is that there may be an insufficient range of foreground and background values to fully decouple flexibility (Kolling, Wittmann, et al., 2016) from choice difficulty accounts (Shenhav, Cohen, et al., 2016; Shenhav et al., 2014). Subjects are unlikely to stay in patches until the

current value is dramatically lower than the average experienced reward rate, whereas it is much easier to present a low value foreground offer against a high value set of alternatives in accept-reject studies. The use of directed encounters in accept-reject designs may thus be useful in considering further tests of dACC theories. In such designs, participants could have the choice about whether to engage with a two-option foreground gamble, with a set of alternatives visible in the background (Kolling et al., 2012; Shenhav et al., 2014; Shenhav, Straccia, et al., 2016). Rejecting the foreground offer, the subject would be able to select an option from a set of background alternatives as one of the next foreground options, in contrast to previous studies where the new foreground options were randomly drawn from the set of alternatives. Note that while it is conceivable subjects could continue to reject offers until they got offered a specific high value item in these previous studies (Kolling et al., 2012; Shenhav et al., 2014; Shenhav, Straccia, et al., 2016), reject decisions do not seem to be influenced by highest value alternative in this previous task design (Kolling et al., 2012). This might be because repeatedly rejecting offers to reach high value items would constitute a much more indirect form of the directed encounter design we propose, which enables direct selection from the background set for the subsequent two-option foreground gamble. In a similar manner to stay-switch designs, one prediction in the design we suggest is that rejection thresholds will no longer be set based on the average value of the alternatives (a global threshold) but on the most valuable alternative among the set (a local threshold). Note that in this case the global threshold refers to the average value of presented alternatives, whereas in stay-switch designs it refers to the average experienced reward rate. If dACC activity correlates with choice difficulty (Shenhav, Cohen, et al., 2016; Shenhav et al., 2014), we should observe a difference in activity for the same set of alternatives. Specifically, if the

value of the alternatives being considered are different under random and directed encounter conditions (e.g., background average vs. best alternative), this would make the decision to reject more or less difficult, and this difference should be reflected in dACC activity under the choice difficulty account. Therefore, while many of our practical suggestions have been framed under the broad view that dACC contributes to behavioral flexibility (Hayden et al., 2011; Kolling et al., 2012; Kolling, Wittmann, et al., 2016), the example above shows that directed encounters can be applied when considering other theories of dACC. In addition, the development of new foraging approaches is certainly not restricted to understanding dACC function and may be important for supporting the recent trend of investigating other cortical and subcortical contributions to foraging decisions (Barack et al., 2017; Kane et al., 2017; Lottem et al., 2018).

As we will see in the next step chapter, replenishing items can be used to assess how people track the changing value of switching, even when there is just one alternative. This approach is novel in three critical ways. First, it allows us to investigate how item-specific information for alternative options contributes to switch decisions, in contrast to previous studies that have focused on global information about alternative options. Second, it changes the focus of the switch decision, from when the current option should be left (based on its declining reward) to when an alternative option should be exploited (based on its growing reward). Third, it allows us to investigate how the brain tracks reward information that is not observed directly from sensory information. As an alternative item is (by definition) not currently being exploited, its current reward state is 'hidden'. Item-specific properties, such as growth rate, might therefore be used infer its changing value to make good decisions about when to exploit it. This is different from previous foraging studies, which have focused on fully observable reward information

from the current option and average reward information that is computed based on experienced reward feedback (Kane et al., 2017) or observable background options (Kolling et al., 2012; Shenhav et al., 2014; Shenhav, Straccia, et al., 2016).

Conclusion

Neuroscience has made great strides in using foraging as a tool to understand the neural computations that underlie sequential, value-based choice. In reviewing these insights, alongside those from behavioral ecology and computational modelling, we have presented evidence demonstrating that, while useful, current task designs involving random encounters with choice items do not reflect situations in which animals can make use of their knowledge in the environment to encounter items strategically (Fagan et al., 2013; Menzel, 1999; Merkle et al., 2014, 2017; Passingham, 1985; Sayers & Menzel, 2012). Based on this evidence, we propose that task designs allowing subjects to revisit choice items will further improve the ecological validity of foraging approaches used in neuroscience. In addition, we have derived several behavioral and neural predictions centered on the idea that the value of switching away from current options can be higher in directed encounter designs. This is because individuals can base their decisions to switch on a subset of valuable alternatives, modulating global or local thresholds for choice disengagement. In support of this idea, we undertook simple proof-of-concept simulations, showing that directed agents were able to garner more reward and spend less time exploiting items on average, compared with agents that encounter choice items at random. Finally, there is a clear trend towards integrating RL and foraging approaches to understand sequential choice processes in complex settings (Constantino & Daw, 2015; Kolling & Akam, 2017; Wittmann et al., 2016). The recommendations and predictions

outlined in this chapter may be useful starting points for the development of increasingly naturalistic foraging designs, in which elements like patch revisiting, replenish rates and patch distributions are manipulated to further unravel contributions of learning and memory to sequential, value-based choice.

Chapter 4 Appendix

A1. Performance of random and directed agents as a function of the number of choice items.

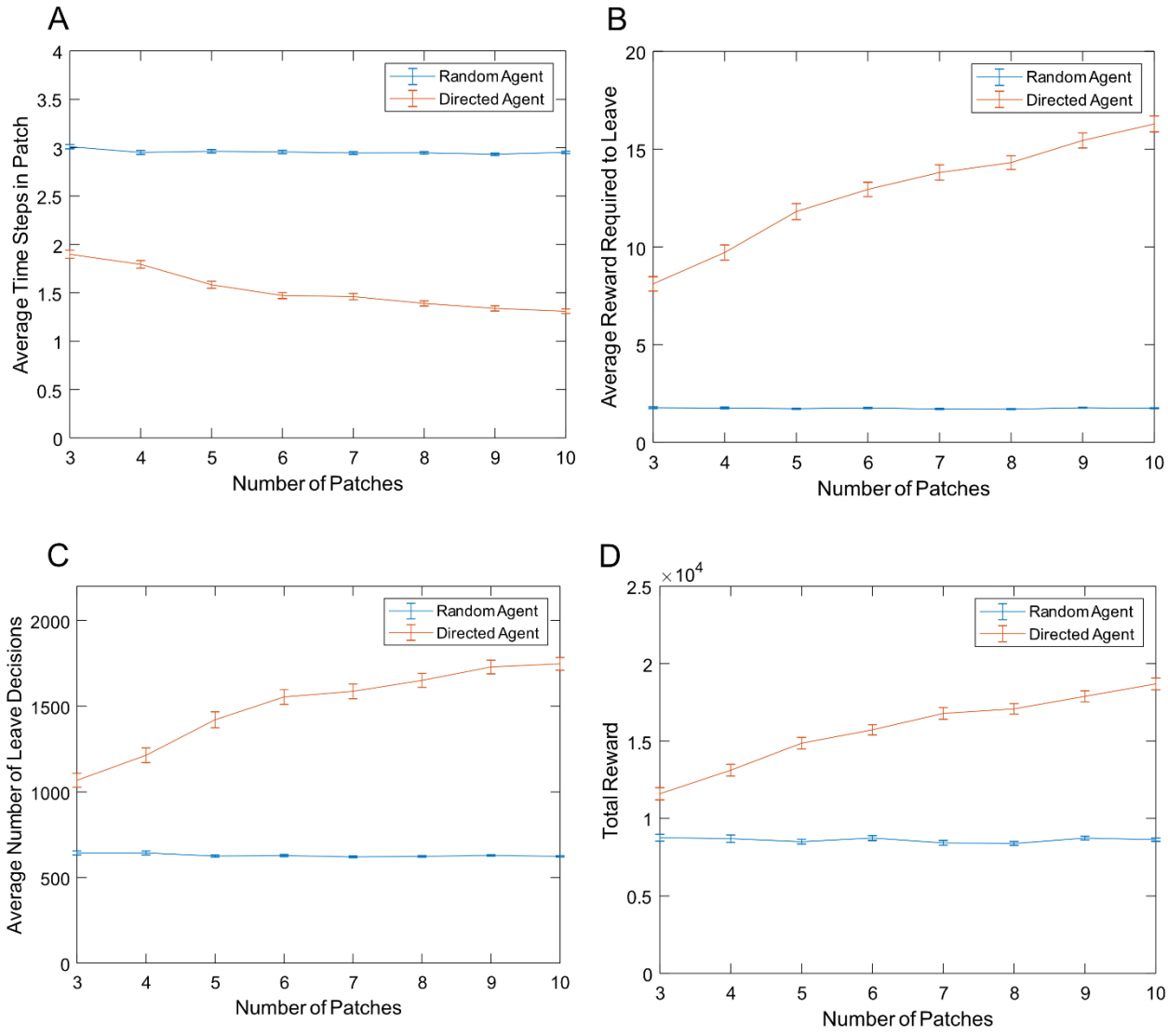


Figure A1.1. Simulation results when replenish rate is high (10) for random and directed agents, as a function of set size. The pattern of differences between random and directed agents reported in the main text holds across different set sizes. The reward benefit of directing encounters increases with set size. A: Average number of time steps in each patch before leaving for random and directed agents. B: Average reward used as a threshold for leaving the current option for random and directed agents. Note that a higher value indicates a lower threshold for leaving because higher values will be met more quickly by reward from the current option, during its exponential decay. C: Number of leave decisions made within the 5000 allotted time steps, averaged separately over random and directed agents. D: Total points accrued during the 5000 time steps, averaged separately over random and directed agents. A-D: Error bars show standard error of mean.

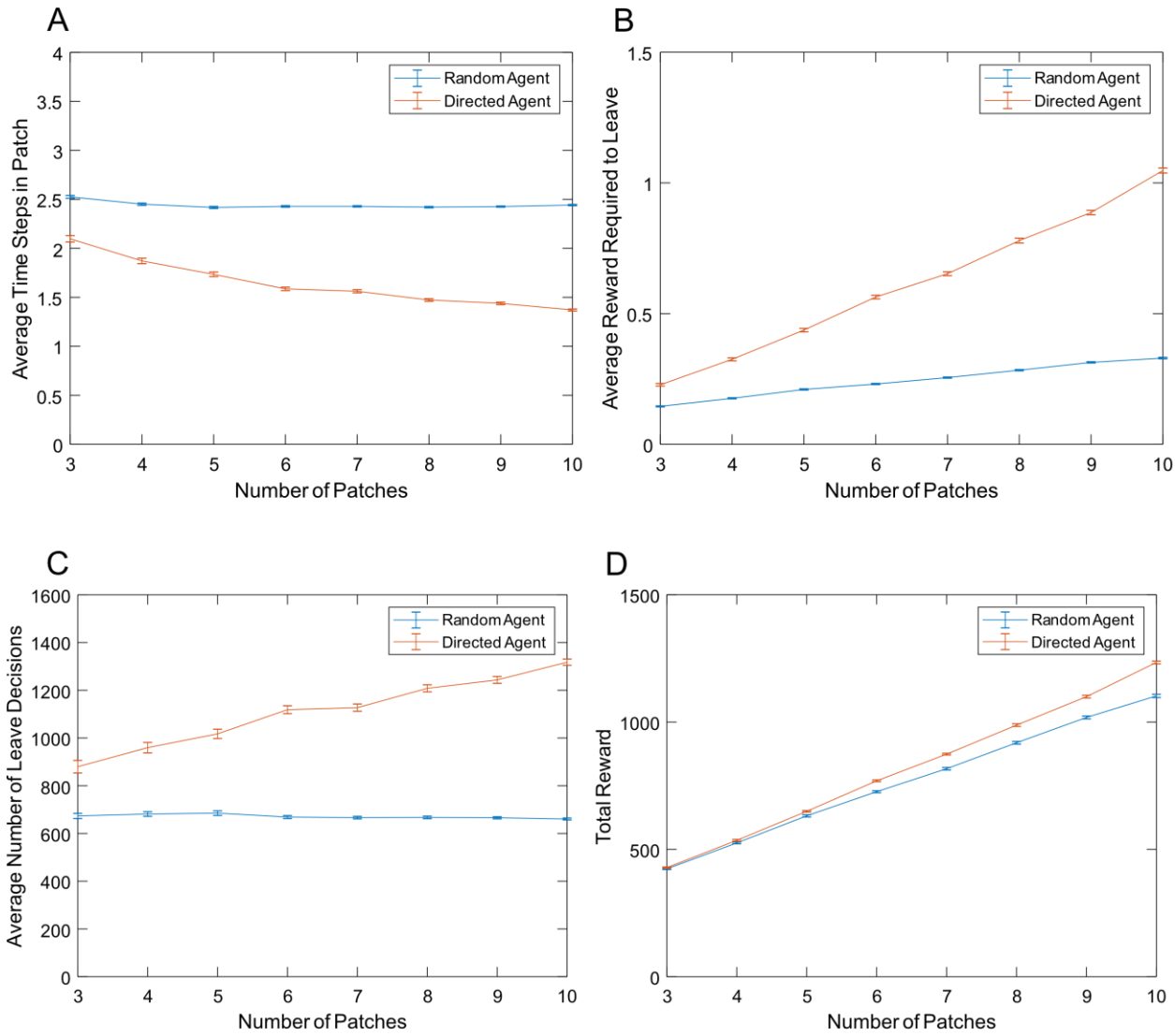


Figure A1.2. Simulation results when replenish rate is high (0.1) for random and directed agents, as a function of set size. The pattern of differences between random and directed agents reported in the main text holds across different set sizes. The reward benefit of directing encounters increases with set size. A: Average number of time steps in each patch before leaving for random and directed agents. B: Average reward used as a threshold for leaving the current option for random and directed agents. Note that a higher value indicates a lower threshold for leaving because higher values will be met more quickly by reward from the current option, during its exponential decay. C: Number of leave decisions made within the 5000 allotted time steps, averaged separately over random and directed agents. D: Total points accrued during the 5000 time steps, averaged separately over random and directed agents. A-D: Error bars show standard error of mean.

A2. Performance of random and directed agents as a function of replenish rate.

