The Neural Correlates of Exploration

by

Cameron Dale Hassall
BSc, University of Alberta, 2001
BSc, University of British Columbia, 2011
MSc, Dalhousie University, 2013

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in Interdisciplinary Studies

Supervisory Committee

The Neural Correlates of Exploration

by

Cameron Dale Hassall
BSc, University of Alberta, 2001
BSc, University of British Columbia, 2011
MSc, Dalhousie University, 2013

Supervisory Committee

Dr. Olave E. Krigolson (School of Exercise Science, Physical and Health Education)
Supervisor

Dr. Clay B. Holroyd (Department of Psychology)
Co-Supervisor

Dr. Adam Krawitz (Department of Psychology)
Additional Member

Abstract

Supervisory Committee
Dr. Olave E. Krigolson (School of Exercise Science, Physical and Health Education)
Supervisor
Dr. Clay B. Holroyd (Department of Psychology)
Co-Supervisor
Dr. Adam Krawitz (Department of Psychology)
Additional Member

Like other animals, humans explore to learn about the world, and exploit what we have learned in order to maximize reward. The trade-off between exploration and exploitation is a widely-studied topic that cuts across multiple domains, including animal ecology, economics, and computer science. This work approaches the explore-exploit dilemma from the perspective of cognitive neuroscience. In particular, how are our decisions to explore or exploit represented computationally? And how is that representation implemented in the brain? Experiment 1 examined neural signals following outcomes in a risk-taking task. Explorations – defined as slower responses – were preceded by an enhancement of the P300, a component of the human event-related brain potential thought to reflect a phasic release of norepinephrine from locus coeruleus. Experiment 2 revealed that the same neural signal precedes feedback in a learning task called a two-armed bandit. There, a reinforcement learning model was used to classify responses as either exploitations or explorations; exploitations were driven by previous rewards, and explorations were not. Experiments 3 and 4 extended these results in three important ways. First, evidence is presented that the neural signal observed in Experiments 1 and 2 was driven not only by the upcoming decision, but also by the preceding decision (perhaps even more so). Second, Experiments 3 and 4 involved increasingly larger action spaces. Experiment 3 involved choosing from among either 4, 9, or 16 options. Experiment 4 involved searching for rewards in continuous two-

dimensional map. In both experiments, the feedback-locked P300 was enhanced following exploration. Third, exploitation was the more common strategy in Experiments 1 and 2. Thus, it was unclear whether the exploration-related P300 enhancement observed there was due to exploration per se, to exploration rate, or to the fact that exploration was rare compared to exploitation. Experiment 3 partially address this by eliciting different rates of exploration; the exploration-related P300 effect correlated with rate of exploration. In Experiment 4, exploration was more common than exploitation (in contrast to Experiments 1–3); even so, exploration was followed by a P300 enhancement. Together, Experiments 1–4 suggest the presence of a general neural system related to exploration that operates across multiple task types (discrete to continuous), regardless of whether exploration or exploitation is the more common task strategy. The proposed purpose of this neural signal is to interrupt one mode of decision-making (exploration) in favour of another (exploitation).

Table of Contents

Supervisory Committee ii
Abstract iii
Table of Contents v
List of Tables vii
List of Figures viii
Acknowledgments ix
Dedication x
Chapter 1: General Introduction 1
Overview of Experiments 7
Previously Published Material 7
Chapter 2: Experiment 1 9
Methods 15
Participants. 15
Apparatus and procedure. 15
Data collection. 17
Data analysis. 17
Results 20
Behavioural data. 20
Electroencephalographic data. 22
Discussion 25
Computational framework. 26
The P300 and exploratory behaviour. 27
The P300 and reward magnitude. 30
Conclusions. 31
Chapter 3: Experiment 2 33
Methods 39
Participants. 39
Apparatus and procedure. 39
Data collection. 42
Computational model. 43
Data analysis. 45
Results 52
Modelling data. 52
Behavioural data. 53
Electroencephalographic data. 53
Discussion 57
Neural response to feedback. 58
Neural response to bandits. 59
Conclusions. 62
Chapter 4: Experiment 3 63
Methods 67
Participants. 67
Apparatus and procedure. 67
Data collection. 69

Computational models. 69
Data analysis. 72
Results 79
Modelling data. 79
Behavioural data. 80
Electroencephalographic data. 84
Discussion 86
Chapter 5: Experiment 4 92
Methods 95
Participants. 95
Apparatus and procedure. 96
Data collection. 98
Computational models. 98
Data analysis. 103
Results 110
Modelling data. 110
Behavioural data. 111
Electroencephalographic data. 113
Discussion 114
Chapter 6: General Discussion 118
Bibliography 122
Appendix A 140
Appendix B 141

List of Tables

Table 1 Best-fitting model participant counts (N = 30) 79
Table 2 Effects of current/next trial decision 82
Table 3 Behavioural summary (means and 95% confidence intervals) 83
Table 4 ERP summary (means and 95% confidence intervals) 85
Table 6 Effects of current/next trial decision 111
Table 7 Behavioural means, with 95% confidence intervals 113
Table 8 ERP scores, with 95% confidence intervals 114

List of Figures

Figure 1. Experimental design, along with timing details. 16
Figure 2. Time between pumps for Subject 14, Balloon 10. 21
Figure 3. Mean exploration rate. 21
Figure 4. Decision to explore or exploit. 23
Figure 5. sLORETA source analysis of exploration trials compared to exploitation trials at 400 ms post decision. 23
Figure 6. Correlation between decision time (time between pumps) and magnitude of the peak of the ERP in the P300 time range, r(28) = .51, p = .01. 24
Figure 7. Averaged ERP waveforms recorded at channel Cz for low- and high- value bursts and inflations. 25
Figure 8. Two-armed bandit task. 42
Figure 9. Behavioural results. 45
Figure 10. Reward positivity preceding decisions to exploit and explore. 48
Figure 11. Feedback-locked P300 waveforms and scalp distributions preceding decisions to exploit and explore. 49
Figure 12. Choice-locked N200 waveform and scalp distributions. 51
Figure 13. Summary of results. 54
Figure 14. Relationship between each participant's trial-to-trial P300 and the model-generated likelihood (softmax) of the upcoming decision. 56
Figure 15. Task, with timing details. 69
Figure 16. Behavioural and EEG data by current/next trial type. 76
Figure 17. Feedback-locked P300 responses following decisions to either exploit or explore, with difference waveforms and scalp topographies. 77
Figure 18. Feedback-locked P200 responses following decisions to either exploit or explore, with difference waveforms and scalp topographies. 78
Figure 19. Comparison of model fits. 80
Figure 20. Behavioural data. 83
Figure 21. Mean number of prior responses, for each decision type and task. Explorations tended to be preceded by fewer responses of the same choice. Particular response options were sampled less often in the larger decision space. 84
Figure 22. Correlations between P300 difference (exploit minus explore) and likelihood of exploring. 86
Figure 23. Task with timing details. 98
Figure 24. Sample responses and model representations. 101
Figure 25. Model fit results. 103
Figure 26. Behavioural and EEG data by current/next trial type. 107
Figure 27. Feedback-locked waveforms (left) and scalp topographies (right). 107
Figure 28. Feedback-locked waveforms (left) and scalp topography of the explore-minus-exploit difference scores (right). 109
Figure 29. Trial classification. 110
Figure 30. Behavioural results. 113

Acknowledgments

Financial support for this work was provided by the Natural Sciences and Engineering Research Council of Canada, the Neuroeducation Network at the University of Victoria, the University of Victoria Fellowship program, and various University of Victoria donors. I would also like to acknowledge those collaborators who contributed directly to this work: Katharine Holland, Craig McDonald, and Tom Ferguson.

Thank you to Chad Williams for all the great discussions. Thank you to Adam Krawitz for being on my committee and for always having an open door. Thank you to Clay Holroyd for agreeing to be my co-supervisor and for welcoming me into his laboratory. Finally, thank you to my supervisor and friend Olav Krigolson for his generosity and tireless mentorship.

Dedication

To my wife, Aisling, for her unconditional love and support

and

To my children, Henry and James, without whom this work would have been completed two years earlier
(but they are worth it)

Chapter 1: General Introduction

How do we learn to make decisions in an unfamiliar environment? How should decision-making change with learning? These are complicated questions, but part of the answer might relate to the trade-off between *exploration* and *exploitation*. Explorations are decisions that tell us more about the world. Exploratory actions themselves may or may not be rewarding, but the main value in exploration is that it reduces uncertainty and informs future decision-making. With learning, the decision maker may also choose to exploit; they may repeat a previously-rewarded action, thus ensuring some gains. Over-exploitation is risky, though – what if one's picture of the world is inaccurate or incomplete? What if the world changes? Over-exploration, on the other hand, might lead to poor long-term rewards. Thus, optimal decision-making involves managing a trade-off between exploration and exploitation, a process that presumably occurs in the brain, but about which little is currently known.

The importance of a proper balance between exploration and exploitation can be illustrated by an examining these behaviours within the framework of infant and child development. Exploration is seen as an important component of attachment theory, the idea that an emotional connection with a caregiver is essential for healthy development (Bowlby, 1988). There, exploration is defined as excursions away from the caregiver to play independently. Under times of stress, however, the child will return to the caregiver, or "secure base" – in other words, they will exploit. Both behaviours – exploring the world, and having a secure base to return to – are seen as essential for the mental health of both infants/children (Bowlby, 1988) and adults (Feeney, 2004).

Research on the explore-exploit trade-off cuts across different research domains. In economics, for example, these behaviours are of interest at both the level of the individual (e.g., employee or manager) and organization (e.g., company). Here, exploration is synonymous with innovation, discovery, and financial loss. Exploitation, on the other hand, involves idea refinement and – eventually – product development and financial gain. An effective manager or company strikes a good balance between these two pursuits (Laureiro-Martínez, Brusoni, & Zollo, 2010; March, 1996). Animal ecologists also study this balance, in the form of foraging trade-offs (Crawley & Krebs, 1992). Animal foraging models involve many factors, including the role of predators, disease, climate, and food distribution. For example, the worm *Caenorhabditis elegans* tends to exploit more when exposed to a food source, but will shift to an exploratory mode if no food is found (Hendricks, 2015). Interestingly, *C. elegans* individuals raised on small food patches tend to explore more than those raised in richer environments – impressive behaviour from a nervous system with only 302 neurons (Calhoun et al., 2015).

In monkeys, there is evidence the neuromodulator norepinephrine (NE) may help manage the explore-exploit trade-off. Originating mainly in locus coeruleus (LC), NE projects widely to the cortex. Both tonic (baseline) and phasic NE activity are thought to be relevant here. In particular, phasic NE activity is thought to facilitate exploitation, while tonic NE activity is thought to facilitate exploration (Aston-Jones & Cohen, 2005). The two modes of NE activity (phasic/tonic) appear to trade off – during phasic activity, tonic activity tends to be low; during periods of higher tonic activity, phasic bursts are less likely to occur. In signal detection tasks, this manifests behaviourally as good

performance during phasic periods and poor performance (distractibility) during tonic periods. In the context of reward learning, tonic activity may actually be adaptive though because it facilitates disengaging from one response option in favour of something better (exploration). This is seen in reversal learning tasks, in which phasic activity drops and tonic activity increases following a reward reversal. Phasic activity is resumed once the new target has been found (Aston-Jones, Rajkowski, & Kubiak, 1997). Thus, tonic NE activity – usually associated with poor performance and distractibility – may facilitate exploring new response options, and phasic NE activity may indicate that a new preferred response has been found. There is some evidence for this trade-off in humans. For example, Jepma and Nieuwenhuis (2011) used pupillometry – an indirect measure of NE levels – to show that tonic NE activity in humans increases just before they explore. Evidence for the role of phasic NE in managing the explore-exploit trade-off is less clear, however.

How might phasic NE be involved in decisions to explore or exploit? One possible answer is the notion of *neural interrupt*, a mechanism proposed by Dayan and Yu (2006) by which animals can detect and respond to environmental uncertainty. Yu and Dayan (2005) earlier proposed that acetylcholine and NE track different forms of task uncertainty. In particular, they proposed that phasic NE bursts signal the detection of *unexpected* uncertainty, typified as reversals in a reversal-learning task (Yu & Dayan, 2005). Dayan and Yu (2006) later referred to this as a neural interrupt signal, the purpose of which is to halt ongoing cognitive processes in favour of a new task state.

In its original form, the neural interrupt hypothesis describes a model for how an individual detects task switches. For example, a monkey might learn to respond to rare

targets and ignore frequent distractors. Following a switch (the distractor is now the target), a neural interrupt should eventually occur, allowing the monkey to adapt its responses (Dayan & Yu, 2006). In this case, the signalling event – the frequency of each stimulus – is externally-determined. Does the neural interrupt hypothesis apply to internally-determined events, such as decisions to explore or exploit? I argue here that it might – that the interruption of one mode of decision-making (e.g., exploration) in favour of another (e.g., exploitation) is controlled by same neural system that detects task switches – the NE system.

To test this hypothesis, I will require a means to measure phasic NE activity. The method of event-related potentials (ERPs) allows for this, indirectly. Specifically, the P300 ERP component has been linked to phasic NE activity originating in locus coeruleus (the LC-NE hypothesis: Nieuwenhuis, Aston-Jones, & Cohen, 2005). The P300 itself is a positive-deflection in the ERP that typically peaks 300-500 ms post stimulus (Polich, 2007; S. Sutton, Braren, Zubin, & John, 1965). In practice, I will focus on feedback events, such as wins and losses, because of their critical role in reinforcement learning, one of the methods by which humans learn about their environment (R. S. Sutton & Barto, 2018). I will examine the feedback-locked P300 across four tasks in the hopes of discovering a common neural mechanism related to exploration.

Given a particular task, how does one define exploration? This will be a critical question across my four experiments. One strategy that will be used is to examine participant response times; there is evidence that decisions to explore take longer than decisions to exploit (Beharelle, Polanía, Hare, & Ruff, 2015; Pleskac & Wershbale, 2014). It may not always be feasible to rely on response time differences, however – ERP

studies often involve a “go cue” in order to separate stimulus-locked activity from response-locked activity (Luck, 2014). In these cases, we might expect little or no difference between the time it takes explore compared to the time it takes to exploit, as the response has already been prepared well before the action. Additionally, although response times might tell us something general about exploration (e.g., that it requires additional processing time), they do not really suggest much about how decisions to explore/exploit might be represented in the brain. For this, we turn to computational modelling.

Seminal work by Daw, O’Doherty, Dayan, Seymour, and Dolan (2006) provides an excellent template for a *computational neuroscience* approach to studying the explore-exploit dilemma. Daw and colleagues (2006) presented participants with four slot machines, each with a different payout probability. Through trial-and-error learning participants learned, over time, which slot machine was more likely to yield a high reward. The main contribution of this work was that it considered various computational models in terms of their ability to account for participant choices. They considered reinforcement learning models generally, which use feedback to track values associated with each response option. Once a suitable model was found (i.e., one that did a good job of predicting which slot machines would be chosen), the authors were able to classify participant decisions as either explorations or exploitations (Daw et al., 2006). They did this by considering exploitations as those decisions that maximized the model-generated value of the chosen options. Other decisions (i.e., those not drive by value) were classified as explorations. This allowed the authors to determine the cortical locations of neural activity associated with each decision type (Daw et al., 2006).

A similar strategy will be followed for the majority of my work. The appeal of a computational approach (as compared to an examination of response times, for example) is that it will require us to be explicit about how decisions to explore/exploit are represented in the brain. In general, I will use the same technique as Daw et al. (2006); for each of my tasks, I will look at how well various computational models account for, or predict, my participants' behaviour. My main goal in doing so is not to make any claim about a "true" neural representation. Rather, my main goal is to inform my ERP analyses by classifying trials as either explorations or exploitations. Although some inferences about likely neural representations will be made, I will note that the best-fitting model may depend on both the individual and the task.

Various tasks have been used to examine human foraging behaviours. They typically involve trial-and-error learning, with a participant choosing from among several discrete options. Participants can choose to stick with a choice (exploit) or try something new (explore). Of note, laboratory foraging tasks tend to have little-to-no ecological validity. Mobbs, Trimmer, Blumstein, and Dayan (2018) recently explored this issue, concluding that current practices fail to consider *ethology*, the study of animals in their natural environment. They point out several known factors in animal foraging that remain understudied in humans. These include physical costs, competition, and predation. One unanswered question they raise relates to the effect of the foraging environment on human behaviour (Mobbs et al., 2018). For example, how are decisions to explore affected by the number of options or the reward distribution? These questions will be addressed in two of my experiments.

Overview of Experiments

In summary, I propose that decisions to explore are accompanied by a neural interrupt signal, measurable on the scalp as an enhancement of the P300 ERP component. In Experiment 1, I will examine decisions to explore/exploit in the balloon analogue risk task (the BART). In the BART, participants attempt to inflate a virtual balloon as much as possible without breaking. Based on previous computational work, an exploration here is defined as a pump action preceded by a longer-than-usual pause. Experiment 2 will use a typical feedback-learning task called a two-armed bandit. This task has more distinctive win and loss events compared to the BART and will allow a more direct examination of the feedback-locked neural response. In Experiments 3 and 4 I will enlarge the action space from Experiment 2 in an attempt to determine whether neural foraging signals generalize to more ecologically-valid (and *ethologically*-valid) tasks. Experiment 3 simply adds more response options to the two-armed bandit paradigm: either 4, 9, or 16 choices. Finally, in Experiment 4 participants will search for spatially-correlated rewards within a virtual map – a continuous version of Experiments 2 and 3.

Previously Published Material

Experiment 1 was previously published in Neuroscience as "What Do I Do Now? An Electroencephalographic Investigation of the Explore/Exploit Dilemma" (https://doi.org/10.1016/j.neuroscience.2012.10.040). I am the first author. I designed the study, collected and analyzed the data, and wrote the paper. The second author, Katharine Holland, collected the data. The third author, Olav Krigolson, conceived the study and reviewed/edited the paper.

Experiment 2 was previously published in Brain Research as “Ready, Set, Explore! Event-related Potentials Reveal the Time-course of Exploratory Decisions” (https://doi.org/10.1016/j.brainres.2019.05.039). I am the first author. I conceived and designed the study, collected and analyzed the data, and wrote the paper. The second author, Craig McDonald, reviewed/edited the paper. The third author, Olav Krigolson, reviewed/edited the paper.

Chapter 2: Experiment 1

Abstract

To maximize reward, we are faced with the dilemma of having to balance the exploration of new response options and the exploitation of previous choices. Here, I sought to determine if the event-related brain potential (ERP) in the P300 time range is sensitive to decisions to explore or exploit within the context of a sequential risk-taking task. Specifically, the task I used required participants to continually explore their options – whether they should "push their luck" and keep gambling or "take the money and run" and collect their winnings. My behavioural analysis yielded two distinct distributions of response times: a larger group of short decision times and a smaller group of long decision times. Interestingly, these data suggest that participants adopted one of two modes of control on any given trial: a mode where they quickly decided to keep gambling (i.e. exploit), and a mode where they deliberated whether to the take the money they had already won or continue gambling (i.e. explore). Importantly, I found that the amplitude of the ERP in the P300 time range was larger for explorative decisions than for exploitative decisions and, further, was correlated with decision time. My results are consistent with a recent theoretical account that links changes in ERP amplitude in the P300 time range with phasic activity of the locus coeruleus-norepinephrine (LC-NE) system and decisions to engage in exploratory behaviour.

What Do I Do Now? An Electroencephalographic Investigation of the Explore/Exploit Dilemma

In Mill's Utilitarianism (1863/2008), he argued that humans have an inherent desire to maximize utility. As such, the decisions that we make on a day-to-day and moment-to-moment basis typically reflect a desire to maximize reward. However, as Dennett (1986) and others have pointed out, calculating the utility of decisions in the real world can be challenging because the potential consequences of our actions are not always known. Even if utility calculations are restricted to the near future, complex or novel situations may arise that require exploring options with unknown consequences. Exploration is inherently risky but necessary in order to assess new response options or reassess old ones. The knowledge gained through exploration can later be exploited to improve subsequent decisions, and thus yield even greater increases in utility. However, one cannot always engage in exploratory behaviour. Rather, one must balance exploratory behaviour with exploitation – selecting the most rewarding response option as much as possible. Therefore, an optimal decision strategy for maximizing utility would entail utilizing an exploitative mode of control most of the time with occasional instances of exploratory behaviour.

Experimentally, decisions to explore or exploit can be studied in tasks such as the Balloon Analog Risk Task (BART: Lejuez et al., 2002). During performance of the BART, participants must continually explore their options – either take the money they have already earned or continue gambling. The key manipulation of the BART is that, for each pump of the balloon (gamble), the amount of money earned increases along with the probability of losing all earned money. This manipulation makes each gamble

increasingly risky. Thus, there is an optimal response in the BART (i.e. total number of balloon pumps) that is based on the risk and reward structure of the task (Lejuez et al., 2002), and as such, to maximize reward, participants need to explore in order to determine the optimal number of balloon pumps. Computational models of the BART suggest that people make a risk assessment prior to each pump: a decision to continue pumping or collect their accumulated reward (Wallsten, Pleskac, & Lejuez, 2005). The Wallsten et al. (2005) model's predictions were recently corroborated by Pleskac and Weshbale (2014) who observed two distinct distributions of response times in human BART performance. Specifically, they observed that people generally made automatic, rapid responses in the BART, but occasionally paused to assess whether or not they should continue gambling. Pleskac and Weshbale (2014) hypothesized that these pauses represent the assessments predicted by earlier modelling work (Pleskac, 2008; Wallsten et al., 2005). Interestingly, the number of assessments that participants made during the BART decreased over time. Importantly, this change in assessment rate is consistent with theoretical models of the exploration/exploitation dilemma. Early in learning, people need to explore more often in order to determine the reward structure of a task (e.g., the optimal number of pumps in the BART). However, once the reward structure is known, people exploit more frequently. With all of this in mind, Pleskac and Weshbale (2014) likened fast BART responses to exploitation and slower responses to exploration.