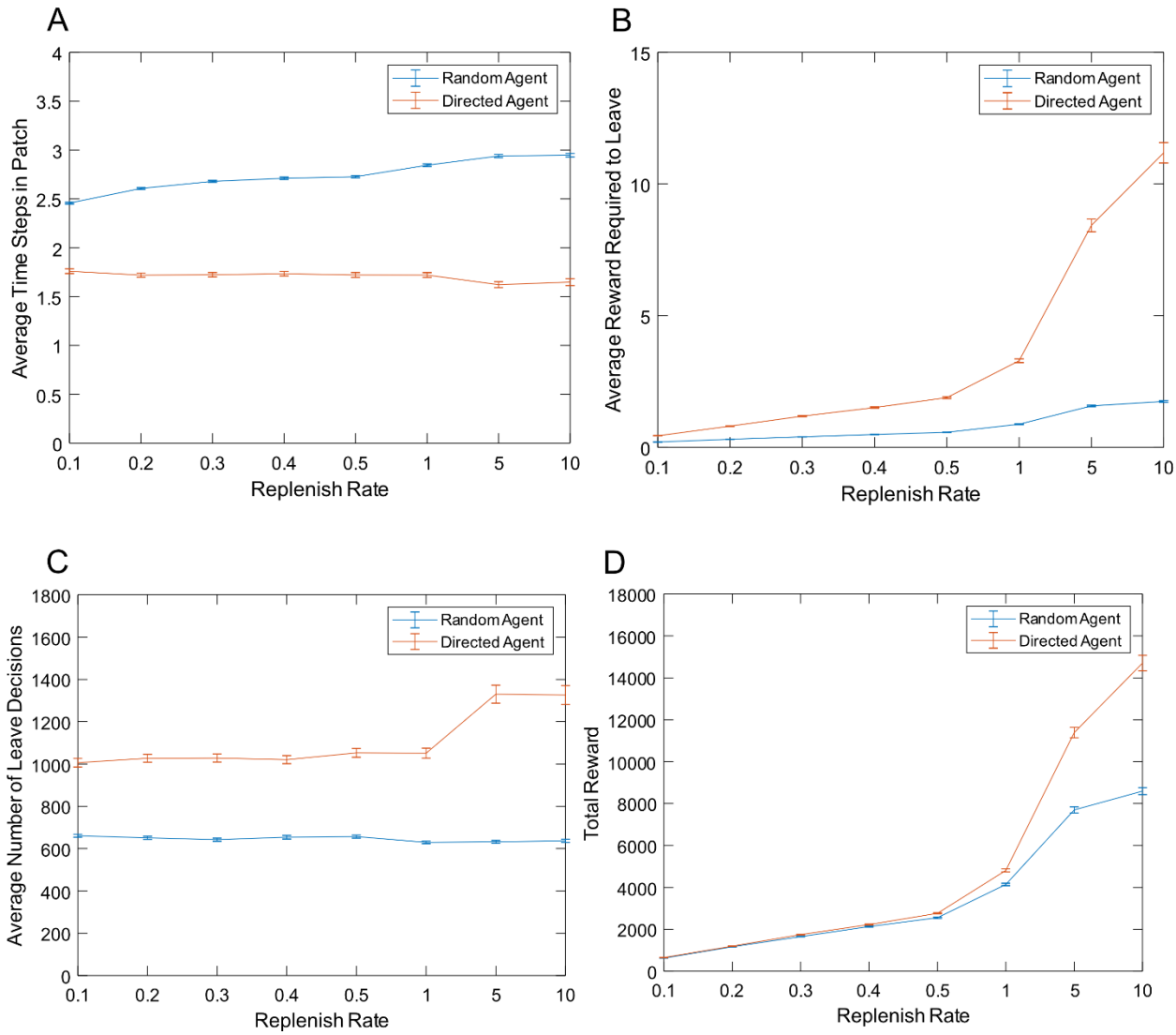


Figure A2.1. Simulation results for random and directed agents in a five patch environment, as a function of replenish. The pattern of differences between random and directed agents reported in the main text holds across different replenish rates. The reward benefit of directing encounters increases with replenish rate. A: Average number of time steps in each patch before leaving for random and directed agents. B: Average reward used as a threshold for leaving the current option for random and directed agents. Note that a higher value indicates a lower threshold for leaving because higher values will be met more quickly by reward from the current option, during its exponential decay. C: Number of leave decisions made within the 5000 allotted time steps, averaged separately over random and directed agents. D: Total points accrued during the 5000 time steps, averaged separately over random and directed agents. A-D: Error bars show standard error of mean.

A3. Performance of a directed agent that selects which item to exploit next but uses a global threshold for deciding when to leave.

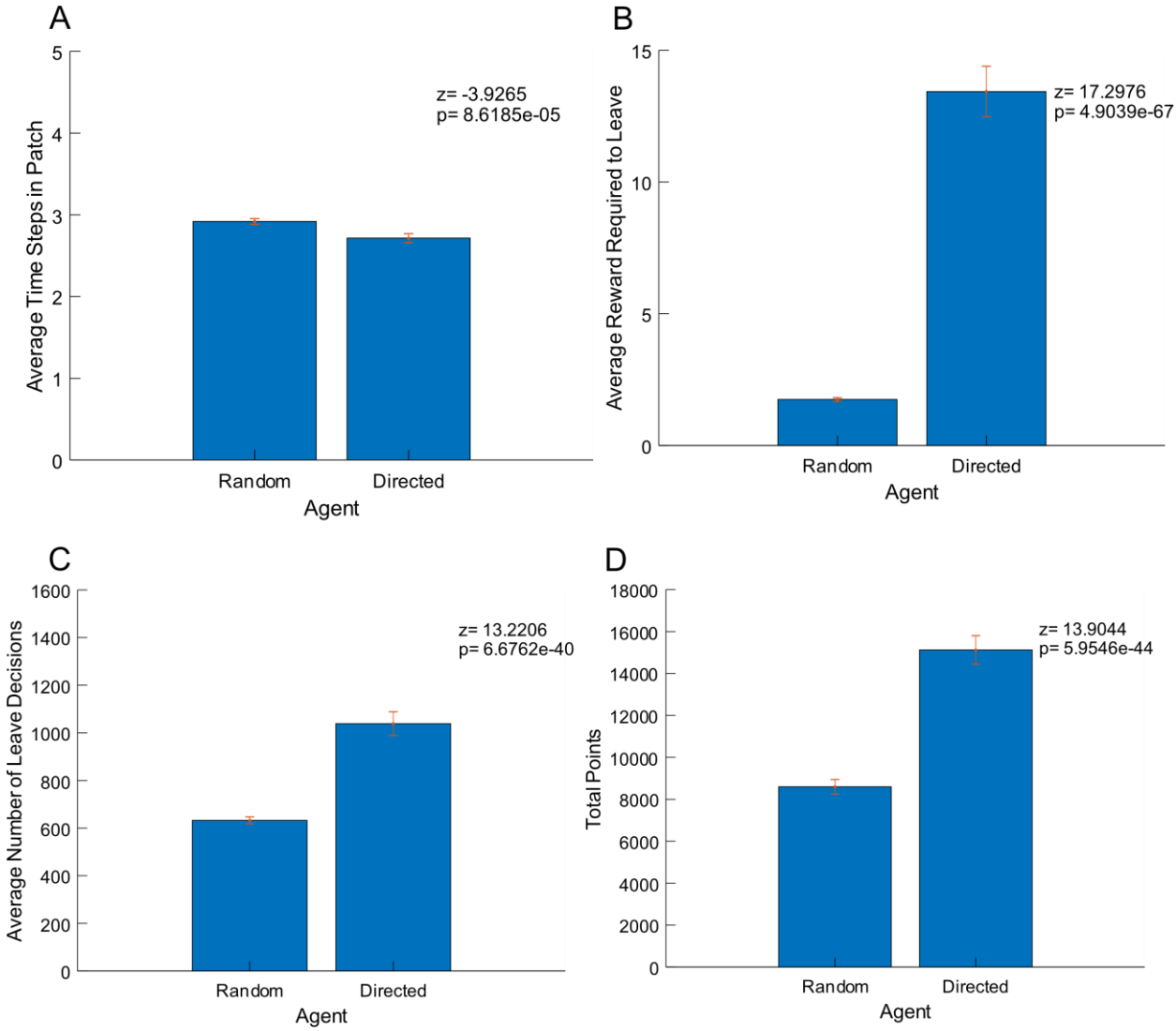


Figure A3.1. Simulation results when replenish rate is high (10) for random and directed agents, where the directed agent a global threshold for deciding when to leave the current patch. The pattern of differences between random and directed agents reported in the main text holds for directed agents using a global leaving threshold. A: Average number of time steps in each patch before leaving for random and directed agents. B: Average reward used as a threshold for leaving the current option for random and directed agents. Note that a higher value indicates a lower threshold for leaving because higher values will be met more quickly by reward from the current option, during its exponential decay. C: Number of leave decisions made within the 5000 allotted time steps, averaged separately over random and directed agents. D: Total points accrued during the 5000 time steps, averaged separately over random and directed agents. A-D: Error bars show 95% confidence intervals.

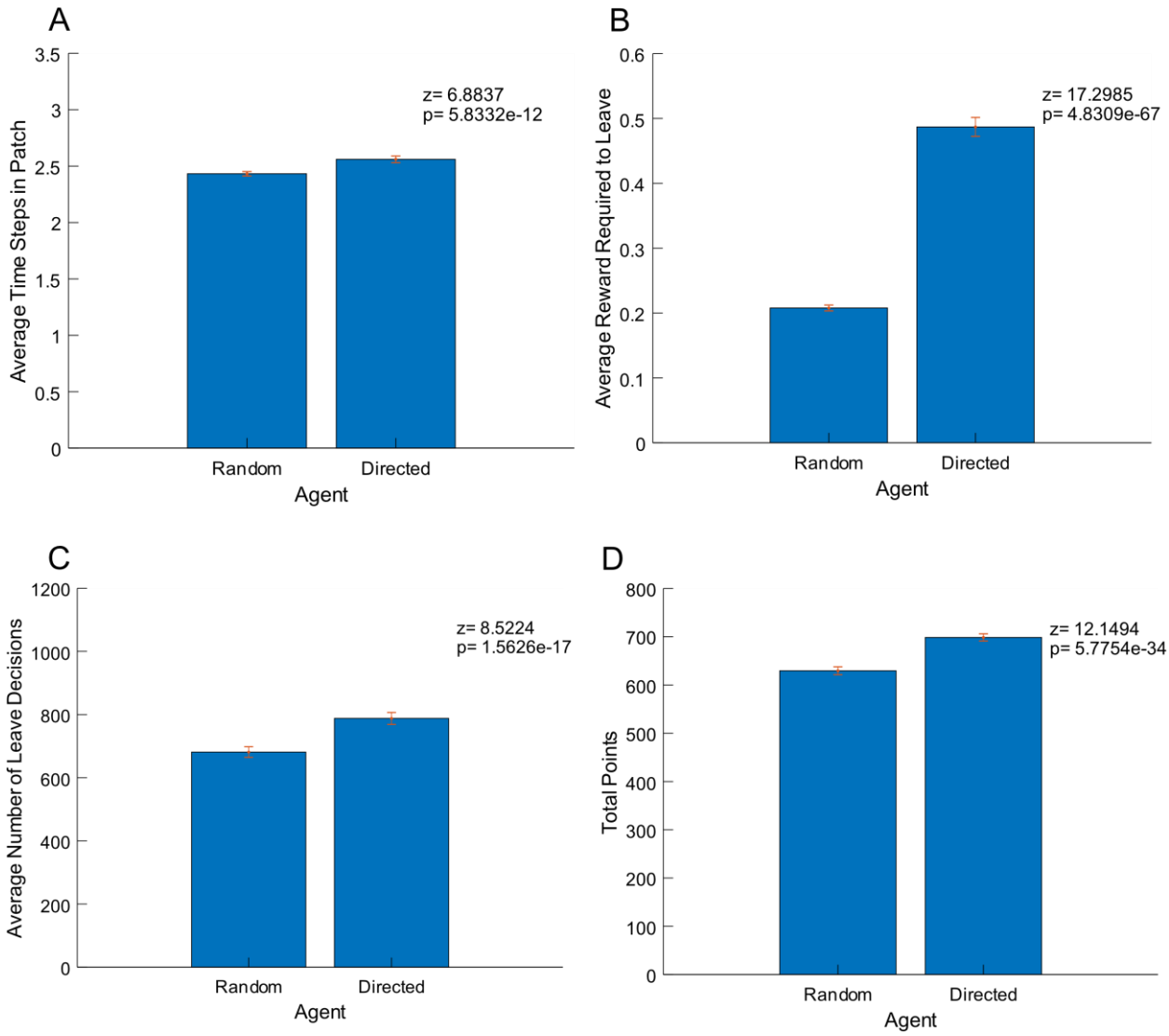


Figure A3.2. Simulation results when replenish rate is high (0.1) for random and directed agents, where the directed agent a global threshold for deciding when to leave the current patch. The pattern of differences between random and directed agents reported in the main text holds for directed agents using a global leaving threshold, with the exception that these directed agents spend longer in patches at this low replenish rate. A: Average number of time steps in each patch before leaving for random and directed agents. B: Average reward used as a threshold for leaving the current option for random and directed agents. Note that a higher value indicates a lower threshold for leaving because higher values will be met more quickly by reward from the current option, during its exponential decay. C: Number of leave decisions made within the 5000 allotted time steps, averaged separately over random and directed agents. D: Total points accrued during the 5000 time steps, averaged separately over random and directed agents. A-D: Error bars show 95% confidence intervals.

Chapter 5. Neural encoding of hidden reward information and switch decisions

Abstract

Deciding when to adjust behaviour is critical when engaged in many goal-directed activities. One way to make such decisions is to use information about the reward prospect for switching from one pattern of behaviour to another. In many real-life situations, the reward prospect for switching is not directly cued by the environment (i.e., sensory input), but must be inferred from learned reward contingencies and how these change over time. In experiment 5.1, we investigated how human participants decide when to switch to an option with changing reward value, in the absence of direct sensory cues. Thirty-two participants performed a sequential decision making task, which involved deciding between waiting and selecting a stimulus to gain its reward. Entering a wait response allowed the stimulus to increase in value towards a maximum. Critically, however, there was no change in the physical stimulus to signal changes in reward. Participants' decisions about when to switch from waiting to selecting the stimulus showed significant sensitivity to properties governing overall reward dynamics, including starting value and growth rate. Representational similarity analyses (RSA) of electroencephalographic data recorded during the task showed preliminary evidence for neural decoding of all task variables needed to compute the hidden reward prospect for selecting the stimulus. Furthermore, preliminary analyses indicated that hidden reward prospect could be decoded, beyond a linear combination of the main task variables. The extent to which participants encoded hidden reward was associated with the higher average reward per trial run. These results require verification with additional analyses. Nonetheless, they provide initial evidence that, in certain situations, hidden reward information may be used to guide decisions about maintaining and adjusting behaviour.

Introduction

Deciding when to change our behaviour is critical for effective goal pursuit. For example, when we have been working on a project for a long time, it is important to decide whether it is worth pursuing further or whether we should switch to something new.

Decisions about whether to maintain or adjust behavior have been investigated using Optimal Foraging Theory (Stephens & Krebs, 1986). Most studies of stay-switch behaviour in this area have been based on the Marginal Value Theorem (MVT), which provides a normative model for when to switch away from an option with monotonically depleting reward (Charnov, 1976). This model has been successful in accounting for switch decisions in a range of species, including mice, rats, non-human primates and humans (Barack et al., 2017; Constantino & Daw, 2015; Hayden et al., 2011; Kane et al., 2017). Corresponding neural work has indicated that dorsal anterior cingulate cortex and posterior cingulate cortex in monkeys (Barack et al., 2017; Hayden et al., 2011), as well as locus coeruleus and dorsal raphe nucleus in rodents (Kane et al., 2017; Lottem et al., 2018) might be important for helping to regulate switch decisions.

Despite their success, studies based on MVT focus on a comparison between a current option's instantaneous reward and the average reward rate for an environment. This focus has neglected the potential role that additional forms of reward information make to switch decisions, such as the starting value and growth rate of the option being switched towards. In addition, studies based on MVT tend to involve updating current and average reward values based on reward feedback. However, in real life, we often need to update reward information in the absence of direct perceptual cues.

This experiment therefore investigated how people decide when to exploit a specific alternative option, in the absence of sensory cues to indicate its changing reward value. Provided such changes in values follow a lawful function, one solution to this dilemma is to use information about an option's starting value and growth rate to estimate its *hidden* reward value over time. This is possible when the option's reward function has been learned and when starting value and growth rate can be used as inputs to the function to simulate future values or update reward values at each time step. This solution implies two possible neural mechanisms to regulate switch decisions.

The first neural mechanism would be that the brain tracks latent reward value and uses this latent estimate to decide when to switch. The plausibility of this mechanism is supported through three lines of evidence showing neural tracking of latent task information. First, activity in human orbitofrontal frontal cortex shows encoding of latent task states, which are not cued directly but must be inferred from a combination of sensory variables (Schuck & Niv, 2019). Second, neurons in human dorsal anterior cingulate cortex show encoding of latent reward trends, which can be used to regulate stay-switch decisions (Wittmann et al., 2016). Third, neurons in monkey mid-cingulate cortex encode information about the hidden reward for an option, around the time of decisions to engage with an on-going task or select the option, either to check its current reward value or to gain its reward when it has reached a certain magnitude (Stoll et al., 2016).

The second neural mechanism would be that the brain determines when to switch in advance and then tracks the distance to this critical *goal-state*. The plausibility of this mechanism is supported through two lines of evidence. First, electrophysiological recordings from rats show that firing rates in hippocampal cells scale with the distance to goal states, where reward can be obtained. Stronger coupling between firing rates and distance to goal

is also associated with better task performance (Spiers et al., 2018). Second, human hippocampus and dorsomedial PFC (dmPFC) encode proximity to goal states during virtual navigation, with the strength of encoding in dmPFC tracking the influence of goal distance on reaction times (Balaguer et al., 2016). While in the context of navigation, recent interest in the non-spatial function of brain systems involved in navigation (Behrens et al., 2018) raise the possibility that these results reflect a broader mechanism, where tracking the distance to critical behavioural states is used to help regulate goal-directed behaviour.