Research examining the neural basis of decisions to explore or exploit is limited (see Cohen, McClure, & Yu, 2007 for a review). In one recent study, Cavanagh, Figueroa, Cohen, and Frank (2011) suggested increased frontal theta-band oscillation as a possible neural marker of uncertainty-driven exploration. Specifically, Cavanagh and

colleagues (2011) observed a correlation between medial-frontal theta power and the parameters of their reinforcement-learning model during exploration in a decision-making task. From their results, Cavanagh et al. (2011) hypothesized that midbrain regions were responsible for exploitation but that frontal brain regions took control when deciding to explore in uncertain situations. The Cavanagh et al. (2011) hypothesis is consistent with an earlier functional magnetic resonance imaging (fMRI) study that showed enhanced frontal brain activity during exploratory decisions in a four-armed bandit task (Daw et al., 2006). Cavanagh and colleagues' (2011) hypothesis is also consistent with work by Frank, Doll, Oas-Terpstra, and Moreno (2009) that associated a prefrontal cortex (PFC) dopamine gene (COMT) with exploratory decisions. In particular, Frank et al. (2009) showed an effect of COMT gene dose (which they defined as the amount of methionine-encoding or *met* allele present) on uncertainty-driven exploration. The presence of the met allele is linked to increased PFC dopamine levels compared to the presence of the valine-encoding or *val* allele. Although Frank et al. (2009) were uncertain about the exact role of COMT in exploratory behaviour, they suggested that the observed and known effects of the met allele implicate the PFC as the controller of uncertainty-driven exploration. Taken together, these studies suggest that switching from an exploitative to an explorative mode of control involves the intervention of frontal cognitive systems over midbrain lower-level reward-processing systems (see Mars, Sallet, Rushworth, & Yeung, 2011, for more examples of cognitive control) .

Currently, there are no definitive electroencephalographic correlates differentiating decisions to explore or exploit. Having said that, there are good reasons to

hypothesize that the event-related brain potential (ERP) in the time range of the P300 may be sensitive to this distinction. The P300 is a high-amplitude, positive ERP component with peak latency 300–500 ms following the presentation of a stimulus (S. Sutton et al., 1965) that has been associated with several different cognitive functions (Polich, 2007). One influential account – the context-updating hypothesis – states that the P300 reflects the updating of an internal model of the probabilistic structure of the world (Donchin, 1981; Donchin & Coles, 1988). Donchin's (1981) account arose out of earlier observations that the P300 is sensitive to stimulus frequency (Duncan-Johnson & Donchin, 1977). Consistent with the context-updating hypothesis, Nieuwenhuis, Aston-Jones, and Cohen (2005) recently suggested that ERP changes in the P300 time range reflect the locus coeruleus-norepinephrine (LC-NE) system's response to internal decision-making processes regarding task-relevant stimuli (Aston-Jones & Cohen, 2005; Nieuwenhuis, De Geus, & Aston-Jones, 2011; also see Pineda, Foote, & Neville, 1989, for early work linking the LC and the P300). The LC contains noradrenergic neurons and provides the only source of NE to the hippocampus and neocortex (Berridge & Waterhouse, 2003). Increases in LC activity, and the associated rise in NE, are linked to increased exploratory behaviour in monkeys (Aston-Jones & Bloom, 1981; Aston-Jones & Cohen, 2005; modelled by McClure, Gilzenrat, & Cohen, 2006; Usher, Cohen, Servan-Schreiber, Rajkowski, & Aston-Jones, 1999). Importantly, a series of lesion, psychopharmacological, and electroencephalographic studies support the link between an ERP difference in the P300 time range and phasic changes in the activity of the LC-NE system (see Nieuwenhuis et al., 2005, for a review). Thus, given the link between the LC-NE system and exploration, and the link between the LC-NE system and the P300, it

stands to reason that the amplitude of the ERP in the P300 time range may differentiate decisions to explore or exploit.

Our main purpose here was to determine whether or not ERP amplitude in the P300 time range would be sensitive to decisions to explore or exploit. To accomplish this, I had participants perform a modified version of the BART while electroencephalographic (EEG) data were recorded. In terms of behaviour, I expected to observe a similar distribution of response times as Pleskac and Wershbale (2014). In particular, I expected to see two distinct distributions of response times: one distribution of fast responses indicative of exploitation, and a second distribution of slow responses indicative of exploration. Importantly, I predicted that the amplitude of the ERP in the P300 time range preceding decisions to explore would be greater than the ERP amplitude in the same time range preceding decisions to exploit – a prediction derived from Nieuwenhuis and colleagues' (2005) hypothesis that ERP modulation in the P300 time range is driven by phasic changes in LC-NE activity linked to internal decision-making processes.

There is a growing body of evidence that the amplitude of the P300 is also modulated by reward magnitude (Bellebaum & Daum, 2008; Hajcak, Moser, Holroyd, & Simons, 2006; Y. Wu & Zhou, 2009; Yeung & Sanfey, 2004). The P300's sensitivity to reward magnitude is of particular importance here because the purpose of exploration is to specify or update values associated with actions, and the purpose of exploitation is to take advantage of current value assessments (Sutton & Barto, 2018). As such, I also hypothesized that the amplitude of the P300 elicited by balloon bursts would scale with

the magnitude of the amount of lost reward, reflecting an update of participants' model of the probabilistic reward structure of the task.

Methods

Participants. Fourteen right-handed university-aged participants (2 male, mean age: 21.5 +/- 1.5) with no known neurological impairments and with normal or corrected-to-normal vision took part in the experiment. All of the participants were volunteers who received monetary compensation for their participation. The participants provided informed consent approved by the Office of the Vice-President, Research, Dalhousie University, and the study was conducted in accordance with the ethical standards prescribed in the 1964 Declaration of Helsinki.

Apparatus and procedure. Participants were seated comfortably 75 cm in front of a computer monitor and used a standard USB controller to perform a computerized risk-taking task (written in MATLAB [Version 7.14, Mathworks, Natick, U.S.A.] using the Psychophysics Toolbox Extension, Brainard, 1997). To perform the task, participants pushed a button on the controller to inflate a "balloon" (initially a 2.8 cm diameter green circle, subtending 2.1 degrees of visual angle) and earn money. Each trial began with the presentation of a fixation cross for one second. After one second, a green-coloured balloon appeared behind the fixation cross, cuing participants to begin self-paced pumping. With each pump, the balloon either "grew" (increasing in size by 0.3 degrees of visual angle) and the participant won five cents, or the balloon "exploded" (turned red – see below for more detail on the probability of the balloon exploding) and the participant lost all of the money he or she had won during that trial. As such, prior to each pump, participants had to decide whether or not to pump and potentially earn more money, or to

stop the trial and take the money that they had already won (see Figure 1 for timing details). After each group of ten trials, participants were given a self-paced rest break. The experiment consisted of 300 trials in total. All trials were paid at a rate of 20:1 so that the average total payoff was $9.37 +/- $0.16, with individual total payoffs ranging from $8.27 to $10.42.

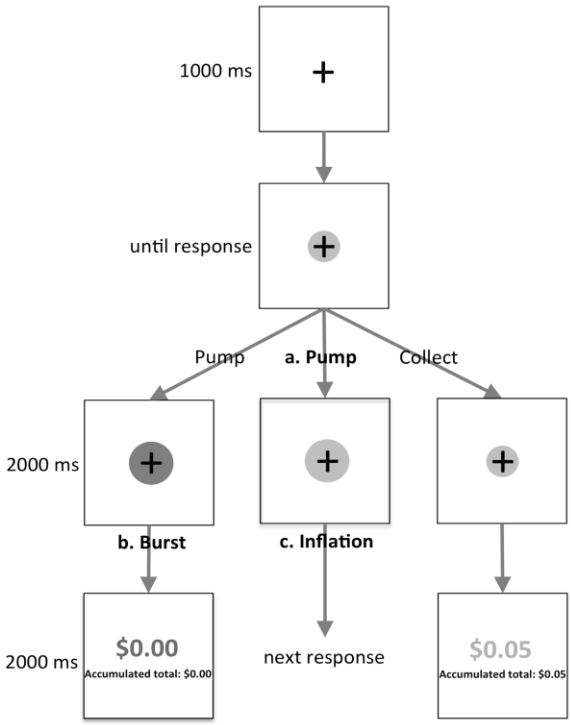


Figure 1. Experimental design, along with timing details. Participants could respond by either pumping the balloon or collecting the accumulated reward. Pumps could result in a successful inflation, or a balloon burst, in which case the accumulated reward for that balloon was lost. Relevant EEG data were recorded at (a) decisions to pump that were followed by a balloon inflation, (b) balloon bursts, and (c) balloon inflations.

Participants were informed that they would play 300 trials, but were given no prior information on the probability structure that governed the balloon exploding; rather, they were only informed “it is up to you to decide how much to pump each balloon –

some may pop after one pump, and some may not pop until the balloon fills the whole screen." In reality, and unbeknownst to participants, the computer program allowed a maximum of 30 pumps, and the balloon exploded randomly with a probability of $(31 - n)^{-1.4}$ on trial n.

Data collection. The experimental program recorded response time (elapsed time from the previous button press or start of trial, in ms), decision type (pump or collect), and whether or not the balloon grew or exploded. The electroencephalogram (EEG) was recorded from 64 electrodes using BrainVision Recorder software (Version 1.20, Brainproducts, GmbH, Munich, Germany). The electrodes were mounted in a fitted cap with a standard 10-20 layout and were recorded with an average reference built into the amplifier (see www.neuroeconlab.com for the exact electrode configuration). Vertical and horizontal electrooculograms were recorded from electrodes placed above and below the right eye and on the outer canthi of the left and right eyes. Electrode impedances were kept below 20 kΩ at all times. The EEG data were sampled at 1000 Hz, amplified (Quick Amp, Brainproducts, GmbH, Munich, Germany), and filtered through a passband of 0.017–67.5 Hz (90 dB octave roll off).

Data analysis. For each response (balloon pump), a response time defined as the elapsed time since the previous response was recorded. Balloon pumps with a response time less than 100 ms or greater than 2000 ms were excluded from subsequent analysis. Next, I classified each balloon pump as corresponding either to a decision to explore or a decision to exploit. Based on Pleskac and Wershbale (2014), I classified response times more than three standard deviations above the mean as decisions to explore. Thus, the increase in balloon size for a successful pump prior to a long response time was marked

as the time point at which participants began “exploring” or, in other words, considering their options. All other balloon pumps were classified as “exploitations”, with the preceding increase in balloon size marked as the time point following which a decision was made to exploit. Thus, I was able to relabel the EEG data following data collection, and then use these revised labels to epoch the EEG data into segments containing decisions to explore or exploit.

The preprocessing of the EEG data began with the application of a 0.1–20 Hz phase shift-free Butterworth filter, following which the continuous EEG data were re-referenced to the average of the two mastoid channels. As mentioned previously, my ERP hypotheses concerned two events: decisions to explore or exploit, and balloon bursts. To test whether the amplitude of the ERP in the P300 time range was sensitive to the decision to explore or exploit, 800 ms epochs of data (from 200 ms before the increase in balloon size to 600 ms after the increase in balloon size) were extracted from the continuous EEG for each trial, channel, and participant, for each condition (explore/exploit). Following isolation of the epoched data, ocular artifacts were corrected using the algorithm described by (Gratton, Coles, & Donchin, 1983). Subsequent to this, all trials were baseline corrected using a 200 ms epoch prior to stimulus onset. Finally, trials in which the change in voltage in any channel exceeded 10 μV per sampling point, or the change in voltage across the epoch was greater than 100 μV, were discarded. In total, less than 2% of the data were discarded.

Our preprocessing resulted in far more exploitation than exploration segments; as such, only exploitation segments that immediately preceded exploration segments were used in the subsequent ERP analysis. Specifically, my average ERP waveforms only

included the 100 epochs corresponding to the 100 longest exploration periods and the 100 epochs (i.e., exploitation periods) immediately preceding them. Subsequent to the creation of the average ERP waveforms for each participants and condition (explore/exploit) I created difference waveforms for each participant and channel by subtracting the average exploitation waveforms from the average exploration waveforms. A visual examination of the grand average difference waveforms and a review of recent research (Duncan et al., 2009; Nieuwenhuis et al., 2011; Polich, 2007) led to a decision to quantify the magnitude of the ERP in the P300 time range as the maximum positive deflection of the difference waveform 300–450 ms following the increase in balloon size at the centro-parietal channel where the difference was maximal (channel CP2). The resulting ERP amplitudes were then statistically tested against zero using a single-sample t-test, with an assumed alpha level of .05.

To evaluate whether the amplitude of the P300 was sensitive to accumulated reward magnitude, 800 ms epochs of data (from 200 ms before balloon burst/growth onset to 600 ms after burst/growth onset) were extracted from the continuous EEG for each trial, channel, and participant for early and late balloon bursts (i.e., losses) and for the increase in balloon size immediately preceding the balloon bursts (i.e., potential gains). Early balloon bursts/growths were defined as bursts that were preceded by between 1 and 15 successful pumps. Late bursts/growths were preceded by between 16 and 30 successful pumps. I then preprocessed the EEG data in an identical manner as outlined above. Following preprocessing, ERPs were created by averaging the EEG data by condition for each electrode channel and participant separately for early and late gains and losses.

To quantify the P300 evoked by balloon bursts, I created a difference waveform for each participant and channel by subtracting the gain (growth) waveforms from the subsequent loss (burst) waveforms for both early and late balloons (see above). As before, the P300 was defined as the maximum positive deflection in the difference waveforms 300–450 ms following stimulus onset for each balloon burst (early/late) at electrode site Cz, where the difference was maximal. P300 amplitudes were then statistically tested against zero using a single-sample t-test, with an assumed alpha level of .05.

Results

Behavioural data. A visual examination of the behavioural data revealed a subset of trials with longer response times – presumably, trials in which participants deliberated whether to take their accumulated money or continue playing (i.e., exploration). Long decision times (long inter-pump times) were defined as those more than three standard deviations above the mean. See Figure 2 for a set of sample responses. Explore decision points were defined as increases in balloon size preceding long inter-pump times. All other increases in balloon size were classified as exploitations – trials in which the response time was short, suggesting an exploitative mode of control. This criterion created two separate distributions of decision times, each with a different mean ($p < .01$): shorter decision times for exploitative decisions (404 +/- 31 ms), and longer decision times for exploratory decisions (798 +/- 50 ms), consistent with Pleskac and Wershbale (2014). Also consistent with Pleskac and Wershbale (2014),

participants explored less (3 +/- 0.4% of trials for balloons numbered 51–300 compared to 15 +/- 3% for balloons 1–50) as they became more familiar with the task (Figure 3).

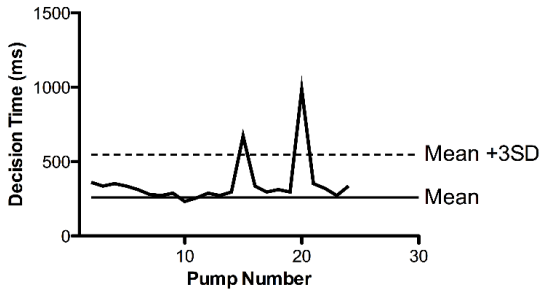


Figure 2. Time between pumps for Subject 14, Balloon 10. Mean response time was characterized by short, somewhat automatic pumps (exploitations). Response times more than 3 standard deviations above the mean were classified as explorations.

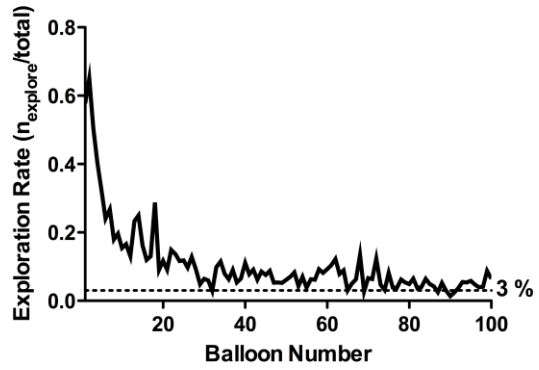


Figure 3. Mean exploration rate. The mean number of explorations per balloon decreased over time. Only the first 100 out of 300 balloons are shown to emphasize the change in exploration rate over the first few balloons. A horizontal line is shown at 3%, the mean exploration rate for balloons 51–300.

Electroencephalographic data.

Exploration. Recall that I predicted exploration would lead to a larger ERP response in the P300 time range preceding longer response times, as I believed that this reflected deliberation of the decision to explore or exploit. Indeed, my analysis of the ERP waveforms in the P300 time range supported my hypothesis as I found a difference between explorative and exploitative trials that was maximal at electrode CP2. Specifically, I found a larger (more positive) ERP response in the P300 time range for exploration trials (1.79 μV +/- 0.40 μV) relative to exploitation trials (0.47 μV +/- 0.39 μV), $t(13) = 5.202$, $p < .01$ (see Figure 4)[1]. I then localized the source of the voltage difference between exploration and exploitation trials using standardized low-resolution brain electromagnetic tomography software (sLORETA: Pascual-Marqui, 2002). An sLORETA analysis at 400 ms post decision (when the ERP response in the P300 time range was maximal) indicated maximal current sources in Brodmann Areas 6 and 10 within the superior frontal gyrus (Figure 5). Finally, ERP amplitude in the P300 time range for both exploration and exploitation trials correlated positively with decision time, $r(28) = .51$, $p = .01$ (see Figure 6).

[1] We also statistically tested whether the N1 was sensitive to decisions to explore/exploit. No difference was seen between decisions to explore/exploit in the N1 time range (130–190 ms post stimulus: $t(13) = .46$, $p = .67$).

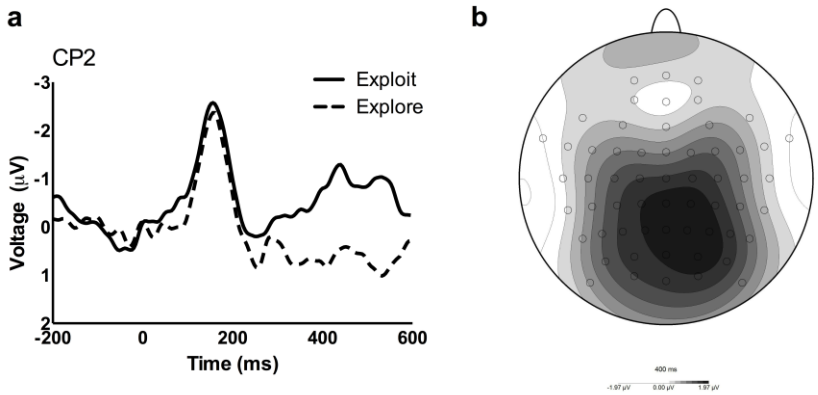


Figure 4. Decision to explore or exploit. Note that 0 ms corresponds to the onset of the decision (balloon pump). Negative voltages are plotted up by convention. (a) Averaged ERP waveforms recorded at channel CP2 for exploration and exploitation decisions. (b) ERP topography map for the difference waveform (explore minus exploit) at 400 ms post decision.

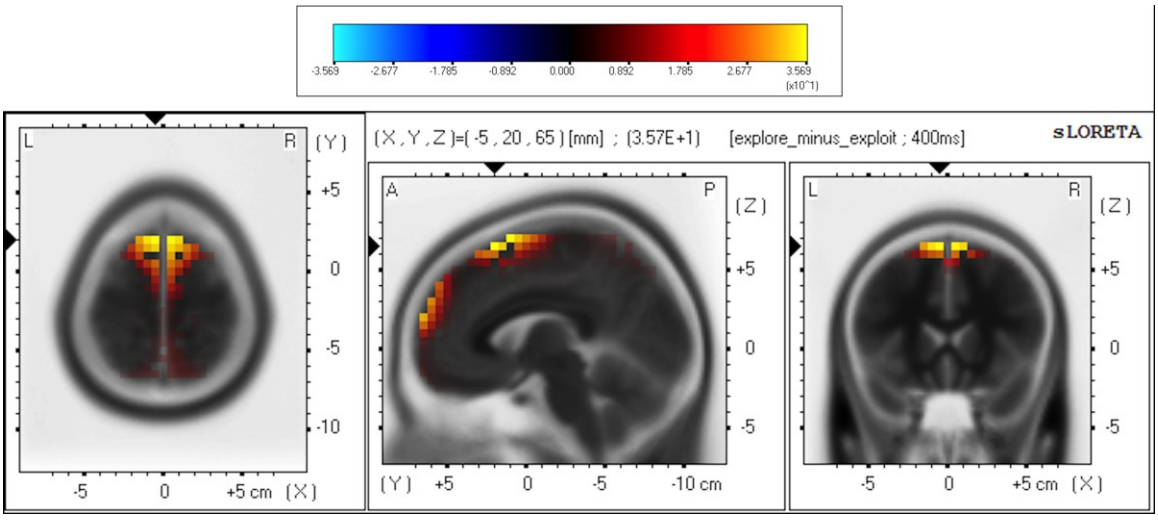


Figure 5. sLORETA source analysis of exploration trials compared to exploitation trials at 400 ms post decision. Statistical nonparametric mapping (SnPM) at a significance level of .05 revealed differences localized in Brodmann Areas 6 (sLORETA value = 35.7) and 10 (sLORETA value = 31.8) within the superior frontal gyrus.

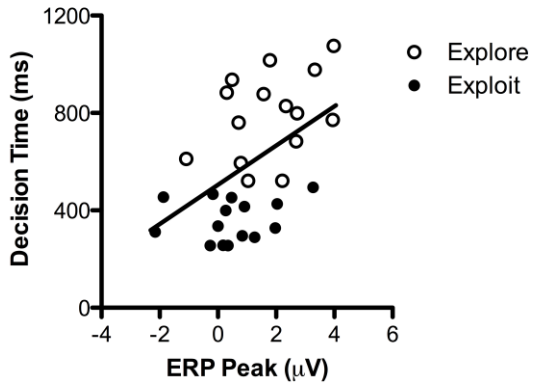


Figure 6. Correlation between decision time (time between pumps) and magnitude of the peak of the ERP in the P300 time range, $r(28) = .51$, $p = .01$.

Balloon bursts. I also wanted to see if the P300 following balloon bursts was sensitive to accumulated reward magnitude, since balloon bursts later in a trial sequence reflected a loss of more money as more money had accumulated. On average, there was an equal number of early bursts (53.0 +/- 2.2) compared to late bursts (54.1 +/- 3.5), $p = .8$. In line with my prediction, I found that the amplitude of the P300 scaled to reward magnitude: late high-valued pumps (defined as pumps 16–30: 35.53 μV +/- 1.69 μV) versus early low-valued pumps (defined as pumps 1–15: 28.24 μV +/- 2.15 μV), $t(13) = 5.00$, $p < .001$ (see Figure 7).