In this experiment, we tested how human participants decide when to switch to an alternative option as it becomes increasingly valuable, in the absence of sensory cues for changes in reward. Based on latent reward tracking accounts (Stoll et al., 2016; Wittmann et al., 2016), we hypothesized that the brain would track hidden reward information and use it to decide when to switch. We compared this with a competing mechanism (Balaguer et al., 2016; Patai et al., 2019; Spiers et al., 2018), in which the brain determines when to switch and monitors distance to this critical goal state. To test these mechanisms, we developed a sequential decision making task. Participants were presented with an option that became more valuable with each successive trial, up to a maximum. Trial-to-trial, participants decided between letting the reward for the option update (staying) or exploiting the option (switching). We applied multivariate pattern analyses to electroencephalographic data recorded during the task, to test for neural encoding of hidden reward and distance to goal information.

To summarise the main results, we found preliminary evidence for neural encoding of all task variables needed to calculate hidden reward. We also found preliminary evidence for hidden reward information that could be detected in neural recordings above and beyond a linear combination of these task variables. While distance to goal information was

also encoded, hidden reward was found to be a significantly better predictor of residual variance that was not explained by core task variables. The extent to which participants encoded hidden reward information prior to making a response was significantly correlated with average reward accrued. This association was not detected for distance to goal coding.

Method

Participants

We set a target sample of 32 participants. During recruitment five participants were excluded. One participant was excluded due to substantially lower behavioural performance than the sample distribution (more than three standard deviations below the mean percentage of total points: 89.29%). Four participants were excluded due to excessive artefacts in the EEG recording, which lead to the rejection of more than 400 trials. To replace these subjects and meet the 32 participant target, we therefore collected five more participants. The final sample were between 18 and 35 (mean age=25, 19 female). All participants reported normal or corrected-to-normal vision (including normal colour vision) and no history of neurological or psychiatric illness. Participants received £10 per hour or course credit for taking part and could earn up to £20 extra for task performance. The study was approved by the Central University Research Ethics Committee at the University of Oxford (R58489/RE001) and all participants signed informed consent before taking part.

Materials

Stimuli were presented on a 24-inch screen with a spatial resolution of 1920 x 1080 and refresh rate of 100Hz. Stimulus presentation was controlled using Psychophysics Toolbox-3 (Kleiner, Brainard & Pelli, 2007, RRID: SCR_002881) in MATLAB (RRID: SCR_001622, version R2015b). Stimuli were circles containing a coloured sector. These stimuli had approximate visual angles of 3.79° (150x150 pixels), calculated based on an approximate viewing distance of 60cm. The center point of each circle was presented at 3.16° (125 pixels) above and 3.16° to the left or right of the screen's center. F and J keys on a standard QWERTY keyboard were used to record left and right hand responses. EEG data were recorded with 61 Ag/AgCl sintered electrodes (EasyCap, Herrsching, Germany), a

NeuroScan SynAmps RT amplifier, and Curry 7 acquisition software (RRID: SCR_009546). Electrode impedance was reduced to <10 kΩ before recording. EEG data were pre-processed in EEGLAB (RRID: SCR_007292, version 14.1.1b) (Delorme & Makeig, 2004).

Experimental Design

Participants performed a sequential decision making task, which required a simple choice on each trial: maintain or adjust behaviour (Figure 1). On each run, participants were presented with a stimulus that became more valuable on each successive trial, up to a maximum of 500 points. After reaching the maximum, the stimulus value dropped with each successive trial, creating a narrow window for the stimulus to be exploited at its maximum reward value. On each trial within a run, participants decided between a stay response (waiting), which allowed the stimulus value to update for the next trial, or selecting the stimulus, to gain the reward associated with it. Critically, there was no sensory change to the stimulus to indicate the change in value. This task therefore required participants to progress through a series of partially observable (hidden) reward states and decide when the stimulus was valuable enough to switch away from the default action (waiting).

Each stimulus was a circle containing a coloured sector. The arc length of the coloured sector indicated the starting value for the stimulus (0-350 points). The colour of the sector indicated the growth rate (the increase in points that would be gained by waiting on each trial until the maximum was reached). Each trial had three phases. In the first phase, the circle stimulus was presented on the left or right of the screen for 800ms. Prior to the experiment, participants were instructed to use this phase (of the trial?) to decide whether they should wait (a stay response) or 'cash in' (a switch response). In the second phase, there was a 250ms blank delay, after which the same stimulus was presented on the left or right of the screen, along with a central fixation cross. To decouple neural activity

related to the decision and the motor response, there was a 0.5 probability the stimulus would appear on the same or opposite side of the screen from the previous phase. Participants were instructed to press the button congruent with the side of stimulus presentation to cash in (e.g. F if the stimulus was presented on the left) and to press the button incongruent with stimulus presentation to wait (e.g. F if the stimulus was presented on the right). The third phase of each trial was a feedback phase. If the participant decided to wait on that trial (a stay response), they would receive +1 point. If they decided to 'cash in' (a switch response), they would receive whatever the stimulus had accrued (0-500points). Feedback was presented for 500ms. Trials within a run continued until the participant selected the stimulus, up to a maximum of 16 trials.

The reward associated with each stimulus could grow to a maximum of 500 points. The reward was computed with the following linear function: $reward\ value = s + k * t$, where s is the starting value, k is the growth rate and t is the trial number within the run. The reward value was updated at the end of each trial. Eight different starting values were used in this study (0, 50, 100, 150, 200, 250, 300, 350). On trial runs where the starting value was 0, a coloured sector corresponding to 7.5 points was used so that participants could see the coloured sector and determine the growth rate. Five different growth rates were used in this study (35, 40, 50, 60, 80). These were selected to maximize behavioural variability, producing an even spacing in optimal switch points (15, 13, 11, 9, 7 trial steps), from a starting value of 0 points. Two colours were used per growth rate. In addition, the mapping between colours and growth rates was balanced across participants. Subjects were pseudo-randomly assigned to one of four colour-growth rate mappings, so that there were approximately equal numbers of participants for each mapping. On trials where the updated stimulus value exceeded 500 points, the stimulus value was rounded down to the 500

maximum. If participants did not select the stimulus on the trial where it reached its maximum, the reward associated with the stimulus would begin an exponential decay given by the function: $rd\ value = \max value * e^{-lambda * tm}$, where max value is a constant set at 500 points for the current experiment, lambda is a decay constant set at 0.5 and tm is the number of trials since the stimulus reached the 500 point maximum. The reward value was updated at the end of each trial. The tm count therefore included the trial at which 500 points was reached, to ensure each stimulus had just one trial at maximum reward within a run.

Participants first completed a training sequence of 40 trial runs. If participants scored below 70% or wished to do more practice, participants completed an additional 40-80 trial runs. Eight participants completed 40 practice runs, twenty-one participants completed 80 practice runs and one participant completed 120 practice runs. The main purpose of the practice was to learn the correspondence between the 10 colours and the five growth rates. To facilitate this, participants were presented with an on screen colour legend during task practice, which indicated colours for each of the growth rates. This colour legend was not presented during the main task. Before beginning the main experiment, participants performed a colour ranking task, where they had to indicate the relative growth rate for each colour (1-5). If any ranking was incorrect, the task was repeated; participants needed to score 100% on the ranking task before starting the main experiment. During EEG recording, participants performed 240 trial runs, with equal numbers of each trial type. As participants decided when to 'cash in' on each trial run, this led to different numbers of trials for each subject. On average, we recorded 1660 trials per participant (min=1412, max=1898).

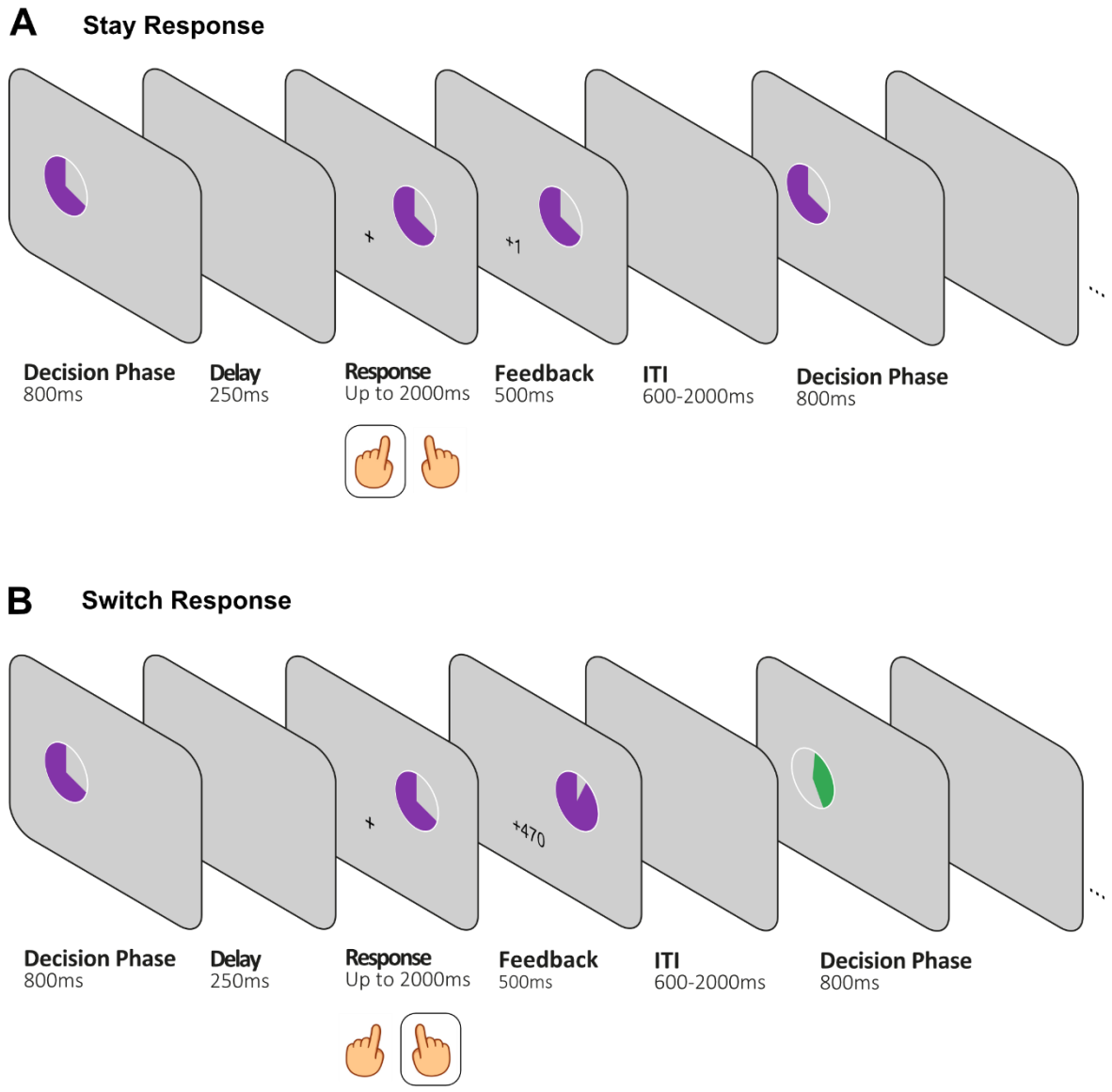


Figure 1. Task design. Participants were presented with a stimulus that had a specific starting value (colour) and growth rate (arc length). A: Participants could choose to let the stimulus value grow, by selecting the blank side of the screen during the response phase. The stimulus value was then updated by the growth rate and the participant performed another trial with the same item. Importantly, the same starting value and growth rate information were presented for each trial within a run, even though the underlying stimulus value was changing. B: Participants could choose to end the trial run, by selecting the side of the screen with the stimulus during the response phase. Participants then received the reward the stimulus had grown to and moved on to a new trial run, with a new stimulus. Trial runs lasted up to a maximum of 16 trials and item values decay exponentially if the stimulus had grown to its maximum value but participants continued to make stay responses.

EEG Pre-processing

EEG data were down-sampled from 1000 to 250Hz and filtered using 40Hz low-pass and 0.01Hz high-pass filters. For each participant, channels with excessive noise were identified by visual inspection and replaced via interpolation, using a weighted average of the surrounding electrodes. Data were then re-referenced by subtracting the mean activation across all electrodes from each individual electrode at each time point. Data were divided into epochs from -1 to +5 seconds from the first stimulus onset of each trial. Epochs containing artefacts (such as muscle activity) were rejected based on visual inspection. Epochs were baseline-corrected using the full epoch time window, which has been shown to improve the reliability of Independent Component Analysis (Groppe et al., 2009). Data were then subjected to an Independent Component Analysis. Structured noise components, such as eye blinks, were removed, resulting in the data set used for subsequent analyses.

Behavioural Analyses

To test whether switching behaviour was sensitive to task variables, we used linear regression. We first tested whether starting value and growth rate could be used to predict the number of trials waited until a switch was made. This used the following regression equation: $(1)\ Trials\ until\ switch = \beta 0\ (constant) + \beta 1 * starting\ value + \beta 2 * \frac{1}{growth\ rate} + \varepsilon\ (residual\ error)$. Note that the $\frac{1}{growth\ rate}$ in this expression arises from the fact that: $Reward = starting\ value + growth\ rate * trial\ number$. This means that the trials until an option reaches maximum reward can be computed as: $Trial\ number = \frac{maximum\ reward - starting\ value}{growth\ rate}$. Following the regression, we multiplied β2 by -1 to align the +/- sign of β1 and β2 estimates. This was done so that the direction of starting value and growth rate betas could be interpreted the same way. In exploratory

analyses, we tested whether the participants' tendency to switch earlier than optimal linear model (the action bias) was greater for slower trial runs (i.e. those that took longer to reach maximum reward). The action bias was computed by subtracting the optimal number of trials to wait until switching from the empirical number of trials waited until switching, on each trial run. The vector was then multiplied by -1 to convert the action bias to a positive scale. The following regression equation was used for this analysis: $(2)\ Action\ bias = \beta 0\ (constant) + \beta 1 * optimal\ trials\ until\ switch + \varepsilon\ (residual\ error)$. We also tested whether the amount of reward lost on each trial run could be predicted from the optimal number of trials until a switch for each condition. The reward loss was computed by subtracting 500 from the amount of reward earned on each run and multiplying the subsequent vector by -1 to convert the reward loss to a positive scale. This resulted in the following regression equation: $(3)\ Reward\ loss = \beta 0\ (constant) + \beta 1 * optimal\ trials\ until\ switch + \varepsilon\ (residual\ error)$. Omission trials were excluded from all behavioural analyses. Dependent and independent variables were z-scored prior to running regressions.

EEG Analyses

EEG analyses aimed to test for the presence of hidden reward information in the neural signal. The common features of all EEG analyses were as follows. Analyses excluded omission and switch trials. Data were baselined 250-50ms prior to trial onset and z-scored over the trial dimension. Task variables were decoded from neural activity using Pearson correlation distances, which are computed as 1 minus the Pearson correlation between condition patterns at a given time point. We initially used Mahalanobis distances for these analyses, as used throughout chapter 3. However, we found that Mahalanobis distances resulted in a bias when decoding trial position information, showing significant decoding of

trial position during the baseline period. This bias was reduced – though not eliminated - by matching trial numbers between conditions (see appendix). In contrast, we found that correlation distances did not show this bias. We therefore present correlation distance results in this chapter, based on the observation that these appear to be the most accurate for the current design. Nonetheless, the differences between distance metrics is something we are continuing to investigate, which is why we refer to results in the current chapter as preliminary. Correlation distances between conditions were regressed against the model dissimilarity matrices for different task variables, in a similar manner to the representational similarity analysis (RSA) approach used in chapter 3. Subsequent regression time courses were smoothed with a 12ms Gaussian kernel immediately prior to non-parametric cluster-based permutation testing, which was used to correct for multiple comparisons (Maris & Oostenveld, 2007; Sassenhagen & Draschkow, 2019; Spaak et al., 2017). The first 100ms was excluded from statistical testing to reduce the influence of stimulus processing.