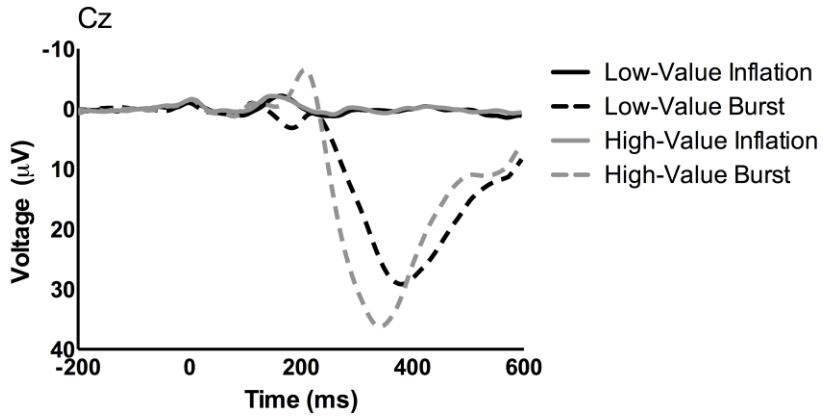


Figure 7. Averaged ERP waveforms recorded at channel Cz for low- and high- value bursts and inflations. Note that 0 ms corresponds either to the onset of the balloon burst or the onset of the balloon inflation. Negative voltages are plotted up by convention.

Discussion

In the present study, decisions to explore in a sequential risk-taking task elicited a larger ERP response in the time range of the P300 – a component sensitive to cognitive processing (Donchin, 1981; Donchin & Coles, 1988) and linked to phasic activity of the LC-NE system (Nieuwenhuis et al., 2005). Supporting my ERP result, my behavioural data mirrored previous work (Pleskac & Weshbale, 2014). I observed that response times in a sequential risk-taking task followed one of two distributions: longer response times indicative of exploration and shorter response times indicative of exploitation. Furthermore, I found that participants explored less over time as they became familiar with the probabilistic structure of the task, a result consistent with observations by Pleskac and Wershbale (2014) and reinforcement-learning theory in general (Sutton & Barto, 2018).

Computational framework. Like earlier work on exploration in humans (Cavanagh et al., 2011; Daw et al., 2006), I relied on a theoretical model (Walsten et al., 2005; Pleskac, 2008; Pleskac & Wershbale, 2014) to identify participants' decisions to explore or exploit during task performance. Recall, decisions preceding fast responses were classified as exploitatory, while decisions preceding long responses were classified as exploratory. The validity of this criterion is critical when interpreting my findings because, while my difference wave in the P300 time range for explore/exploit decisions statistically differed from zero, it was computed by averaging over a post hoc selection of EEG segments derived from this classification system.

Previous research justifies my approach. In a seminal study, Wallsten et al. (2005) evaluated several models of BART performance by comparing their simulated outputs to human behavioural data. Wallsten et al. (2005) found some variation in exploratory behaviour among individual human participants, with some participants continuing to gamble after the optimal number of pumps. To account for this, Wallsten and colleagues' model included components that decided how many pumps to make and whether to stop or keep going prior to each individual pump. In a later improvement of the Wallsten et al. (2005) model called the BSR (Bayesian sequential risk-taking model), Pleskac (2008) included an individual response bias that changed over time (see Busemeyer & Pleskac, 2009, for a review of the different components of dynamic decision-making models). Pleskac and Wershbale (2014) later amended the BSR to account for observed delays in response times so that assessments (decisions to either continue or stop) only occurred on a subset of trials. The trials associated with exploratory behaviour were preceded by longer response times – explained as an increase in cognitive load linked to the decision

process. Notably, and in line with human data, the model predicted that participants would tend to make fewer assessments over time, a prediction consistent with both exploratory behaviour and the pattern of results I observed in my data. The most recent version of the BSR (Pleskac & Wershbale, 2014) provided a good fit for human BART data, including between-subject variation in response selection, and within-subject variation in response-time. In the present experiment, my participants' response time distributions mirrored Pleskac and Wershbale's (2014), thus providing strong support for the use of a response-time criterion to classify participant EEG segments as either containing decisions to explore or exploit.

The P300 and exploratory behaviour. My result that ERP amplitude in the P300 time range was larger for decisions to explore is consistent with the context-updating hypothesis of the P300 (Donchin, 1981; Donchin & Coles, 1988). Under this theoretical framework, a P300 is observed whenever new information requires an update to one's internal mental model of the world – specifically, the probabilistic framework of a particular task (Donchin & Coles, 1988). In my case, to maximize utility, participants had to learn the optimal number of pumps to undertake, a challenging task taking into account the value of a given pump and the risk associated with different balloon sizes (i.e. that larger balloons entailed greater risk). Each pump, whether it resulted in a balloon burst or successful balloon inflation, thus provided information for participants. This notion is corroborated by earlier modelling work (i.e., Pleskac & Wershbale, 2014) suggesting that participants consider new information and review their potential actions at various points throughout a sequential decision-making task – points marked by longer-than-normal response times. It is at these assessment points, I claim, that

participants incorporate new information into their model of the BART and then decide whether or not to continue pumping. As such, at assessment points a larger ERP in the P300 time range is observed, reflecting the incorporation of new information into the internal model and a subsequent exploratory decision. Interestingly, the length of the assessment period correlated with the amplitude of the subsequent ERP in the P300 time range (Figure 6) – a result that further supports my hypothesis that the ERP in the P300 time range is sensitive to decisions to explore or exploit. Finally, an sLORETA source analysis (Pascual-Marqui, 2002) revealed a difference in frontal brain regions for exploration trials compared to exploitation trials, consistent with earlier research (Daw et al., 2006; Frank et al., 2009; Cavanagh et al., 2011).

An unavoidable limitation in this study arose because participants were asked to respond as quickly as they wanted to. As such, the mean response time corresponding to decisions to exploit (404 +/- 31 ms) suggests that some of the EEG segments containing decisions to exploit might have overlapped with the following decision. However, that participant responses were self-paced seems an important part of the BART design, especially if a clear distinction between explorations and exploitations is to be achieved (Lejuez et al., 2002; Pleskac & Wershbale, 2014). Although there are versions of the BART that introduce timing delays (Rao, Korczykowski, Pluta, Hoang, & Detre, 2008; Fukunaga, Brown, & Bogg, 2012), those versions do not, to my knowledge, produce the two distributions of response times necessary to classify responses as explorations or exploitations (Pleskac & Wershbale, 2014).

An alternative explanation for my findings relates to Nieuwenhuis and colleagues' (2005) hypothesis that the P300 time range is modulated by phasic activity of the LC-NE

system. Interestingly, research by Usher et al. (1999) suggests that modulatory activity of the LC is responsible for regulating exploratory behaviour in monkeys. Extending from this, Nieuwenhuis et al. (2005) proposed that the LC may regulate exploratory behaviour in humans through the release of NE, with the change to an exploratory mode of control marked by a related increase in ERP magnitude in the P300 time range. Supporting this contention, Aston-Jones and Cohen (2005) suggested that LC phasic activity is driven by computations about value in the orbitofrontal cortex (OFC) and anterior cingulate cortex (ACC). They further suggested that the purpose of LC phasic release of NE is to break out of one behavioural routine (e.g. exploitation) to engage in a different behaviour (e.g. exploration). Importantly, my data support Aston-Jones and Cohen's (2005) suggestion and the hypothesized link between the LC and the P300 (i.e., Nieuwenhuis et al., 2005) as I observed an increase in the amplitude of the ERP in the P300 time range when participants changed to an exploratory mode of control.

A second alterative explanation for my results relates to response time. Recently, Grinband et al. (2011)suggested that time on task, rather than an increase in cognitive control, might be responsible for increased frontal cortex activity. Grinband et al. (2011) asked participants to balance speed and accuracy in a Stroop task and observed that response times were slower and frontal cortex activity greater on incongruent trials compared to congruent trials. However, when slow and fast congruent trials were compared, Grinband et al. (2011) noted increased frontal activity for slower trials, even though congruency was controlled for. This somewhat controversial finding (e.g., Yeung, Cohen, & Botvinick, 2011) is relevant to the current study since I used response times to categorize decisions as explorations or exploitations. I observed an enhanced P300 for

longer response times (classified as explorations). This is consistent with Grinband and colleagues' (2011) result, provided one is willing to extend a conflict-monitoring result to the exploration/exploitation dilemma (see Ishii, Yoshida, & Yoshimoto, 2002; Khamassi et al., 2011, for some arguments supporting this comparison).

Although the body of research on the EEG correlates of the exploration/exploitation dilemma is sparse, it is growing. For example, Tzovara and colleagues (2012) recently used EEG to study the Daw et al. (2006) gambling paradigm and observed increased frontal brain activity prior to exploratory decisions, which they were able to define based on a computational model. Like us, Tzovara et al. (2012) compared ERPs to feedback prior to participant decisions to explore or exploit, and observed a difference. However, because Tzovara et al. (2012) only examined responses to feedback it is unclear whether their observed difference was due to the result of a decision to explore, reward evaluation, or both. Interestingly, Tzovara et al. (2012) observed that feedback ERP differences (including P300) predicted whether or not participants explored on subsequent trials. This lends further support to my second hypothesis that the P300 scales with reward magnitude, and my speculation that changing representations of value (as indexed by the P300) drive exploration (Sutton & Barto, 2018). A major strength of the present study, and one that distinguishes it from earlier work on the explore/exploit dilemma, is that I was able to examine ERP responses to the explore/exploit decisions themselves, as opposed to responses to feedback alone.

The P300 and reward magnitude. I also observed that the amplitude of the P300 was sensitive to reward magnitude. Specifically, I found a larger P300 amplitude for high-valued losses (balloon bursts) compared to low-valued losses – a result reflective

of a neural representation of the magnitude of the value of taking different actions. In this case, the aforementioned representation related to the negative value associated with losses following early low-valued pumps versus later high-valued pumps. This finding is consistent with earlier work showing that the amplitude of the P300 is (a) sensitive to the magnitude of both wins and losses (Yeung & Sanfey, 2004), and (b) could be related to the motivational significance of feedback (Nieuwenhuis et al., 2005; Niuewenhuis, 2011). Simply put, high-valued rewards and losses are more motivationally significant than low-valued rewards and losses. Of particular relevance here, Yeung and Sanfey (2004) speculated that the P300 might be impacted by the magnitude of actual and alternate outcomes (what might have been) – in other words, they speculated that the P300 reflects an objective representation of reward magnitude, regardless of whether or not reward was actually received. In the present study, losses represented alternate outcomes: what participants might have won had they collected their money instead of gambled. Thus, my result that the P300 amplitude scaled with what might have been won supports the idea that the P300 reflects an objective representation of reward.