Decoding information needed to compute hidden reward

We first tested whether information related to starting value, growth rate and trial position within a run could be decoded using RSA (Kriegeskorte et al., 2008). Trials were divided into two sets based on the colour used to indicate the growth rate. One colour for each growth rate was assigned to data split A and the other colour for each growth rate was assigned to data split B. This allowed us obtain a cross-validated measure of neural similarity, which could not be attributed to comparing trials with the same colour or from the same trial runs. We iterated through each trial position-starting value-growth rate combination and computed the mean pattern across relevant trials, separately for split A and split B data. We then computed the correlation distance between each mean in split A and each mean in split B from -0.250s to +1.5s relative to decision phase onset, resulting in a

participant-specific representational dissimilarity matrix (RDM) for each time point between -0.250s to +1.5s. Next, we constructed a set of model dissimilarity matrices to reflect expected dissimilarity structure based on different task variables. The model for starting value was computed as the difference in ordinal starting value (1= lowest starting value to 8 = highest starting value) between each condition, as ordered in the data RDM. The model for grow rate was computed as the difference in ordinal growth rate (1= lowest growth rate to 5 = highest growth rate) between each condition, as ordered in the data RDM. The trial position model was computed as the difference in trial position for each condition, as ordered in the RDM. These models were combined into a three-column matrix and z-scored across trials. The data RDM was transformed into a vector and z-scored. The models and data RDM were entered into a multiple regression at each time point, with the models as independent variables and neural condition dissimilarities as dependent variables. A constant was used to model the intercept of the linear regression equation. This procedure resulted in one regression coefficient for each model at each time point. The resulting participant betas were tested for using a one-sample t-test at each time point. The test results were corrected for multiple comparisons using the same cluster-based permutation testing procedure used in chapter 3. We interpret the time course of regression coefficients as an index of neural coding for a given task variable. Control analyses for this section additionally included models reflecting the difference in trial number used to compute the mean for each condition in split A and split B, as well as the total number of trials in a condition, summed over split A and split B.

Decoding hidden reward and distance to goal

To decode hidden reward information above and beyond a linear combination of the three task variables above, we constructed a model dissimilarity matrix for hidden reward

value. This was computed as the difference in reward value between conditions, as ordered in the data RDM. The hidden value model was transformed into a vector and used in a regression analysis that regressed the hidden reward model against residual variance, not accounted for by the starting value, growth rate or trial position coding models in the previous section. For this analysis, the hidden reward model was the independent variable and the residual variance was the dependent variable. The hidden value model was z-scored prior to running the regression. The analysis did not use a constant to model the intercept of the linear regression equation, given a constant had been used during the regression in the previous section. A similar approach was taken to assess distance to goal coding. We constructed a model dissimilarity matrix for distance to goal, which was computed as the distance between current trial position and the trial position where the stimulus was at maximum reward. The distance to goal model was transformed into a vector, z-scored and regressed against the residual variance. Due to the high correlation between hidden reward and distance to goal coding models (Spearman's Rho averaged over subjects = 0.6456), we ran separate regressions for hidden value and distance to goal against the residual variance. Participant regression coefficients for all analyses in this section were tested for significance using cluster-based permutation testing.

Relationship between hidden reward coding and task performance

We tested the relationship between hidden reward betas and task performance at two distinct phases of the trial. In one case, betas were averaged over time points in the decision phase (100-800ms from trial onset). In the second case, betas were averaged over time points in the first 350ms of the response phase (1050-1400ms from trial onset). Each vector of participant betas was then correlated with: a) the average amount of reward gained per trial run; b) the average amount of reward lost per trial run; c) the average bias

to switch earlier than the optimal linear model (the action bias). The same procedure was performed to test the relationship between distance to goal betas and task performance. The alpha level for these tests was adjusted to correct for multiple comparisons for the six tests within each neural regressor using the Bonferroni-Holm correction.

Results

Decisions to switch were influenced by starting value and growth rate

To assess participant choices, we ran a regression that tested whether the number of trials waited before making a behavioural switch could be predicted by starting value and/or growth rate. The analysis revealed that switch decisions showed significant sensitivity to both starting value (mean beta=-0.6391, *p*=8.5034e-27) and growth rate information (mean beta=-0.4667, *p*=8.6550e-20), with higher starting values and growth rate predicting earlier switches. This information is summarised in Figure 2, which shows mean number of trials until a switch response, for each starting-value growth rate combination.

Participants were biased to switch early when items were slow to reach the maximum value

In addition to the results above, participants showed a bias to switch earlier than the optimal linear model, in conditions with the lowest starting values and growth rates. This can be seen in Figure 2C. To formally test this, we ran a regression that examined whether the bias to switch early could be predicted from optimal switching times. This confirmed that participants tended to show larger bias to switch early when a condition required more trials to reach the maximum value (mean beta=0.4399, *p*=5.5034e-14). The bias to switch early on slow trials resulted in a loss in points compared to the optimal linear model. The reward loss (Figure 2D) could also be predicted from the optimal switching times (mean beta=0.2645, *p*=8.8234e-13), indicating larger reward losses for items that were slower to reach the maximum value. To summarise, our behavioural analyses indicate participant choices were influenced by the reward structure in this task. In addition, participants showed a bias to make behavioural switches earlier than the optimal model when faced with items that were slower to reach the maximum value.

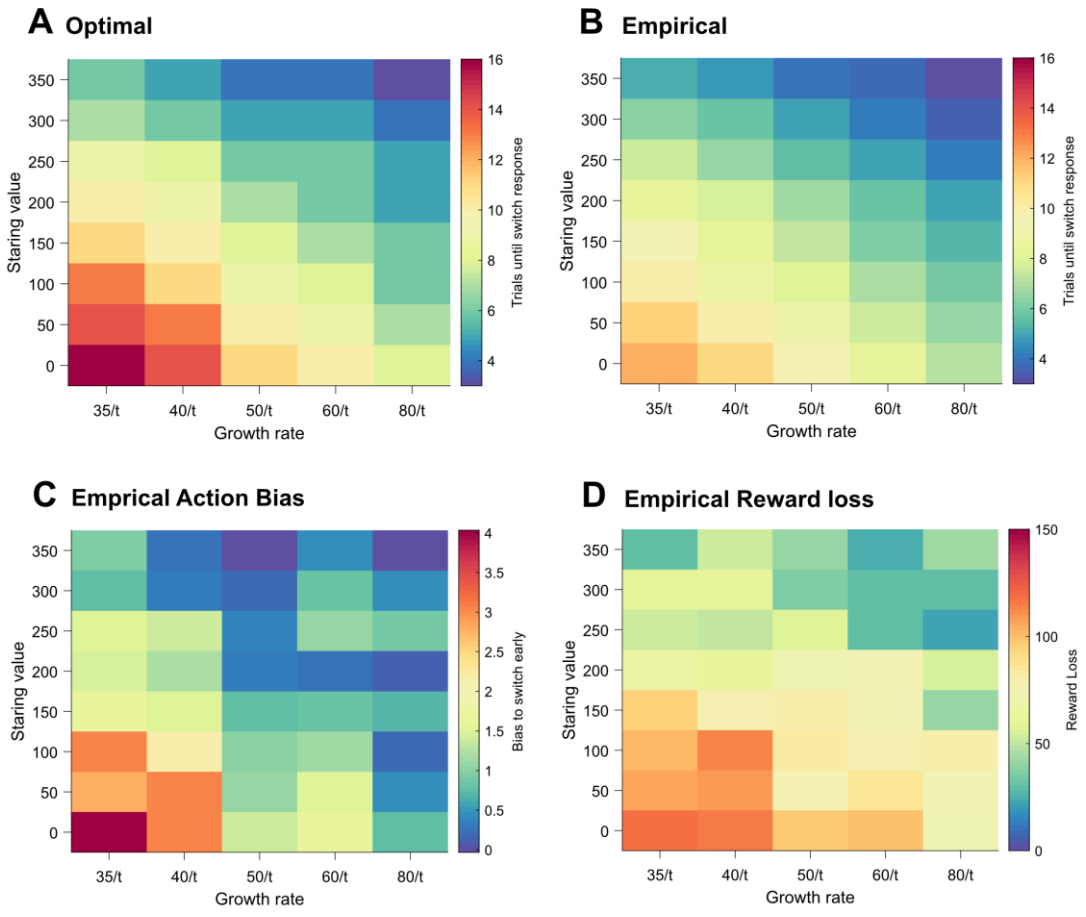


Figure 2. Summary of behavioural switching for each starting value-growth rate combination. A. The optimal number of trials until a behavioural switch should be made to get the maximum number of points. B. The mean number of trials participants wait until making a behavioural switch. C. The difference between the mean number of trials participants wait before a behavioural switch and the optimal number of trials that should be waited to maximise reward. D. The difference between the mean number of points earned and the maximum number that could be earned.

Information needed to compute hidden reward value can be decoded from neural activity

Having established that participant choices were sensitive to the reward structure of our task, we tested whether task variables were encoded in patterns of neural activity. Information related to the starting value of the choice item (the arc length) could be decoded shortly after the item was presented during the decision phase (window tested=100-1500ms, cluster window=100-348ms, *p*=3.9996e-4). The information was also

detected prior to and during the response phase (window tested=100-1500ms, second cluster window=1004-1108ms, *p*=0.0330, third cluster window = 1256-1376ms, *p*=0.0224). Encoding of abstract growth information, independent of stimulus colour, was detected during the second half of the decision phase (window tested=100-1500ms, cluster window=480-632ms, *p*=0.0076). In addition, information about the trial position within the trial run was detect throughout the trial period (window tested=100-1500ms, cluster window=100-1500ms, *p*=1.998e-4). In summary, we verified that task information needed to compute a choice item's hidden reward value was encoded in multivariate patterns of neural activity.

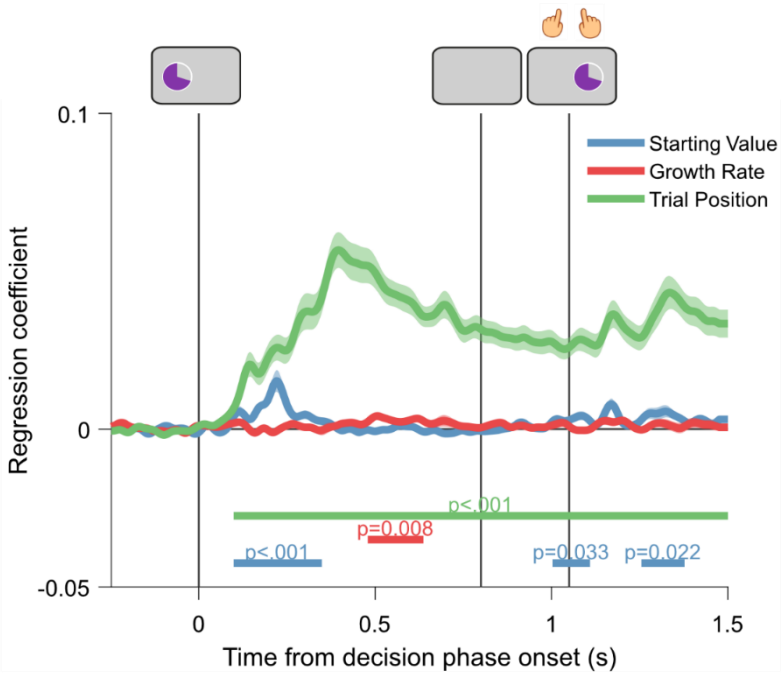


Figure 3. Time resolved regression coefficients for neural coding of starting value, growth rate and trial position related information. Vertical lines show onset of the decision phase, delay and response phase respectively. Shading around principle lines indicates standard error of mean. Cluster-corrected p-values are shown below time courses.

Hidden reward information can be decoded from neural activity

Having established the presence of all variables needed to compute hidden reward value, we tested for hidden reward information in the neural signal. We detected neural coding of hidden reward information, above and beyond a linear combination of starting value, growth rate and trial position, during the decision and response phases (window tested=100-1500ms, first cluster window=132-1020ms, p=1.9998e-4; second cluster window=1040-1260ms, p=0.0112; third cluster window=1276-1500, p=0.0072). Follow up analysis that included a nuisance regressor for the difference in trial number between conditions replicated this result (window tested=100-1500ms, first cluster window=132-1024ms, p=1.9998e-4; second cluster window=1036-1264ms, p=0.0096; third cluster window=1276-1500ms, p=0.0076), as did a similar analysis that further included a nuisance regressor for the total number of trials in each condition (window tested=100-1500ms, first cluster window=132-1024ms, p=1.998e-4; second cluster window=1036-1500ms, p=9.999e-4). To summarise, we found preliminary evidence for neural coding of hidden reward information that could not be accounted for by a linear combination of starting value, growth rate and trial position.

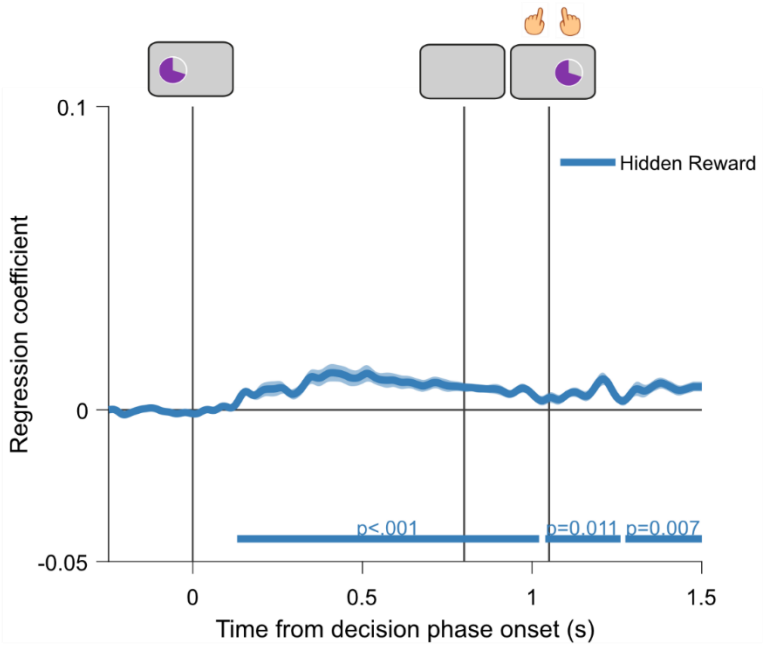


Figure 4. Time resolved regression coefficients for neural coding of hidden reward value. The figure is based on a fit to residual variance not explained by a linear combination of starting value, growth rate and trial position coding models. Vertical lines show onset of the decision phase, delay and response phase respectively. Shading around principle lines indicates standard error of mean. Cluster-corrected p-values are shown below time courses.

Hidden reward explained more residual variance than distance to goal

Thus far we have focused on neural encoding of hidden reward value. However, another neural mechanism to solve this task would be to determine the optimal switch trial and to monitor the trial distance to this critical goal state. We therefore tested how a distance to goal coding model compared to hidden reward coding when regressed against residual variance in the neural signal. Note that due to the high correlation between hidden reward and distance to goal coding models (Spearman's Rho averaged over subjects = 0.6456), we ran separate regressions against the residual variance. The results of this preliminary analysis indicated the distance to goal model was a significant predictor of residual variance (window tested=100-1500ms, first cluster window=120-292ms, p=0.0246;

second cluster window=316-1008ms, *p*=1.9998e-4; third cluster window=1084-1500ms, *p*=0.0012). However, the hidden value model was found to be a significantly better predictor than the distance to goal model (window tested=100-1500ms, first difference cluster=244-724ms, *p*=3.9996e-4; second difference cluster=740-904ms, *p*=0.0100).

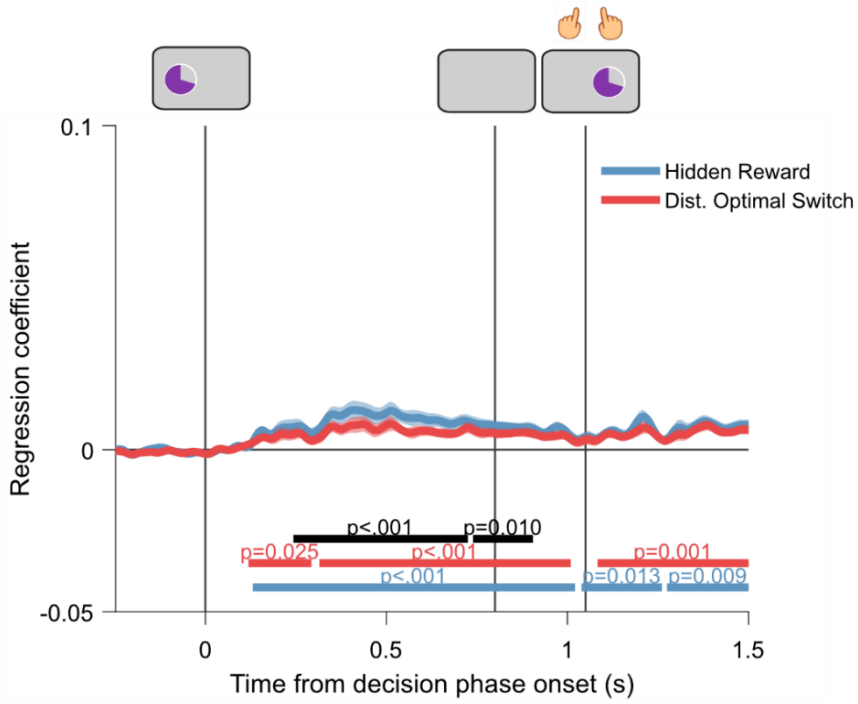


Figure 5. Time resolved regression coefficients for neural coding of hidden reward value and distance from the current trial position to the optimal switch trial for a given run using correlation distances. Vertical lines show onset of the decision phase, delay and response phase respectively. Shading around principle lines indicates standard error of mean. Cluster-corrected p-values are shown below time courses.

Neural coding of hidden reward was associated with better task performance

Having identified neural encoding hidden reward information that could not be explained by a linear combination of task variables, we tested whether this reward information tracked subject performance. Results showed that participants with higher hidden reward coding during the response phase (1050-1400ms) tended to gain more reward on average (Spearman's Rho=0.4754, Bonferroni-Holm corrected p=0.0389) and lost less reward on average, compared to the 500 points that could be gained from each choice item (Spearman's Rho=0.4710, Bonferroni-Holm corrected p=0.0389). In contrast, the analysis did not detect significant associations between hidden value coding during the decision phase and average reward (Spearman's Rho=-0.0784, Bonferroni-Holm corrected p>0.99) or average reward loss (Spearman's Rho=-0.1419, Bonferroni-Holm corrected p>0.99). Furthermore, no significant associations were detected between the bias to switch earlier than the optimal linear model and hidden reward coding, either during the decision phase (Spearman's Rho=0.0429, Bonferroni-Holm corrected p>0.99) or the response phase (Spearman's Rho=-0.0550, Bonferroni-Holm corrected p>0.99). In summary, these results indicate that neural coding of hidden reward immediately prior to making a response was associated with decision making performance.

In contrast, we did not detect significant associations between neural encoding of distance to the optimal switch trial and behavioural performance. This was the case for associations between distance to optimal switch coding during the decision phase (averaged over 100-800ms) and the average reward per run (Spearman's Rho=0.0521, Bonferroni-Holm corrected p>0.99), the average reward loss per run (Spearman's Rho=-0.0128, Bonferroni-Holm corrected p>0.99), or the bias to switch earlier than the optimal linear model (Spearman's Rho=-0.0319, Bonferroni-Holm corrected p>0.99). The same pattern of

results was seen for associations between distance to optimal switch coding during the response phase (averaged over 1050-1400ms) and the average reward per run (Spearman's Rho=0.3160, Bonferroni-Holm corrected p=0.4709), the average reward loss per run (Spearman's Rho=0.3061, Bonferroni-Holm corrected p=0.4709) and the bias to switch earlier than the optimal linear model (Spearman's Rho=-0.0869, Bonferroni-Holm corrected *p*>0.99).

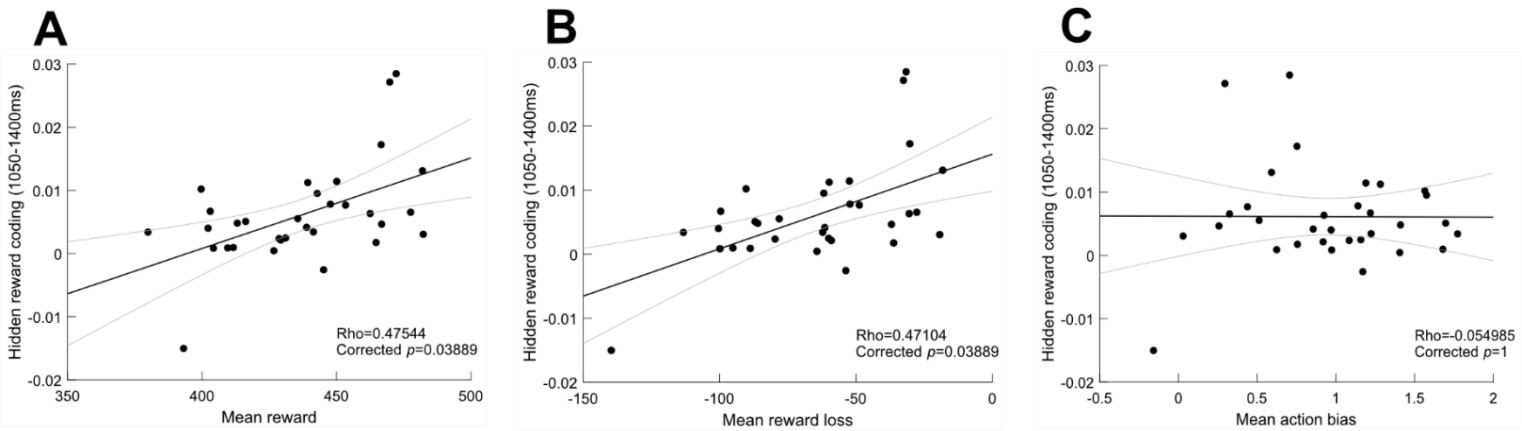


Figure 6. Spearman correlations between mean hidden reward coding during the response phase and task performance. A: Correlation between hidden reward coding and the mean reward gained per trial run. B: Correlation between hidden reward coding and mean amount of reward lost per trial run. C: Correlation between hidden reward coding and the bias to switch earlier than the optimal linear model. A-C: Black lines indicate linear fits to the data and grey lines indicate 95% confidence intervals of the fits.

Discussion

The present study investigated how people track the reward value of a specific option, in the absence of direct sensory cues to signal changes in its reward value. We hypothesized that the option's starting value and growth rate could be used as inputs to its reward function, allowing the option's hidden reward value to be tracked across the trial run. We also tested a competing mechanism, where participants pre-determined when to switch and tracked the trial distance to this critical goal state. Consistent with our latent reward coding hypothesis, we found preliminary evidence for neural encoding of all task variables needed to compute an option's hidden reward value. We also found preliminary evidence for neural coding of hidden reward information above and beyond a linear combination of these variables. Hidden value coding appeared to be a significantly better predictor of residual variance than distance to goal coding, although the two models were highly correlated. The extent to which participants encoded hidden reward was also correlated with the average amount of reward obtained per trial run as well as the average reward lost per trial run.

The present results add to a growing body of neuroscience studies that have used Optimal Foraging Theory to investigate stay-switch behaviour (Barack et al., 2017; Constantino & Daw, 2015; Hayden et al., 2011; Kane et al., 2017; Kolling et al., 2012; Lottem et al., 2018). The majority of these studies have investigated a specific form of stay-switch behaviour, involving decisions about when to switch away from an option with decreasing reward and where there is uncertainty about the next option that will be encountered after making a switch (Hall-McMaster & Luyckx, 2019). We extend this work by focusing on a very different form of stay-switch behaviour, where participants decided when to switch away from an option with stable reward, to exploit an alternative option that increases in value in

absence of direct sensory cues. One reason this is theoretically important is that many studies mentioned above have shown stay-switch decisions involve a comparison between instantaneous reward from the current option and the mean reward rate for the environment (Charnov, 1976). The present results show that other forms of stay-switch decisions, such as those where switching occurs towards a known alternative, might be based on additional choice information, such as the starting value and growth rate of an option being switched to.

Our preliminary results also add to a broader literature on neural encoding of latent task variables to guide flexible behaviour. The initial results in this chapter suggest that in addition to neural coding of latent, non-rewarded task states (Schuck & Niv, 2019) and positive or negative trends in an option's reward value (Wittmann et al., 2016), the human brain may regulate certain switch decisions by using initial sensory information to estimate an option's hidden reward prospect. Recent work has found evidence for a similar phenomenon in Rhesus monkeys (Stoll et al., 2016). Stoll and colleagues (2016) showed that around the time of decisions about whether to engage with an on-going task or check whether an alternative option had grown to a sufficient level to provide a reward bonus, there was a transient increase in midcingulate cortex activity. This increase in activity was found to carry information about the alternative option's current state in relation to the bonus threshold, consistent with the idea that hidden reward prospect information might be involved in regulating stay-switch decisions. While our preliminary results are consistent with this work, the experimental designs do have relevant differences. For instance, monkeys did not know how long it will take for the alternative option to reach the bonus threshold at the beginning of the trial run in Stoll et al. (2016). This means that checking the alternative option during a trial run carries high information value, helping the monkey

determine how close they are to receiving a bonus for switching. Thus, one possibility in Stoll et al. (2016) could be that midcingulate activity does not reflect the alternative option's hidden state per se, but uncertainty about whether it has reached threshold. In contrast, the current experiment presented participants with all information needed to compute hidden reward prospect, providing no external uncertainty about when to switch. However, it is possible that uncertainty from an internal source could have contributed to our results. For example, uncertainty in hidden reward estimates due to noise accumulation or lapses in attention might explain participants' bias to switch early on trial runs that were slow to reach maximum reward.

The present results provide many possibilities for follow up analysis. To scrutinise the robustness of our hidden value coding results, it will be important to understand why non-correlation distance metrics, such as the Mahalanobis distance, show a decoding bias during the baseline period (Figure A1). Baselining was performed on an individual trial level, so a bias could indicate the presence of position information in the signal that was changing over the baseline period. However, if this were true it would also be expected when using correlations. Differences between distance metrics will therefore be a focus of follow-up analyses. In addition, participants showed a strong bias to switch earlier than the optimal linear model for options that were slowest to reach their maximum value. The results presented in this chapter do not address the neural mechanisms behind this bias to leave early. Follow-up analyses will therefore examine whether this action bias reflects a systematic bias in the hidden value coding, wherein participants have a noisy estimate of when these slow options reach their maximum, or whether hidden value coding is accurate but participants switch early to move onto options with higher reward rates. In addition, the present results have focused primarily on hidden value coding. However, another potential

strategy to solve this task would be to encode the optimal switch trial for a given starting-value growth rate combination and monitor the proximity to this optimal switch trial throughout the trial run. We have some initial evidence that hidden value coding was a better predictor of residual variance in neural activity, not explained by starting value, growth rate and trial position (Figure 5). However, results showing significant correlations between behavioural performance and hidden reward, but not with distance to goal, should be interpreted carefully. This does not imply that distance to goal coding was unrelated to behavioural performance and additional caution should be exercised given the high correlation between hidden value and distance to goal coding models (Spearman's Rho averaged over subjects = 0.6456). It will therefore be important to develop follow up analyses that can further disambiguate these two strategies. This is especially important in light of subjective reports from some participants about using knowledge of when to switch during trial runs and for comparison with previous literature that has focused on neural monitoring of distance to goal states in the regulation of goal-directed behaviour (Balaguer et al., 2016; Patai et al., 2019; Spiers et al., 2018).

Should the present results hold up to additional scrutiny, they present many exciting avenues for future work. For example, one situation where it is useful to track hidden reward information for an alternative is to revisit that option at a later stage. When an option offers high reward but becomes depleted with exploitation, knowing its reward function could be used to decide when that option should be revisited (Hall-McMaster & Luyckx, 2019). A future experiment could therefore build on the present design by having participants exploit an option with monotonically depleting reward and decide when to switch away from it. Participants would then need to decide when the reward for the original option had replenished enough to revisit it, using information about hidden changes

in reward. The distinction between hidden reward coding and distance to goal coding could be disambiguated in a future experiment by using manipulations that influence participants' behavioural strategy. For example, changing the growth rate within a trial run may make hidden reward tracking a more adaptive strategy because the optimal switch trial would no longer be fixed across the run. Future work could also address the neural source of hidden value encoding, which is challenging to achieve with EEG. Should our preliminary results hold, there would be a strong theoretical precedent to investigate hidden reward coding in regions that have been implicated in coding latent variables to guide flexible behaviour, including dorsal anterior cingulate cortex (Wittmann et al., 2016), mid-cingulate cortex (Stoll et al., 2016) and orbitofrontal cortex (Schuck & Niv, 2019).

To conclude, previous work has investigated switch behaviour for options with monotonically depleting reward and uncertainty about what option will be encountered when a switch is made. By investigating a new form of stay-switch behaviour, the present study provides preliminary evidence for neural coding of the reward prospect for switching to a known alternative, which can be inferred from its starting value and growth rate. Encoding such hidden reward information may be used to guide decisions about maintaining or adjusting behaviour between current and alternative options, helping to coordinate efficient goal pursuit.

Chapter 5 Appendix

A1. Bias in decoding during baseline period using Mahalanobis distance metric

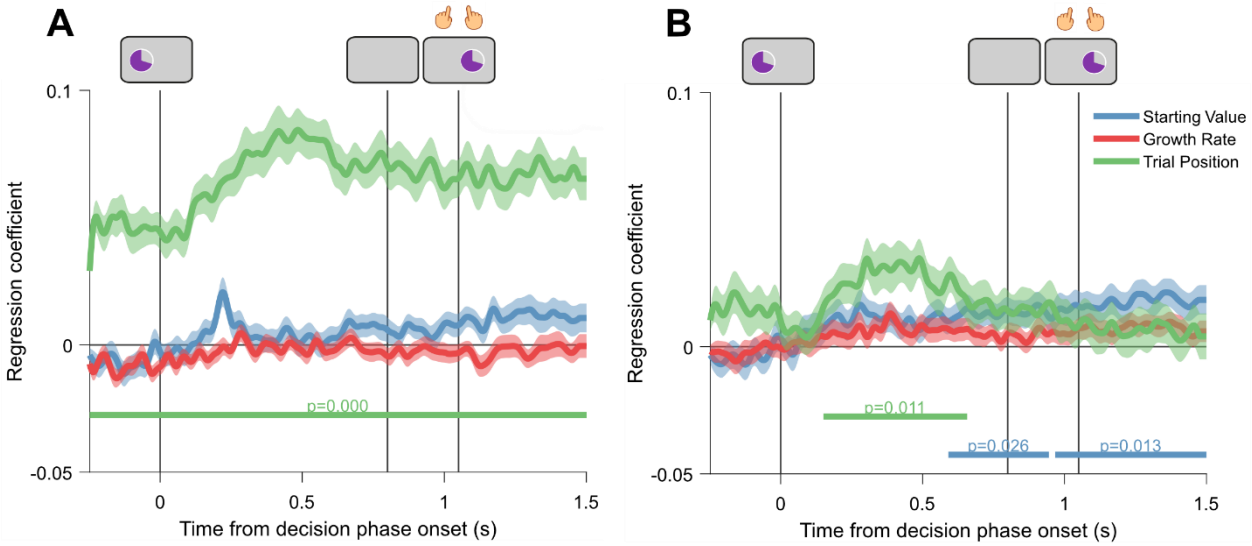


Figure A1. Time resolved regression coefficients for neural coding of starting value, growth rate and trial position using Mahalanobis distances. A: Regression time courses from an analysis that did not match trial numbers between conditions. B: Regression time courses from an analysis that matched trial numbers between conditions. A-B: Trial covariance was estimated with trials used in data split A, each of which had been demeaned with the relevant condition mean, as in chapter 3. Vertical lines show onset of the decision phase, delay and response phase respectively. Shading around principle lines indicates standard error of mean. Cluster-corrected p-values are shown below time courses (window tested=-250-1500ms).