Conclusions. Research on the neural basis of exploration in humans has thus far lacked specific neural markers for this behaviour. Here, I found that decisions to explore or exploit modulated ERP amplitude in the P300 time range in a sequential risk-taking task. Interestingly, this result is in line with a theoretical account that relates ERP amplitudes in the P300 time range to changes in phasic LC-NE activity – changes which are yoked to increased exploratory behaviour (Nieuwenhuis et al., 2005; Aston-Jones & Cohen, 2005). As such, my results (a) suggest that the amplitude of the ERP in the P300 time range is sensitive to decisions to explore or exploit, and (b) relate modulation of the

ERP in the P300 time range to an underlying neural system that is responsible for these changes: the LC-NE system. Of further interest, my results are in line with previous findings (e.g. Yeung & Sanfey, 2004) that demonstrate that the amplitude of the P300 scales to reward magnitude.

Chapter 3: Experiment 2

Abstract

The decision trade-off between exploiting the known and exploring the unknown has been studied using a variety of approaches and techniques. Surprisingly, electroencephalography (EEG) has been underused in this area of study, even though its high temporal resolution has the potential to reveal the time-course of exploratory decisions. I addressed this issue by recording EEG data while participants tried to win as many points as possible in a two-choice gambling task called a two-armed bandit. After using a computational model to classify responses as either exploitations or explorations, I examined event-related potentials locked to two events preceding decisions to exploit/explore: the arrival of feedback, and the subsequent appearance of the next trial's choice stimuli. In particular, I examined the feedback-locked P300 component, thought to index a phasic release of norepinephrine (a neural interrupt signal), and the reward positivity, thought to index a phasic release of dopamine (a neural prediction error signal). I observed an exploration-dependent enhancement of the P300 only, suggesting a critical role of norepinephrine (but not dopamine) in triggering decisions to explore. Similarly, I examined the N200/P300 components evoked by the appearance of the choice stimuli. In this case, exploration was characterized by an enhancement of the N200, but not P300, a result I attribute to increased response conflict. These results demonstrate the usefulness of combining computational and EEG methodologies and suggest that exploratory decisions are preceded by two characterizing events: a feedback-locked neural interrupt signal (enhanced P300), and a choice-locked increase in response conflict (enhanced N200).

Ready, Set, Explore! Event-related Potentials Reveal the Time-course of Exploratory Decisions

Making choices involves managing a trade-off between different decision types, such as risky versus safe, emotional versus logical, and automatic versus deliberative. One such trade-off is deciding whether to exploit previous learning or explore new options (the "explore-exploit dilemma": Gittins & Jones, 1974). Exploration is useful when it reduces our uncertainty about the world and leads to better future outcomes (Behrens, Woolrich, Walton, & Rushworth, 2007). However, in order to experience those positive outcomes, it is also important to exploit what is known, i.e., to forgo exploration in order to make value-maximizing decisions. Humans, like other animals, have evolved neural systems to manage the explore/exploit dilemma, a critical ability in uncertain environments.

Broadly speaking, two neurotransmitters are thought to regulate the explore/exploit dilemma: dopamine and norepinephrine. There is evidence that greater tonic dopamine is associated with exploration (Beeler, 2012; Beeler, Daw, Frazier, & Zhuang, 2010; Frank et al., 2009; Kayser, Mitchell, Weinstein, & Frank, 2015). For example, individuals with greater dopamine levels in prefrontal cortex tend to explore more (Frank et al., 2009). In addition to dopamine, the neurotransmitter norepinephrine has been implicated in exploration (Aston-Jones & Cohen, 2005; Gilzenrat, Nieuwenhuis, Jepma, & Cohen, 2010; Jepma & Nieuwenhuis, 2011; G. A. Kane et al., 2017; Warren et al., 2017). Neurons within the locus coeruleus (LC), the main source of norepinephrine in the brain, show two patterns of firing: phasic bursts of activation in response to task-relevant events, and more gradual tonic (baseline) changes. For example, during a

reversal learning task phasic LC activation to a previous target decreases when that target is no longer rewarding; activation shifts instead to the new target (Aston-Jones, Rajkowski, & Kubiak, 1997). Thus, phasic LC activation is associated with good signal detection and stimulus-response learning in monkeys (Aston-Jones et al., 1997; Clayton, Rajkowski, Cohen, & Aston-Jones, 2004). An increase in tonic LC activation, on the other hand, is associated with poor task performance and high levels of distraction (Aston-Jones & Cohen, 2005). The tonic LC mode may not be maladaptive, however. Converging animal, drug, and pupillometry evidence suggests that high tonic norepinephrine may promote exploration: trying other bandits in a multi-armed bandit task (Jepma & Nieuwenhuis, 2011), leaving a patch while foraging (G. A. Kane et al., 2017), and disengaging from a tone discrimination task when rewards diminish (Gilzenrat et al., 2010).

Investigations into the role of dopamine and norepinephrine in the explore/exploit dilemma have thus far been fruitful. It is therefore surprising that little is known about the electroencephalographic (EEG) correlates of these decisions. This is surprising for two reasons. First, the high temporal resolution of EEG lends itself to the time-course of human decision-making (Heekeren, Marrett, & Ungerleider, 2008). Second, there is evidence that the activity of dopamine and norepinephrine may be indirectly measured via event-related potentials (ERPs) – the averaged EEG response to an event. For example, the reward positivity is an ERP component thought to reflect the effect of phasic dopamine on anterior cingulate cortex (ACC: Holroyd & Coles, 2002; Holroyd & Yeung, 2012). According to Holroyd and Coles (2002), phasic changes in dopamine signify reinforcement learning (RL) prediction errors that modulate the magnitude of the

reward positivity. The ACC, according to this view, is attempting to learn the value of options (sequences of actions: Holroyd & McClure, 2015; Holroyd & Yeung, 2012). Note that the reward positivity is usually thought of as being sensitive to phasic, not tonic, dopamine activity. There is evidence, however, that these two types of dopamine activity are related (Grace, Floresco, Goto, & Lodge, 2007; Niv, Daw, Joel, & Dayan, 2007). Relevant here, the reward positivity is affected by tonic dopamine; greater prefrontal baseline dopamine activity predicts either a decreased reward positivity (Marco-Pallarés et al., 2009) or an increased reward positivity (Foti & Hajcak, 2012).

The reward positivity is actually a special case of another ERP component, the N200 (Baker & Holroyd, 2011; Holroyd, Pakzad-Vaezi, & Krigolson, 2008). While the reward positivity occurs specifically in response to feedback, the N200 is elicited by any task-relevant event, is enhanced for surprising events, and is thought to reflect cortical activity arising from a phasic release of norepinephrine (Hong, Walz, & Sajda, 2014; Mückschel, Chmielewski, Ziemssen, & Beste, 2017; Warren & Holroyd, 2012; Warren, Tanaka, & Holroyd, 2011). Thus, assuming that feedback is unexpected, the amplitude of the reward positivity depends on both reward-related phasic dopamine activity and surprise-related norepinephrine activity. N200 modulation, on the other hand, is tied more to norepinephrine activity alone (Hong et al., 2014; Mückschel et al., 2017; Warren & Holroyd, 2012; Warren et al., 2011). The N200 is often followed by another norepinephrine-dependent ERP component called the P300 (Nieuwenhuis et al., 2005). Like the N200, the P300 is enhanced for infrequent and/or task-relevant stimuli and has also been linked to the phasic release of norepinephrine (Murphy, Robertson, Balsters, & O'connell, 2011; Nieuwenhuis et al., 2005, 2011). In summary, it may be possible to

track phasic changes in norepinephrine via the N200 and P300, and phasic changes in dopamine via the reward positivity.

Previous work on the EEG correlates of exploration and exploitation is sparse. Early work by Bourdaud, Chavarriaga, Galan, and Millan (2008) analyzed EEG recorded from participants performing a four-armed bandit task (Daw et al., 2006). Bourdaud and colleagues (2008) asked simply whether or not pre-response EEG was capable of differentiating decisions to explore and exploit. To answer this question, they showed that machine learning could successfully classify trials as explorations and exploitations based on the frequency content of EEG at frontal and parietal sites (also see Tzovara et al., 2012). Consistent with this result, Cavanagh, Figueroa, Cohen, and Frank (2011) observed a correlation between uncertainty and response-locked medial frontal theta power that was positive for exploratory decisions, but negative for exploitative decisions. Finally, in Experiment 1 I observed an enhancement of the P300 component at the time of exploratory responses compared to exploitative responses during a sequential risk-taking task called the Balloon Analogue Risk Task (BART: Lejuez et al., 2002). Responses and feedback occur simultaneously in the BART, though, so it is unclear which event (response or feedback) led to the P300 effect observed in Experiment 1.

Our goal here was to use EEG to affirm the roles of dopamine and norepinephrine in managing the explore/exploit dilemma. To do this, I examined ERP components locked to two events in a two-armed bandit task: the (feedback-locked) reward positivity/P300 and the (choice-locked) N200/P300. I hypothesized that the enhanced tonic dopamine activity associated with exploration would, when combined with the usual reward-related phasic dopamine activity, affect the reward positivity (either

enhance it or reduce it). In light of conflicting reports (Foti & Hajcak, 2012; Marco-Pallarés et al., 2009) I did not hypothesize as to which decision type would elicit the larger reward positivity, only that there would be a difference. To generate my N200/P300 hypothesis, I considered two somewhat conflicting viewpoints on the role of norepinephrine in regulating the explore-exploit dilemma. As mentioned, previous studies have suggested that the tonic mode of LC activity (low task performance/high distraction) may facilitate exploration, while the phasic mode of LC activity (high task performance/low distraction) may facilitation exploitation (Gilzenrat et al., 2010; Jepma & Nieuwenhuis, 2011; G. A. Kane et al., 2017; Nieuwenhuis et al., 2005; Warren et al., 2017). Based on this interpretation, and since the N200/P300 complex is associated with phasic norepinephrine, one might predict enhancements of those components around the time of exploitations. However, Dayan and Yu (2006) interpreted phasic norepinephrine as a neural interrupt signal, signaling a need to update one's model of the world – or context – and to switch strategies accordingly (also see: Bouret & Sara, 2005; Yu & Dayan, 2005). Indeed, Donchin's (1981) context-updating hypothesis of the P300 can be considered a precursor to the LC-NE hypothesis of the P300 (the LC-NE P3 theory: Nieuwenhuis et al., 2005) as both highlight the motivational/task significance of a stimulus. Under this view, a phasic burst of NE (and large, concomitant P300) to feedback could signal a need to explore the environment, provided that exploitation had been the dominant strategy, such as in a reversal learning task in which reversals are relatively rare (as seen in Aston-Jones et al., 1997, for example). For example, in the BART, used in Experiment 1, explorations were rare, and associated with a greater P300 compared to exploitations. For these reasons, I hypothesized that explorations in my task

would be associated with an enhancement of the feedback-locked P300 and choice-locked N200/P300.

Methods

Participants. Twenty-three university-aged participants (9 male, 1 left-handed, M_{age} = 22, 95% CI [21, 23] with no known neurological impairments and with normal or corrected-to-normal vision took part in the experiment. Participants who did not meet a pre-set accuracy threshold of 60% (as defined below in the Data Analysis subsection) were excluded from the analysis. In total, five participants were excluded from the analysis due to poor performance (mean accuracies of 56%, 51%, 50%, 48%, and 50%). All of the participants were volunteers who received credit in an undergraduate course for their participation. The participants provided informed consent approved by the Health Sciences Research Ethics Board at Dalhousie University.