Chapter 6. General discussion

This thesis has examined how reward prospect contributes to behavioural flexibility. Specifically, it has aimed to address two open questions in this literature. The first is how reward prospect influences task rule updating, needed to switch between different tasks. The second is how reward prospect influences decisions about when to switch from one option to another. This discussion summarises the findings from each chapter, discusses their core implications and bridges across cognitive control and decision making studies.

Summary of experimental findings

Chapter 2 examined whether reward prospect has differential effects on rule switching depending on immediate reward history. It tested the hypothesis that sequential increases in reward might promote rule switching, in comparison to high-to-high reward sequences. The results did not provide conclusive evidence for this hypothesis. Planned analyses did not detect a sequential effect of reward on switching in all three behavioural experiments. Exploratory analyses that included practice trials to boost power showed a strong trend towards a sequential effect in the final study. However, this was due to a difference in the task switch cost for low-to-high and low-to-low (as opposed to high-to-high) reward sequences.

Chapter 3 used multivariate pattern analyses of EEG data to investigate how current reward prospect influences the neural encoding of rule information during task switching. The study first replicated previous work showing that high reward prospect strengthens proactive encoding of task rules and that these changes track reward-driven performance improvements (Etzel et al., 2016). It extended this work by testing whether reward-driven increases in rule coding were more prominent on switch trials, where stronger rule coding

would be especially beneficial for reducing interference with the previous rule. Results indeed showed that high reward prospect strengthened rule coding on switch but not repeat trials. This was followed up with a behavioural study that tested whether high reward outcomes could boost active rule representations and impair rule switching on the subsequent trial. The study did not detect an influence of previous trial reward on rule switching.

Chapter 4 examined previous lab-based and real-world foraging studies, highlighting the critical influence of reward information about specific alternatives in switch decisions. This review was coupled with proof-of-concept simulations showing that agents basing switch decisions on the most valuable alternative tended to leave a current option earlier and earn more reward overall, than agents that based their switch decisions on the average experience reward rate. The impact of an option-specific switching policy on overall reward was especially prominent for environments containing items that replenished quickly after being exploited.

Chapter 5 used multivariate pattern analyses of EEG data to investigate how people decide when to switch to a specific option, in the absence of direct sensory cues to indicate changes in its reward value. Decisions about when to switch showed significant sensitivity to reward information about the option being switched towards, including its starting value and growth rate. EEG analyses showed preliminary evidence that participants tracked the hidden reward prospect for switching. The strength of hidden reward tracking was also associated with the average reward gained per trial run.

Relationship to previous studies of rule coding

Previous studies examining how rule information is coded on switch and repeat trials have shown mixed results. Waskom et al. (2014) found that specific frontoparietal brain regions showed strongest encoding of rule information during switch trials and that the strength of rule coding gradually decreased with each successive repeat. In contrast, Qiao et al. (2017) found better decoding when tasks were repeated. One possibility for these different results is that the predictable timing of rule switches and long preparation intervals allowed for strong rule encoding on switch trials in Waskom et al. (2014), whereas, unpredictable switches and shorter preparation time led to weaker encoding on switch trials in Qiao et al. (2017). While we focused on comparing high and low reward prospect within switch and repeat trials separately, we also performed a supplementary analysis that compared switch and repeat trials within each reward condition. This showed stronger rule coding on switch trials over repeat trials when reward was high, but no difference when reward was low. These results are more consistent with Waskom et al.'s (2014) findings and their proposal that competing information should have more separable neural representations when control demands are high. However, this consistency with Waskom et al. (2014) is somewhat puzzling in light of our task design being more similar to Qiao et al. (2017). The difference in rule coding when switching and repeating task rules is therefore an on-going question for future research to resolve.

Relationship to theories of task switching

The mechanisms of task switching have been a source of intense debate. Prominent theories include the active reconfiguration (updating) of rule information (see Meiran, 1996) and the encoding of rules by separate neuronal populations that interfere with each other

(see Yeung et al., 2006). The data in this thesis cannot arbitrate between these theoretical accounts. The reward-driven boost to rule coding on switch trials (chapter 3) could reflect stronger updating or reduced interference between current and previous rule codes. These possibilities are not mutually exclusive, as stronger updating could help to reduce interference (Vandierendonck et al., 2010). Our results could therefore reflect a combination of neural processes. To separate the contribution of these processes in future work, one would need to use a design that independently manipulates the interval between the context cue and the target (which determines the time for reconfiguration) and the inter-trial interval (which determines the time for previous rule representations to diminish). Alongside these theoretical accounts, it has been proposed that reward prospect could boost attention, which in turn improves rule decoding (Etzel et al., 2016). While this could be the case, our data suggest such a relationship is not straightforward. For instance, if high reward prospect boosted attention one might predict better rule encoding on both switch and repeat trials. The selective effects of reward prospect on switch trials suggest that reward prospect might enhance rule encoding through a different process (such as stronger updating), or that reward prospect's recruitment of attention interacts with cognitive control demands.

Relationship to motivation as a construct

Motivation has been defined as the orienting and invigorating effects of reward prospect on behaviour (Botvinick & Braver, 2015). However, it is difficult to see how this behavioural definition maps to the specific neural changes generated during motivated behaviour. Does orienting correspond to more sustained encoding of goal-relevant information or stronger encoding of goal-relevant information? Does orienting in fact occur

because goal-relevant information is encoded more strongly during perception and therefore seems more salient? This thesis indicates that motivation can have very specific effects on neural processing. The finding that reward prospect boosts rule coding on switch but not repeat trials in chapter 3 suggests that reward motivation strengthens neural coding of goal-relevant information as a function of cognitive control demands. This accords with research showing that motivation has more prominent effects on neural processing in situations with higher control demands, such as when targets contain incongruent information (Padmala & Pessoa, 2011). Together, these results suggest that defining the elusive construct of motivation needs to go beyond broad behavioural terms, such as orienting and invigorating. Defining motivation needs to get specific about how high reward prospect changes neural processing to deal with situations requiring heightened control over behaviour.

Separating contributions of mean reward rate and option-specific properties

One argument advanced in this thesis is that many decisions to switch away from an option might not involve a comparison between current reward rate and the average reward rate (see Barack et al., 2017; Constantino & Daw, 2015; Hayden et al., 2011; Kane et al., 2017). Instead, comparing current reward against the reward for a specific alternative might be used to guide switch decisions. The simulations in chapter 4 showed that decisions to switch based on the most rewarding alternative tended to produce earlier leaving times and higher reward, than a strategy based on mean reward. While this highlights a theoretical importance in using information tied to specific alternatives, these simulations also showed that the reward benefit for doing so diminished when options had lower replenish rates. Given the presumed computational expense of monitoring specific options,

as compared to the environment in general, it might be that the meta-control of these strategies depends on the reward and the effort to implement them. Future work could therefore focus on what factors lead people to use mean reward or specific option reward to make switch decisions. Chapter 5 advanced our argument about the use of option-specific properties, such as starting value and growth rate, in decisions to switch. However, the study examined a simple case with just one alternative option per switch decision, where option specific and mean reward information cannot be submitted to a direct comparison. Therefore, future research will be needed to explicitly separate the contributions of mean and option-specific information to switching. This could be achieved using a design similar to the simulations in chapter 4. In one condition, people could perform a standard patch-leaving task, where they do not know what option they will encounter when switching. This could be contrasted with a second condition, where people can choose a specific option to encounter after switching.

Bridging cognitive control and decision making

How do the cognitive control and decision-making studies in this thesis relate? To answer this question, let us examine previous attempts to understand the relationship between cognitive control and decision-making. One approach has focused on the difference between cued and uncued task behaviour. Such studies have primarily contrasted cued rule switching with voluntary rule switching (Fröber & Dreisbach, 2016; Wisniewski et al., 2016; Zhang et al., 2013). Note that this thesis examined cued task behaviour through cued rule switching and uncued task behaviour through both voluntary rule switching and reward-guided patch leaving. Previous research on cued and voluntary rule switching has shown that sequential increases in reward promote flexibility, reducing cued task switch

costs (Kleinsorge & Rinkenauer, 2012; Shen & Chun, 2011) and increasing the rate of decisions to voluntarily switch tasks (Fröber & Dreisbach, 2016). Neural studies examining the distinction between cued and uncued rule encoding is mixed. Some research shows similar coding of cued and uncued rules in frontoparietal regions (Wisniewski et al., 2016) and other research shows distinct coding (Zhang et al., 2013). The first experiment in chapter 2 aimed to use the distinction between cued and uncued task switching as a window into reward-guided flexibility. Unfortunately, this experiment did not detect an influence of reward on cued or uncued task switching, which could have been due to insufficient statistical power and inconsistent mappings between reward cues and reward probabilities. On reflection, such tasks might offer a well-matched comparison between cognitive control and decision processes, but the ecological validity of having a completely free choice over task rules is unclear. Aside from staving off boredom, what reason do participants have to voluntarily switch rules? The notion that uncued rule switching captures a fully internal decision process is also put into question by results showing that voluntary rule switches are significantly higher following stimulus changes than stimulus repeats (Mayr et al., 2006). To compare flexibility from cognitive control and decision-making perspectives, it therefore seems important that decisions to switch are motivated in some way. Decision tasks should not provide cues about when to switch, but they should still manipulate external factors, such as the reward prospect for switching. This means that approaches like foraging provide a more formal and ecological way to investigate internally-initiated switch behaviour than approaches like voluntary task switching.

A second proposal that links cognitive control and decision making is that cognitive control is itself a form of decision making. This view holds that people decide how much cognitive control to exert depending on the reward at stake. From an empirical perspective,

this view has been advanced from studies adapting cognitive control tasks so that people can choose between responses that involve high and low levels of control for various amounts of reward. This has revealed that people's choices to engage in high control responses scale with their reward prospect, relative to a low control option (Dixon & Christoff, 2012). When people's choices about whether to engage in a more mentally demanding task for various amounts of reward are titrated to an indifference point, where the more demanding task is chosen 50% of the time, the amount of reward needed to choose the demanding task is consistently higher than the low demand task (Westbrook et al., 2013). Such findings have given rise to view that cognitive control carries a subjective cost and that reward can offset that cost, leading people to choose more demanding responses when reward prospect is high (Kool, Shenhav, et al., 2017). Viewing cognitive control as a reward-based decision process appears consistent with the behavioural results in this thesis. Higher reward prospect in our task switching experiment (chapter 3) would be expected to pay the cost of faster and more accurate target responses. Suboptimal leaving times for the longest trial runs in our stay-switch EEG experiment (chapter 5) might be accounted for by the fact that the maximum reward prospect is fixed on every trial run. Thus, the maximum reward prospect could be sufficient to pay the control costs for exerting attention over shorter trial runs but insufficient over longer trial runs.

While the cost-benefit view above might be used to interpret our behavioural results, it is unclear whether it would predict our neural results. For these, proposals focused on signal and noise in neural systems might be more applicable. For example, Manohar and colleagues (2015) propose an optimal control model in which cognitive control and motor performance are improved through reward-driven reductions in neural noise. The reduction in noise under high reward conditions could account for stronger rule

encoding and thereby faster, more accurate behaviour seen in chapter 3. Reward-driven reductions in neural noise might also be expected to result in more precise coding of hidden reward value in the chapter 5 data. This prediction could be tested in a follow up analysis looking at whether hidden reward decoding is stronger on trial runs with a higher reward rate per trial than those with a lower reward rate per trial. Should participants make their decision about when to switch and monitor the distance to this goal state (Balaguer et al., 2016; Patai et al., 2019; Spiers et al., 2018) instead of hidden reward, then a reward-driven noise reduction model might predict stronger encoding of decision information at trial one, for trial runs with higher reward rates.

Conclusion

This thesis has demonstrated that the prospect of reward influences behavioural flexibility through increasing neural encoding of new context information. In addition, reward information about specific alternatives, such as their starting value and growth rate, have a critical influence on decisions about when to switch from one option to another. This thesis shows that reward has a wide-reaching influence on flexibility, both in situations where switching is cued from the environment and when it is internally-initiated. For progress in this area, future research should focus on whether reward influences cued and uncued behavioural flexibility through common or distinct neural mechanisms, as well as how neural systems use latent reward functions and goal progress to initiate flexible changes in behaviour.

References

Aarts, E., Roelofs, A., Franke, B., Rijpkema, M., Fernández, G., Helmich, R. C., & Cools, R. (2010). Striatal dopamine mediates the interface between motivational and cognitive control in humans: Evidence from genetic imaging. *Neuropsychopharmacology*, *35*(9), 1943–1951. https://doi.org/10.1038/npp.2010.68

Allport, D. A., Styles, E. A., & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In C. Umilta & M. Moscovitch (Eds.), *Attention and Performance IV* (pp. 421–452). MIT press.

Amer, T., Campbell, K. L., & Hasher, L. (2016). Cognitive control as a double-edged sword. *Trends in Cognitive Sciences*, *20*(12), 905–915. https://doi.org/10.1016/j.tics.2016.10.002

Aron, A. R., Monsell, S., Sahakian, B. J., & Robbins, T. W. (2004). A componential analysis of task-switching deficits associated with lesions of left and right frontal cortex. *Brain*, *127*(7), 1561–1573. https://doi.org/10.1093/brain/awh169

Arrington, C. M., & Logan, G. D. (2004). The cost of a voluntary task switch. *Psychological Science*, *15*(9), 610–615. https://doi.org/10.1111/j.0956-7976.2004.00728.x

Asaad, W. F., Rainer, G., & Miller, E. K. (2000). Task-specific neural activity in the primate prefrontal cortex. *Journal of Neurophysiology*, *84*(1), 451–459. https://doi.org/10.1152/jn.2000.84.1.451

Aston-Jones, G., & Cohen, J. D. (2005). An Integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience*, *28*(1), 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. *Trends in Cognitive Sciences*, *12*(5), 193–200. https://doi.org/10.1016/j.tics.2008.02.004

Badre, D., & Nee, D. E. (2018). Frontal cortex and the hierarchical control of behavior. *Trends in Cognitive Sciences*, *22*(2), 170–188. https://doi.org/10.1016/j.tics.2017.11.005

Balaguer, J., Spiers, H., Hassabis, D., & Summerfield, C. (2016). Neural mechanisms of hierarchical planning in a virtual subway network. *Neuron*, *90*(4), 893–903. https://doi.org/10.1016/j.neuron.2016.03.037

Barack, D. L., Chang, S. W. C., & Platt, M. L. (2017). Posterior cingulate neurons dynamically signal decisions to disengage during foraging. *Neuron*, *96*(2), 339–347. https://doi.org/10.1016/j.neuron.2017.09.048

Barceló, F., & Knight, R. T. (2002). Both random and perseverative errors underlie WCST deficits in prefrontal patients. *Neuropsychologia*, *40*(3), 349–356. https://doi.org/10.1016/S0028-3932(01)00110-5

Barraquand, F., Inchausti, P., & Bretagnolle, V. (2009). Cognitive abilities of a central place forager interact with prey spatial aggregation in their effect on intake rate. *Animal Behaviour*, *78*(2), 505–514. https://doi.org/10.1016/j.anbehav.2009.06.008

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron*, *100*(2), 490–509. https://doi.org/10.1016/j.neuron.2018.10.002

Blanchard, T. C., & Hayden, B. Y. (2014). Neurons in dorsal anterior cingulate cortex signal postdecisional variables in a foraging task. *Journal of Neuroscience*, *34*(2), 646–655. https://doi.org/10.1523/JNEUROSCI.3151-13.2014

Bogacz, R. (2007). Optimal decision-making theories: Linking neurobiology with behaviour. *Trends in Cognitive Sciences*, *11*(3), 118–125. https://doi.org/10.1016/j.tics.2006.12.006

Botvinick, M., & Braver, T. S. (2015). Motivation and cognitive control: From behavior to neural mechanism. *Annual Review of Psychology*, *66*, 83–113. https://doi.org/10.1146/annurev-psych-010814-015044

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652.

Braem, S., & Egner, T. (2018). Getting a Grip on Cognitive Flexibility. *Current Directions in Psychological Science*, *27*(6), 470–476. https://doi.org/10.1177/0963721418787475

Braver, T. S. (2012). The variable nature of cognitive control: A dual mechanisms framework. *Trends in Cognitive Sciences*, *16*(2), 106–113. https://doi.org/10.1016/j.tics.2011.12.010

Braver, T. S., & Cohen, J. D. (2000). On the control of control: The role of dopamine in regulating prefrontal function and working memory. In S Monsell & J. Driver (Eds.), *Attention and Performance XVIII* (pp. 712–737). MIT Press.

Buckley, M. J., Mansouri, F. A., Hoda, H., Mahboubi, M., Browning, P. G. F., Kwok, S. C., Phillips, A., & Tanaka, K. (2009). Dissociable components of rule-guided behaviour depend on distinct medial and prefrontal regions. *Science*, *325*(July), 52–58.

Bunge, S. A. (2004). How we use rules to select actions: A review of evidence from cognitive neuroscience. *Cognitive, Affective and Behavioral Neuroscience*, *4*(4), 564–579. https://doi.org/10.3758/CABN.4.4.564

Burgess, G. C., & Braver, T. S. (2010). Neural mechanisms of interference control in working memory: Effects of interference expectancy and fluid intelligence. *PLoS ONE*, *5*(9), 1–11. https://doi.org/10.1371/journal.pone.0012861

Chapman, C. A., Chapman, L. J., & McLaughlin, R. L. (1989). Multiple central place foraging by spider monkeys: Travel consequences of using many sleeping sites. *Oecologia*, *79*(4), 506–511. https://doi.org/10.1007/BF00378668

Charnov, E. L. (1976). Optimal foraging, the Marginal Value Theorem. *Theoretical Population Biology*, *9*, 129–136.

Chiew, K. S., & Braver, T. S. (2016). Reward favors the prepared: Incentive and task-informative cues interact to enhance attentional control. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(1), 52–66. https://doi.org/10.1037/xhp0000129

Cohen, M. X., Elger, C. E., & Ranganath, C. (2007). Reward expectation modulates feedback-related negativity and EEG spectra. *NeuroImage*, *35*(2), 968–978. https://doi.org/10.1016/j.neuroimage.2006.11.056

Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, *35*(7), 1024–1035. https://doi.org/10.1111/j.1460-9568.2011.07980.x.How

Constantino, S. M., & Daw, N. D. (2015). Learning the opportunity cost of time in a patch-foraging task. *Cognitive, Affective & Behavioral Neuroscience*, *15*(4), 837–853. https://doi.org/10.3758/s13415-015-0350-y

Cools, R., & D'Esposito, M. (2011). Inverted-U-shaped dopamine actions on human working memory and cognitive control. *Biological Psychiatry*, *69*(12), e113–e125. https://doi.org/10.1016/j.biopsych.2011.03.028

Cubillo, A., Makwana, A. B., & Hare, T. A. (2019). Differential modulation of cognitive control networks by monetary reward and punishment. *Social Cognitive and Affective Neuroscience*, *14*(3), 305–317. https://doi.org/10.1093/scan/nsz006

D'Ardenne, K., Eshel, N., Luka, J., Lenartowicz, A., Nystrom, L. E., & Cohen, J. D. (2012). Role of prefrontal cortex and the midbrain dopamine system in working memory updating. *Proceedings of the National Academy of Sciences*, *109*(49), 19900–19909. https://doi.org/10.1073/pnas.1116727109

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

de Baene, W., & Brass, M. (2011). Cue-switch effects do not rely on the same neural systems as task-switch effects. *Cognitive, Affective and Behavioral Neuroscience*, *11*(4), 600–607. https://doi.org/10.3758/s13415-011-0055-9

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

D'Esposito, M., Postle, B. R., Jonides, J., & Smith, E. E. (1999). The neural substrate and temporal dynamics of interference effects in working memory as revealed by event-related functional MRI. *Proceedings of the National Academy of Sciences*, *96*(13), 7514–7519. https://doi.org/10.1073/pnas.96.13.7514

Dixon, M. L., & Christoff, K. (2012). The decision to engage cognitive control is driven by expected reward-value: Neural and behavioral evidence. *PLoS ONE*, *7*(12). https://doi.org/10.1371/journal.pone.0051637

Dutilh, G., Vandekerckhove, J., Forstmann, B. U., Keuleers, E., Brysbaert, M., & Wagenmakers, E. J. (2012). Testing theories of post-error slowing. *Attention, Perception, and Psychophysics*, *74*(2), 454–465. https://doi.org/10.3758/s13414-011-0243-2

Ebitz, R. B., & Hayden, B. Y. (2016). Dorsal anterior cingulate: A Rorschach test for cognitive neuroscience. *Nature Neuroscience*, *19*(10), 1278–1279. https://doi.org/10.1038/nn.4387

Engelmann, J. B., Damaraju, E., Padmala, S., & Pessoa, L. (2009). Combined effects of attention and motivation on visual task performance: Transient and sustained motivational effects. *Frontiers in Human Neuroscience*, *3*(MAR), 1–17. https://doi.org/10.3389/neuro.09.004.2009

Etzel, J. A., Cole, M. W., Zacks, J. M., Kay, K. N., & Braver, T. S. (2016). Reward motivation enhances task coding in frontoparietal cortex. *Cerebral Cortex*, *26*(4), 1647–1659. https://doi.org/10.1093/cercor/bhu327

Fagan, W. F., Lewis, M. A., Auger-Méthé, M., Avgar, T., Benhamou, S., Breed, G., Ladage, L., Schlägel, U. E., Tang, W. W., Papastamatiou, Y. P., Forester, J., & Mueller, T. (2013). Spatial memory and animal movement. *Ecology Letters*, *16*(10), 1316–1329. https://doi.org/10.1111/ele.12165

Foerde, K., & Shohamy, D. (2011). Feedback timing modulates brain systems for learning in humans. *Journal of Neuroscience*, *31*(37), 13157–13167. https://doi.org/10.1523/JNEUROSCI.2701-11.2011

Frankenhuis, W. E., Panchanathan, K., & Barto, A. G. (2018). Enriching behavioral ecology with reinforcement learning methods. *Behavioural Processes*. https://doi.org/10.1016/j.beproc.2018.01.008

Fröber, K., & Dreisbach, G. (2016). How sequential changes in reward magnitude modulate cognitive flexibility: Evidence from voluntary task switching. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(2), 285–295. https://doi.org/10.1037/xlm0000166

Fröber, K., Pfister, R., & Dreisbach, G. (2019). Increasing reward prospect promotes cognitive flexibility: Direct evidence from voluntary task switching with double registration. *Quarterly Journal of Experimental Psychology*, *72*(8), 1926–1944. https://doi.org/10.1177/1747021818819449

Garber, P. A. (1987). Foraging strategies among living primates. *Annual Review of Anthropology*, *16*, 339–364. https://doi.org/10.1146/annurev.anthro.16.1.339

Goschke, T., & Bolte, A. (2014). Emotional modulation of control dilemmas: The role of positive affect, reward, and dopamine in cognitive stability and flexibility. *Neuropsychologia*, *62*, 403–423. https://doi.org/10.1016/j.neuropsychologia.2014.07.015

Groppe, D. M., Makeig, S., & Kutas, M. (2009). Identifying reliable independent components via split-half comparisons. *NeuroImage*, *45*(4), 1199–1211. https://doi.org/10.1016/j.neuroimage.2008.12.038

Hall-McMaster, S., & Luyckx, F. (2019). Revisiting foraging approaches in neuroscience. *Cognitive, Affective and Behavioral Neuroscience*, *19*(2), 225–230. https://doi.org/10.3758/s13415-018-00682-z

Hall-McMaster, S., Millar, J., Ruan, M., & Ward, R. D. (2017). Medial orbitofrontal cortex modulates associative learning between environmental cues and reward probability. *Behavioral Neuroscience*, *131*(1), 1–10. https://doi.org/10.1037/bne0000178

Haxby, J. v., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annual Review of Neuroscience*, *37*(1), 435–456. https://doi.org/10.1146/annurev-neuro-062012-170325

Hayden, B. Y. (2018). Economic choice: the foraging perspective. *Current Opinion in Behavioral Sciences*, *24*, 1–6. https://doi.org/10.1016/j.cobeha.2017.12.002

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

Hayden, B. Y., & Walton, M. E. (2014). Neuroscience of foraging. *Frontiers in Neuroscience*, *8*, 1–3. http://www.journals.uchicago.edu/doi/10.1086/284489

Hefer, C., & Dreisbach, G. (2017). How performance-contingent reward prospect modulates cognitive control: Increased cue maintenance at the cost of decreased flexibility. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(10), 1643–1658. https://doi.org/http://dx.doi.org/10.1037/xlm0000397

Heilbronner, S. R., & Hayden, B. Y. (2016). Dorsal anterior cingulate cortex: A bottom-up view. *Annual Review of Neuroscience*, *39*(1), 149–170. https://doi.org/10.1146/annurev-neuro-070815-013952

Hickey, C., Chelazzi, L., & Theeuwes, J. (2010). Reward changes salience in human vision via the anterior cingulate. *Journal of Neuroscience*, *30*(33), 11096–11103. https://doi.org/10.1523/JNEUROSCI.1026-10.2010

Hickey, C., & Peelen, M. v. (2015). Neural mechanisms of incentive salience in naturalistic human vision. *Neuron*, *85*(3), 512–518. https://doi.org/10.1016/j.neuron.2014.12.049

Hippmann, B., Kuhlemann, I., Bäumer, T., Bahlmann, J., Münte, T. F., & Jessen, S. (2019). Boosting the effect of reward on cognitive control using TMS over the left IFJ. *Neuropsychologia*, *125*(February), 109–115. https://doi.org/10.1016/j.neuropsychologia.2019.01.016

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychol. Rev*, *109*, 679–709. https://doi.org/10.1037/0033-295X.109.4.679

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., Nystrom, L., Mars, R. B., Coles, M. G. H., & Cohen, J. D. (2004). Dorsal anterior cingulate cortex shows fMRI response to internal and external error signals. *Nature Neuroscience*, *7*(5), 497–498. https://doi.org/10.1038/nn1238

Jacob, S. N., Ott, T., & Nieder, A. (2013). Dopamine regulates two classes of primate prefrontal neurons that represent sensory signals. *Journal of Neuroscience*, *33*(34), 13724–13734. https://doi.org/10.1523/JNEUROSCI.0210-13.2013

Janmaat, K. R. L., Ban, S. D., & Boesch, C. (2013). Chimpanzees use long-term spatial memory to monitor large fruit trees and remember feeding experiences across seasons. *Animal Behaviour*, *86*(6), 1183–1205. https://doi.org/10.1016/j.anbehav.2013.09.021

Jimura, K., Locke, H. S., & Braver, T. S. (2010). Prefrontal cortex mediation of cognitive enhancement in rewarding motivational contexts. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(19), 8871–8876. https://doi.org/10.1073/pnas.1002007107

Joshi, S., Li, Y., Kalwani, R. M., & Gold, J. I. (2016). Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron*, *89*(1), 221–234. https://doi.org/10.1016/j.neuron.2015.11.028

Jurczyk, V., Fröber, K., & Dreisbach, G. (2018). Increasing reward prospect motivates switching to the more difficult task. *Motivation Science*, *5*(4), 295–313. https://doi.org/10.1037/mot0000119

Kane, G. A., Vazey, E. M., Wilson, R. C., Shenhav, A., Daw, N. D., Aston-Jones, G., & Cohen, J. D. (2017). Increased locus coeruleus tonic activity causes disengagement from a patch-foraging task. *Cognitive, Affective and Behavioral Neuroscience*, *17*(6), 1073–1083. https://doi.org/10.3758/s13415-017-0531-y

Kleinsorge, T., & Rinkenauer, G. (2012). Effects of monetary incentives on task switching. *Experimental Psychology*, *59*(4), 216–226. https://doi.org/10.1027/1618-3169/a000146

Kolling, N., & Akam, T. (2017). (Reinforcement?) Learning to forage optimally. *Current Opinion in Neurobiology*, *46*, 162–169. https://doi.org/10.1016/j.conb.2017.08.008

Kolling, N., Behrens, T. E. J., Mars, R. B., & Rushworth, M. F. S. (2012). Neural mechanisms of foraging. *Science*, *335*(6077), 95–98. https://doi.org/10.1126/science.1216930

Kolling, N., Behrens, T. E. J., Wittmann, M. K., & Rushworth, M. F. S. (2016). Multiple signals in anterior cingulate cortex. *Current Opinion in Neurobiology*, *37*, 36–43. https://doi.org/10.1016/j.conb.2015.12.007

Kolling, N., Wittmann, M. K., Behrens, T. E. J., Boorman, E. D., Mars, R. B., & Rushworth, M. F. S. (2016). Value, search, persistence and model updating in anterior cingulate cortex. *Nature Neuroscience*, *19*(10), 1280–1285. https://doi.org/10.1038/nn.4382

Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological Science*, *28*(9), 1321–1333. https://doi.org/10.1177/0956797617708288

Kool, W., Shenhav, A., & Botvinick, M. M. (2017). Cognitive control as cost-benefit decision making. In T. Egner (Ed.), *The Wiley Handbook of Cognitive Control* (pp. 167–189). John Wiley & Sons.

Kostandyan, M., Bombeke, K., Carsten, T., Krebs, R. M., Notebaert, W., & Boehler, C. N. (2019). Differential effects of sustained and transient effort triggered by reward – A combined EEG and pupillometry study. *Neuropsychologia*, *123*(April 2018), 116–130. https://doi.org/10.1016/j.neuropsychologia.2018.04.032

Kouneiher, F., Charron, S., & Koechlin, E. (2009). Motivation and cognitive control in the human prefrontal cortex. *Nature Neuroscience*, *12*(7), 939–945. https://doi.org/10.1038/nn.2321

Krebs, R. M., Boehler, C. N., Roberts, K. C., Song, A. W., & Woldorff, M. G. (2012). The involvement of the dopaminergic midbrain and cortico-striatal-thalamic circuits in the integration of reward prospect and attentional task demands. *Cerebral Cortex*, *22*(3), 607–615. https://doi.org/10.1093/cercor/bhr134

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, *103*(10), 3863–3868. https://doi.org/10.1073/pnas.0600244103

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 1–28. https://doi.org/10.3389/neuro.06.004.2008

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*, 365–411.