Apparatus and procedure. Participants were seated comfortably 75 cm in front of a computer display and used a standard USB keyboard to perform a computerized gambling task, written in MATLAB (Version 7.14, Mathworks, Natick, USA) using the Psychophysics Toolbox Extension (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997). Participants received both verbal and written instructions and were encouraged to maintain a central fixation and to minimize head movements and eye blinks. Participants were told that the goal of the task was to win as many points as possible.

The experimental task was a two-choice gambling game (i.e., a two-armed bandit: R. S. Sutton & Barto, 2018). Comprised of two one-armed bandits, or slot machines, my two-armed bandit required that participants choose between one of two possible gambles, represented by coloured squares presented to the left and right of a central fixation point.

Participants were told beforehand that one of the choices had a higher average win payout than the other. The loss amounts associated with each choice were equivalent. Thus, participants were gambling based on the win payouts; the overall proportion of wins to losses was not task relevant. At the beginning of a trial, a 1.1 cm fixation cross subtending 0.84 degrees of visual angle was presented for 900 – 1100 ms. Subsequent to this, two coloured squares appeared, each 2.8 cm across and subtending 2.14 degrees of visual angle, equidistant on either side of the fixation cross. The squares were 11.3 cm apart, center-to-center, or 8.62 degrees of visual angle. After the squares were presented for 900–1100 ms, the fixation cross changed colour to cue participants to respond by selecting either the left square (‘a’ key) or right square (‘l’ key). If participants responded too early, points were deducted from the total won. Similarly, participants were told that points would be deducted if they responded too late (after two seconds). This ensured that all valid responses occurred within a 0–2000 ms window following the go cue.

After a valid response, the squares were removed, leaving only a fixation cross on the display for 400–600 ms. Participants then viewed a feedback stimulus indicating the amount of points won or lost on that trial for 1000 ms. As explained to participants in the instructions, half of the trials resulted in a win, and half of the trials in a loss; specific outcomes were determined by a pseudorandom number generator. Thus, the chance of winning any given trial was 50%. This ensured that a similar number of win and loss trials would be available for later analysis (Holroyd & Krigolson, 2007; Krigolson, 2017). Loss trials resulted in a 5-point deduction, regardless of which square was selected. Wins, on the other hand, always paid a positive amount that was dependent on which square was selected. Each square, or bandit, paid amounts selected from Gaussian

distributions with identical variances ($\sigma^2 = 1$), but with different means. A block consisted of 20 gambles, or trials, after which a new block began with two new random colours and two new reward distributions. Participants were told that after each block the squares reset (new colours and payouts), and that they then had to relearn which square was the higher-paying choice in order to win as much as possible. Participants completed 50 blocks in total and were given a self-paced rest break every 10 blocks.

To ensure that the task presented a similar level of difficulty for all participants, payout distributions were initially quite different (means of 6 and 12 points). The mean of the lower valued square was increased by one after every block as long as participants were able to achieve an accuracy of 80% (defined as selecting the higher valued option in the second half of a block at least 8 times out of 10). The payout distributions were fixed once participant accuracy dropped below 80%, i.e. once an appropriate level of difficulty for that participant was achieved. See Figure 8 for timing details.

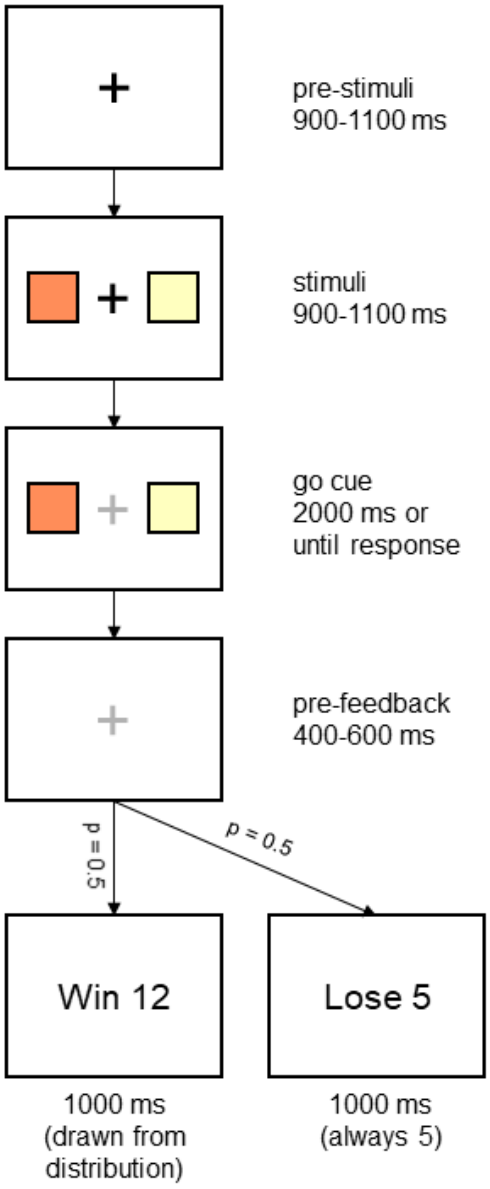


Figure 8. Two-armed bandit task. Participants were given feedback after selecting one of two coloured squares, or bandits. On average, one bandit paid more points than the other. Losses were always the same magnitude (5 points).

Data collection. The experimental program recorded participant choice (higher or lower valued square) and response time. The EEG was recorded from 64 electrode locations using Brain Vision Recorder software (Version 1.20, Brain Products, GmbH,

Munich, Germany). The electrodes were mounted in a fitted cap with a standard 10-20 layout and were recorded with an average reference built into the amplifier. The vertical and horizontal electrooculograms were recorded from electrodes placed above and below the right eye and on the outer canthi of the left and right eyes. Electrode impedances were kept below 20 kΩ. The EEG data were sampled at 1000 Hz and amplified (Quick Amp, Brainproducts, GmbH, Munich, Germany).

Computational model. My analysis depended on classifying participant decisions as either exploitations or explorations. To achieve this, I modeled each participant's responses, trial by trial. My model, used previously in Krigolson, Hassall, and Handy (2014), maintained a value for each bandit stimulus on each trial: $v_t(1)$ and $v_t(2)$. The probability on trial t of selecting stimulus i (that is, the likelihood of making an action a_i) was computed as per the *softmax* equation:

$$P_t(a_i) = \frac{e^{v_t(i)/\tau}}{e^{v_t(1)/\tau} + e^{v_t(2)/\tau}}$$

where τ (temperature) determined the degree of bias towards choosing high-valued stimuli (greater bias for lower τ). On each "win" trial, following feedback R_t, a prediction error δ_t was generated for the selected stimulus s according to:

$$\delta_t = R_t - v_t(s)$$

The value of the chosen bandit s was then updated using the following learning rule:

$$v_{t+1}(s) = v_t(s) + \alpha\delta_t$$

in which prediction errors were scaled by α. The value of the unselected stimulus was unchanged. Losses, designed to be uninformative in this task, did not result in any prediction error computation or model update. (Recall that losses occurred with 50%

probability, regardless of bandit choice, and only ever resulted in a loss of 5 points.) To support this design choice, I compared my model to one in which both wins and losses prompted model updates.

The temperature and learning rate were tuned for each participant. These parameters (τ, α) were tuned using the MATLAB function fmincon (Optimization Toolbox, Release 2018a, Mathworks, Natick). Specifically, I constructed an objective function (the function to be minimized) as the negative log-likelihood of a participant's set of responses. Log-likelihood was computed as:

$$\sum_{t} \log(P_t(a_s))$$

where $P_t(a_s)$ was the softmax probability associated with the selected bandit s on trial t. This was a post hoc fit criterion – the goal was to maximize the likelihood of the actual participant responses (i.e., to predict the one-step-ahead predictions: Ahn, Busemeyer, Wagenmakers, & Stout, 2008).

To reiterate: model tuning was done for each participant. Thus, the model-tuning procedure generated learning parameters (τ, α) for each participant. Additionally, I classified trials as exploitations or explorations using the softmax result on each trial. Trials on which the participant made the less likely response, according to the softmax equation, were classified as explorations. All other trials – trials in which the higher-probability response was made – were classified as exploitations. As expected, there were more explorations early in learning (Figure 9). These trial classifications (exploit/explore) were used to drive the ERP analysis – see below.

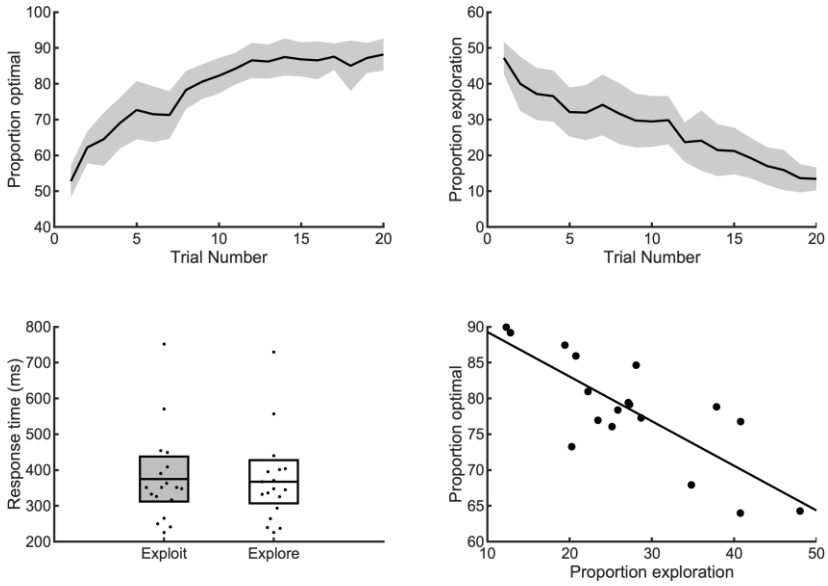


Figure 9. Behavioural results. Accuracy (top left) was defined as the proportion of trials that participants selected the higher-valued option. A computational model classified each trial as an exploration or exploitation (mean exploration proportion: top left). Mean response time did not differ by decision type (bottom left, individual means also shown). In this task, participants who explored more tended to perform worse (bottom right). The shaded regions and error bars show 95% confidence intervals.

Data analysis.

Behavioural data. For each participant and trial (1-20) I computed the mean proportion of times, across all blocks, that the optimal choice was made (i.e., the higher-valued bandit was chosen). I also computed the mean of this proportion across all trials and participants. Similarly, I computed the mean proportion of explorations for each trial (1-20) from each participant's trained model. I then computed the average and standard

deviation of this exploration proportion across all participants. Finally, I computed the mean response times for each decision type (exploit/explore).

Electroencephalographic data. EEG data were downsampled to 250 Hz, filtered through a (0.1 Hz – 30 Hz pass band) phase shift-free Butterworth filter (60 Hz notch), and rereferenced to the average of the two mastoid channels. Next, ocular artifacts were removed using independent component analysis. Subsequent to this, and for each event of interest (stimulus and feedback presentation), 800 ms epochs of EEG data were constructed from 200 ms prior to 800 ms following event onset. All trials were then baseline corrected using a 200 ms pre-event window. Finally, trials in which the change in voltage in any channel exceeded 10 μV per sampling point or the change in voltage across the epoch was greater than 100 μV were discarded. On average, I removed 6.3% of the stimulus-locked epochs (95% CI [3.8, 8.9]) and 6.6% of the feedback-locked trials (95% CI [3.7, 9.5]). My hypothesis concerned two events: feedback given just prior to a decision to exploit/explore (trial N-1), and the choice stimuli that are exploited/explored (trial N). Below I describe how I quantified ERPs for these two events.