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, *81*(3), 687–699. https://doi.org/10.1016/j.neuron.2013.11.028

Leotti, L. A., & Wager, T. D. (2010). Motivational influences on response inhibition measures. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(2), 430–447. https://doi.org/10.1037/a0016802

Locke, H. S., & Braver, T. S. (2008). Motivational influences on cognitive control: Behavior, brain activation, and individual differences. *Cognitive, Affective and Behavioral Neuroscience*, *8*(1), 99–112. https://doi.org/10.3758/CABN.8.1.99

Logan, G. D., & Bundesen, C. (2003). Clever homunculus: Is there an endogenous act of control in the explicit task-cuing procedure? *Journal of Experimental Psychology: Human Perception and Performance*, *29*(3), 575–599. https://doi.org/10.1037/0096-1523.29.3.575

Lottem, E., Banerjee, D., Vertechi, P., Sarra, D., Lohuis, M. O., & Mainen, Z. F. (2018). Activation of serotonin neurons promotes active persistence in a probabilistic foraging task. *Nature Communications*, *9*(1), 1–12. https://doi.org/10.1038/s41467-018-03438-y

Manohar, S. G., Chong, T. T. J., Apps, M. A. J., Batla, A., Stamelou, M., Jarman, P. R., Bhatia, K. P., & Husain, M. (2015). Reward pays the cost of noise reduction in motor and cognitive control. *Current Biology*, *25*(13), 1707–1716. https://doi.org/10.1016/j.cub.2015.05.038

Manohar, S. G., Finzi, R. D., Drew, D., & Husain, M. (2017). Distinct motivational effects of contingent and noncontingent rewards. *Psychological Science*, *28*(7), 1016–1026. https://doi.org/10.1177/0956797617693326

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Mark Williams, S., & Goldman-Rakic, P. S. (1998). Widespread origin of the primate mesofrontal dopamine system. *Cerebral Cortex*, *8*(4), 321–345. https://doi.org/10.1093/cercor/8.4.321

Matsumoto, M., & Takada, M. (2013). Distinct representations of cognitive and motivational signals in midbrain dopamine neurons. *Neuron*, *79*(5), 1011–1024. https://doi.org/10.1016/j.neuron.2013.07.002

Mayr, U., Diedrichsen, J., Ivry, R., & Keele, S. W. (2006). Dissociating task-set selection from task-set inhibition in the prefrontal cortex. *Journal of Cognitive Neuroscience*, *18*(1), 14–21. https://doi.org/10.1162/089892906775250085

McNamara, J. (1982). Optimal patch use in a stochastic environment. *Theoretical Population Biology*, *21*(2), 269–288. https://doi.org/10.1016/0040-5809(82)90018-1

Meiran, N. (1996). Reconfiguration of processing mode prior to task performance. *Journal of Experimental Psychology: Learning Memory and Cognition*, *22*(6), 1423–1442. https://doi.org/10.1037/0278-7393.22.6.1423

Menzel, R. (1999). Memory dynamics in the honeybee. *Journal of Comparative Physiology - A Sensory, Neural, and Behavioral Physiology*, *185*(4), 323–340. https://doi.org/10.1007/s003590050392

Merkle, J. A., Fortin, D., & Morales, J. M. (2014). A memory-based foraging tactic reveals an adaptive mechanism for restricted space use. *Ecology Letters*, *17*(8), 924–931. https://doi.org/10.1111/ele.12294

Merkle, J. A., Potts, J. R., & Fortin, D. (2017). Energy benefits and emergent space use patterns of an empirically parameterized model of memory-based patch selection. *Oikos*, *126*(2), 185–195. https://doi.org/10.1111/oik.03356

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*(1), 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

Milner, B. (1982). Some cognitive effects of frontal-lobe lesions in man. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *298*(1089), 211–226. https://doi.org/10.1098/rstb.1982.0083

Mobbs, D., Trimmer, P. C., Blumstein, D. T., & Dayan, P. (2018). Foraging for foundations in decision neuroscience: Insights from ethology. *Nature Reviews Neuroscience*, *19*, 419–427. https://doi.org/10.1038/s41583-018-0010-7

Monsell, Stephen. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. https://doi.org/10.1016/S1364-6613(03)00028-7

Notebaert, W., Houtman, F., Opstal, F. van, Gevers, W., Fias, W., & Verguts, T. (2009). Post-error slowing: An orienting account. *Cognition*, *111*(2), 275–279. https://doi.org/10.1016/j.cognition.2009.02.002

Oaten, A. (1977). Optimal foraging in patches: A case for stochasticity. *Theoretical Population Biology*, *12*(3), 263–285. https://doi.org/10.1016/0040-5809(77)90046-6

O'Reilly, J. X., Schüffelgen, U., Cuell, S. F., Behrens, T. E. J., Mars, R. B., & Rushworth, M. F. S. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proceedings of the National Academy of Sciences*, *110*(38). https://doi.org/10.1073/pnas.1305373110

Ott, T., Jacob, S. N., & Nieder, A. (2014). Dopamine receptors differentially enhance rule coding in primate prefrontal cortex neurons. *Neuron*, *84*(6), 1317–1328. https://doi.org/10.1016/j.neuron.2014.11.012

Ott, T., & Nieder, A. (2019). Dopamine and cognitive control in prefrontal cortex. *Trends in Cognitive Sciences*, *23*(3), 213–234. https://doi.org/10.1016/j.tics.2018.12.006

Padmala, S., & Pessoa, L. (2011). Reward reduces conflict by enhancing attentional control and biasing visual cortical processing. *Journal of Cognitive Neuroscience*, *23*(11), 3419–3432. https://doi.org/10.1162/jocn_a_00011

Parro, C., Dixon, M. L., & Christoff, K. (2018). The neural basis of motivational influences on cognitive control. *Human Brain Mapping*, *39*(12), 5097–5111. https://doi.org/10.1002/hbm.24348

Passingham, R. E. (1985). Memory of monkeys (Macaco mulatta) with lesions in prefrontal cortex. *Behavioral Neuroscience*, *99*(1), 3–21.

Patai, E. Z., Javadi, A. H., Ozubko, J. D., O'Callaghan, A., Ji, S., Robin, J., Grady, C., Winocur, G., Rosenbaum, R. S., Moscovitch, M., & Spiers, H. J. (2019). Hippocampal and retrosplenial goal distance coding after long-term consolidation of a real-world environment. *Cerebral Cortex*, *29*(6), 2748–2758. https://doi.org/10.1093/cercor/bhz044

Pearson, J. M., Watson, K. K., & Platt, M. L. (2014). Decision making: The neuroethological turn. *Neuron*, *82*(5), 950–965. https://doi.org/10.1016/j.neuron.2014.04.037

Pessoa, L. (2017). Cognitive-motivational interactions: Beyond boxes-and-arrows models of the mind-brain. *Motivation Science*, *3*(3), 287–303. https://doi.org/10.1037/mot0000074

Pessoa, L., & Engelmann, J. B. (2010). Embedding reward signals into perception and cognition. *Frontiers in Neuroscience*, *4*, 1:8. https://doi.org/10.3389/fnins.2010.00017

Poh, J. H., Massar, S. A. A., Jamaluddin, S. A., & Chee, M. W. L. (2019). Reward supports flexible orienting of attention to category information and influences subsequent memory. *Psychonomic Bulletin and Review*, 559–568. https://doi.org/10.3758/s13423-019-01595-9

Poljac, E., & Yeung, N. (2014). Dissociable neural correlates of intention and action preparation in voluntary task switching. *Cerebral Cortex*, *24*(2), 465–478. https://doi.org/10.1093/cercor/bhs326

Possingham, H. P., & Houston, A. I. (1990). Optimal patch use by a territorial forager. *Journal of Theoretical Biology*, *145*(3), 343–353. https://doi.org/10.1016/S0022-5193(05)80114-6

Qiao, L., Zhang, L., Chen, A., & Egner, T. (2017). Dynamic trial-by-trial re-coding of task-set representations in frontoparietal cortex mediates behavioral flexibility. *The Journal of Neuroscience*, *37*(45), 0935–17. https://doi.org/10.1523/JNEUROSCI.0935-17.2017

Ratcliff. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*(3), 510–532.

Richter, F. R., & Yeung, N. (2016). ERP correlates of encoding success and encoding selectivity in attention switching. *PLoS ONE*, *11*(12), 1–23. https://doi.org/10.1371/journal.pone.0167396

Riotte-Lambert, L., Benhamou, S., & Chamaillé-Jammes, S. (2015). How memory-based movement leads to nonterritorial spatial segregation. *The American Naturalist*, *185*(4), E103–E116. https://doi.org/10.1086/680009

Rosenberg, D. K., & McKelvey, K. S. (1999). Habitat selection for central-place foraging animals. *The Journal of Wildlife Management*, *63*(3), 1028–1038. https://www.jstor.org/stable/pdf/3802818.pdf

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

Sassenhagen, J., & Draschkow, D. (2019). Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology*, *56*(6), 1–8. https://doi.org/10.1111/psyp.13335

Sayers, K., & Menzel, C. R. (2012). Memory and foraging theory: Chimpanzee utilization of optimality heuristics in the rank-order recovery of hidden foods. *Animal Behaviour*, *84*(4), 795–803. https://doi.org/10.1016/j.anbehav.2012.06.034

Schuck, N. W., & Niv, Y. (2019). Sequential replay of nonspatial task states in the human hippocampus. *Science*, *364*(6447). https://doi.org/10.1126/science.aaw5181

Serences, J. T. (2008). Value-based modulations in human visual cortex. *Neuron*, *60*(6), 1169–1181. https://doi.org/10.1016/j.neuron.2008.10.051

Serences, J. T., & Saproo, S. (2010). Population response profiles in early visual cortex are biased in favor of more valuable stimuli. *Journal of Neurophysiology*, *104*(1), 76–87. https://doi.org/10.1152/jn.01090.2009

Servan-Schreiber, D., Cohen, J. D., & Steingard, S. (1996). Schizophrenic deficits in the processing of context: A test of a theoretical model. *Archives of General Psychiatry*, *53*(12), 1105–1112. https://doi.org/10.1017/CBO9781107415324.004

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–240. https://doi.org/10.1016/j.neuron.2013.07.007

Shenhav, A., Cohen, J. D., & Botvinick, M. M. (2016). Dorsal anterior cingulate cortex and the value of control. *Nature Neuroscience*, *19*(10), 1280–1285. https://doi.org/10.1038/nn.4382

Shenhav, A., Straccia, M. A., Botvinick, M. M., & Cohen, J. D. (2016). Dorsal anterior cingulate and ventromedial prefrontal cortex have inverse roles in both foraging and economic choice. *Cognitive, Affective & Behavioral Neuroscience*, *53*(9), 1689–1699. https://doi.org/10.1017/CBO9781107415324.004

Shenhav, A., Straccia, M. A., Cohen, J. D., & Botvinick, M. M. (2014). Anterior cingulate engagement in a foraging context reflects choice difficulty, not foraging value. *Nature Neuroscience*, *17*(9), 1249–1254. https://doi.org/10.1038/nn.3771

Shen, Y. J., & Chun, M. M. (2011). Increases in rewards promote flexible behavior. *Attention, Perception, & Psychophysics*, *73*(3), 938–952. https://doi.org/10.3758/s13414-010-0065-7

Siegel, M., Buschman, T. J., & Miller, E. K. (2015). Cortical information flow during flexible sensorimotor decisions. *Science*, *348*(6241), 1352–1355. https://doi.org/10.1126/science.aab0551

Silvetti, M., Alexander, W., Verguts, T., & Brown, J. W. (2014). From conflict management to reward-based decision making: Actors and critics in primate medial frontal cortex. *Neuroscience and Biobehavioral Reviews*, *46*, 44–57. https://doi.org/10.1016/j.neubiorev.2013.11.003

Smith, B. W., Mitchell, D. G. V., Hardin, M. G., Jazbec, S., Fridberg, D., Blair, R. J. R., & Ernst, M. (2009). Neural substrates of reward magnitude, probability, and risk during a wheel of fortune decision-making task. *NeuroImage*, *44*(2), 600–609. https://doi.org/10.1016/j.neuroimage.2008.08.016

Spaak, E., Watanabe, K., Funahashi, S., & Stokes, M. G. (2017). Stable and dynamic coding for working memory in primate prefrontal cortex. *Journal of Neuroscience*, *37*(27), 6503–6516. https://doi.org/10.1523/JNEUROSCI.3364-16.2017

Spiers, H. J., Olafsdottir, H. F., & Lever, C. (2018). Hippocampal CA1 activity correlated with the distance to the goal and navigation performance. *Hippocampus*, *28*(9), 644–658. https://doi.org/10.1002/hipo.22813

Stephens, D. W. (2008). Decision ecology: Foraging and the ecology of animal decision making. *Cognitive, Affective & Behavioral Neuroscience*, *8*(4), 475–484. https://doi.org/10.3758/CABN.8.4.475

Stephens, D. W., & Krebs, J. R. (1986). *Foraging theory*. Princeton University Press.

Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., & Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. *Neuron*, *78*(2), 364–375. https://doi.org/10.1016/j.neuron.2013.01.039

Stoll, F. M., Fontanier, V., & Procyk, E. (2016). Specific frontal neural dynamics contribute to decisions to check. *Nature Communications*, *7:11990*. https://doi.org/10.1038/ncomms11990

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

van den Berg, B., Krebs, R. M., Lorist, M. M., & Woldorff, M. G. (2014). Utilization of reward-prospect enhances preparatory attention and reduces stimulus conflict. *Cognitive, Affective and Behavioral Neuroscience*, *14*(2), 561–577. https://doi.org/10.3758/s13415-014-0281-z

Vandierendonck, A., Liefooghe, B., & Verbruggen, F. (2010). Task switching: Interplay of reconfiguration and interference control. *Psychological Bulletin*, *136*(4), 601–626. https://doi.org/10.1037/a0019791

van Schouwenburg, M. R., den Ouden, H. E. M., & Cools, R. (2010). The human basal ganglia modulate frontal-posterior connectivity during attention shifting. *Journal of Neuroscience*, *30*(29), 9910–9918. https://doi.org/10.1523/JNEUROSCI.1111-10.2010

Vassena, E., Deraeve, J., & Alexander, W. H. (2019). Task-specific prioritization of reward and effort information: Novel insights from behavior and computational modeling. *Cognitive, Affective and Behavioral Neuroscience*, *19*(3), 619–636. https://doi.org/10.3758/s13415-018-00685-w

Vijayraghavan, S., Major, A. J., & Everling, S. (2016). Dopamine D1 and D2 receptors make dissociable contributions to dorsolateral prefrontal cortical regulation of rule-guided oculomotor behavior. *Cell Reports*, *16*(3), 805–816. https://doi.org/10.1016/j.celrep.2016.06.031

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, *137*, 188–200. https://doi.org/10.1016/j.neuroimage.2015.12.012

Waskom, M. L., Kumaran, D., Gordon, a. M., Rissman, J., & Wagner, a. D. (2014). Frontoparietal representations of task context support the flexible control of goal-directed cognition. *Journal of Neuroscience*, *34*(32), 10743–10755. https://doi.org/10.1523/JNEUROSCI.5282-13.2014

Westbrook, A., & Braver, T. S. (2016). Dopamine does double duty in motivating cognitive effort. *Neuron*, *89*(4), 695–710. https://doi.org/10.1016/j.neuron.2015.12.029

Westbrook, A., Kester, D., & Braver, T. S. (2013). What Is the Subjective Cost of Cognitive Effort? Load, Trait, and Aging Effects Revealed by Economic Preference. *PLoS ONE*, *8*(7), 1–8. https://doi.org/10.1371/journal.pone.0068210

Wisniewski, D., Forstmann, B., & Brass, M. (2019). Outcome contingency selectively affects the neural coding of outcomes but not of tasks. *Scientific Reports*, *9*(1), 1–15. https://doi.org/10.1101/375642

Wisniewski, D., Goschke, T., & Haynes, J. D. (2016). Similar coding of freely chosen and externally cued intentions in a fronto-parietal network. *NeuroImage*, *134*, 450–458. https://doi.org/10.1016/j.neuroimage.2016.04.044

Wisniewski, D., Reverberi, C., Momennejad, I., Kahnt, T., & Haynes, J. D. (2015). The role of the parietal cortex in the representation of task–reward associations. *Journal of Neuroscience*, *35*(36), 12355–12365. https://doi.org/10.1523/JNEUROSCI.4882-14.2015

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

Woolgar, A., Afshar, S., Williams, M. A., & Rich, A. N. (2015). Flexible coding of task rules in frontoparietal cortex: An adaptive system for flexible cognitive control. *Journal of Cognitive Neuroscience*, *27*(10), 1895–1911.

Woolgar, A., Hampshire, A., Thompson, R., & Duncan, J. (2011). Adaptive coding of task-relevant information in human frontoparietal cortex. *Journal of Neuroscience*, *31*(41), 14592–14599. https://doi.org/10.1523/JNEUROSCI.2616-11.2011

Worringer, B., Langner, R., Koch, I., Eickhoff, S. B., Eickhoff, C. R., & Binkofski, F. C. (2019). Common and distinct neural correlates of dual-tasking and task-switching: A meta-analytic review and a neuro-cognitive processing model of human multitasking. *Brain Structure and Function*, *224*(5), 1845–1869. https://doi.org/10.1007/s00429-019-01870-4

Yee, D. M., & Braver, T. S. (2018). Interactions of motivation and cognitive control. *Current Opinion in Behavioral Sciences*, *19*, 83–90. https://doi.org/10.1016/j.cobeha.2017.11.009

Yeung, N., Nystrom, L. E., Aronson, J. A., & Cohen, J. D. (2006). Between-task competition and cognitive control in task switching. *Journal of Neuroscience*, *26*(5), 1429–1438. https://doi.org/10.1523/JNEUROSCI.3109-05.2006

Zhang, J., Kriegeskorte, N., Carlin, J. D., & Rowe, J. B. (2013). Choosing the rules: Distinct and overlapping frontoparietal representations of task rules for perceptual decisions. *Journal of Neuroscience*, *33*(29), 11852–11862. https://doi.org/10.1523/JNEUROSCI.5193-12.2013