Feedback-locked electroencephalographic data. To quantify the reward positivity, I averaged the feedback-locked EEG for each participant, channel, feedback condition (win/loss), and decision type (exploit/explore). I then constructed difference waveforms by subtracting the average loss waveforms from the average win waveforms (Krigolson, 2017). To identify a window of analysis, I constructed a "grand-grand" average difference waveform (Kappenman & Luck, 2016) by collapsing across both participant and decision type (exploit/explore). I then identified a window of interest by locating the peak of this difference waveform (maximum voltage, across all timepoints and scalp

locations), and chose as a half-interval the time on the leading edge of the peak at which 75% of the maximum voltage was reached. Thus, the reward positivity was defined as the mean voltage from 240 to 296 ms post feedback at electrode Cz (See Figure 10). A reward positivity score was computed for each participant and decision type (pre-exploit/pre-explore). A similar procedure was followed for the P300, except that the grand-grand average also collapsed across feedback type (i.e., it was the average response to all feedback). The peak of the P300 was defined as the maximum positive deflection, across all timepoints and scalp locations, and the half-interval was defined as the point on the leading edge of the waveform at which 75% of the maximum voltage was reached. This resulted in a P300 defined as the mean voltage from 308-420 ms post feedback at electrode POz (see Figure 11). Thus, a P300 score was computed for each participant, feedback type (win/loss), and decision type (pre-exploit/pre-explore).

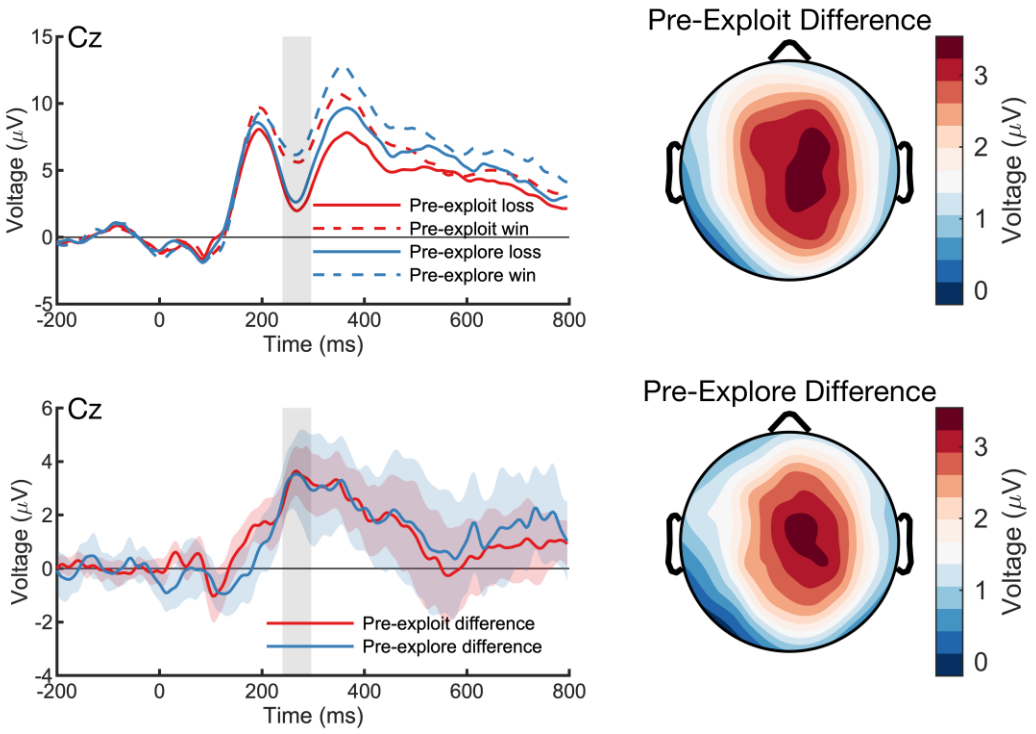


Figure 10. Reward positivity preceding decisions to exploit and explore. Conditional waveforms are show in the top left panel, and difference waves (win minus loss) are shown in the bottom left panel. The vertical shaded rectangle indicates the analysis window. The shaded regions around each difference wave reflect 95% confidence intervals.

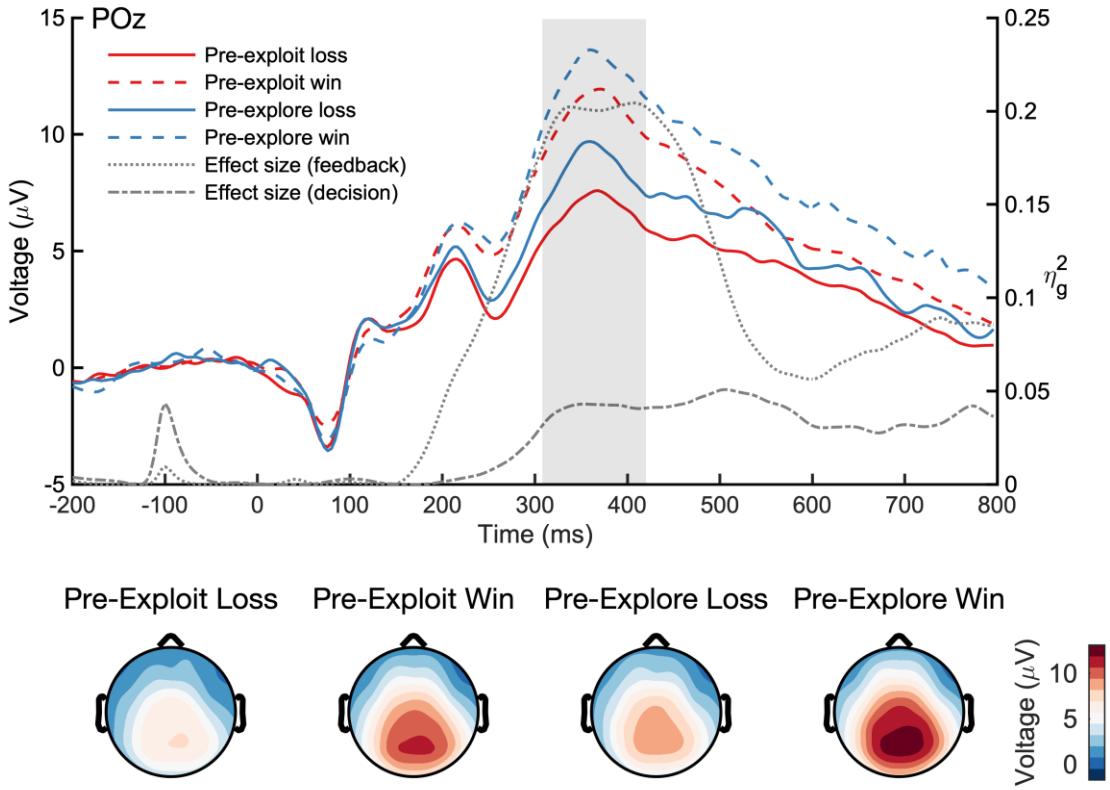


Figure 11. Feedback-locked P300 waveforms and scalp distributions preceding decisions to exploit and explore. The shaded rectangle indicates the analysis window. The grey lines show effect size (η_g^2) for each main effect (feedback: loss/win, decision: exploit/explore) computed on a moving mean (window length: 100 ms).

Choice-locked electroencephalographic data. Preceding a decision to exploit or explore, participants were shown the choice stimuli, or bandits. To analyze the ERPs locked to the bandits, I averaged the choice-locked EEG for each channel and decision type (exploit/explore), for each participant. Only trials with valid behavioural responses were included. To identify the P300 time range, I followed a similar procedure as for the feedback-locked analysis; I collapsed across all participants and conditions (exploit/explore) and found the time/location of greatest voltage. I then took 75% of the

leading edge as the half interval. This resulted in a P300 defined as the mean voltage from 312-392 ms post bandits at electrode POz (i.e., a P300 score for each participant and decision type). My N200 analysis was exploratory, as there was no obvious N200 peak at any anterior electrode site that could be identified when I collapsed across decision type. Instead, I identified the time/location of the greatest difference between my average explore waveform and my average exploit waveform. An interval from 196-244 ms post bandits at electrode FC3 was identified as the time/location of greatest difference (i.e., where the 95% confidence intervals of the difference wave did not overlap with zero). There are two caveats to this exploratory analysis. First, my N200 definition was biased because it was defined using my conditions of interest (exploit/explore). Second, this time range overlapped with a centrally-located P200 ERP (although the difference was maximal at a frontal site – see Figure 12).

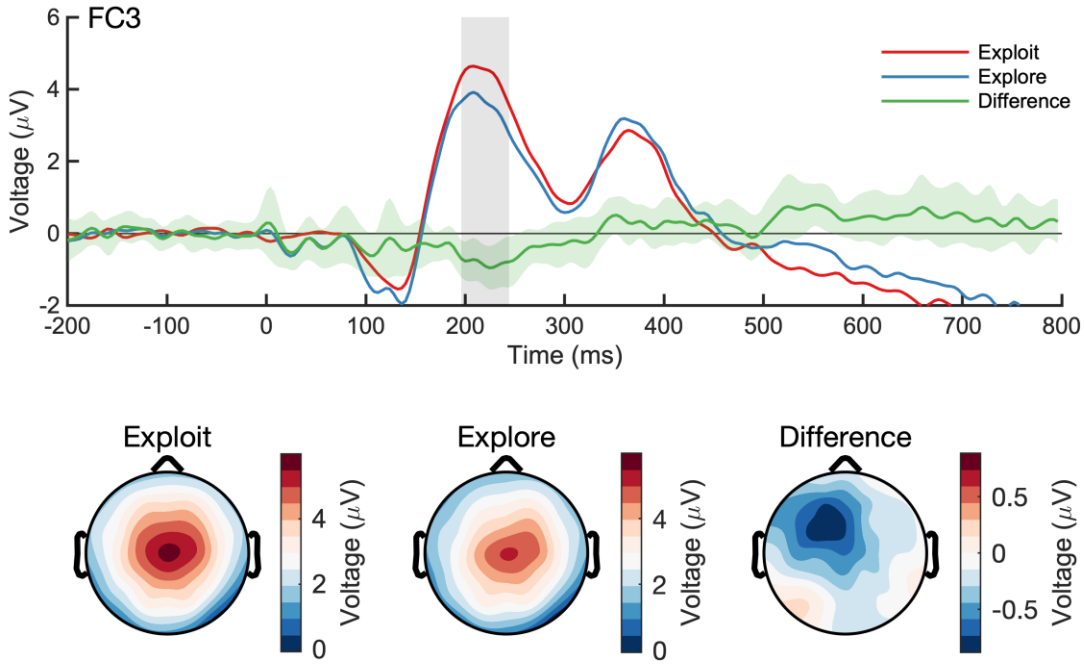


Figure 12. Choice-locked N200 waveform and scalp distributions. The vertical shaded rectangle indicates the analysis window. The shaded region around the difference wave reflects a 95% confidence interval.

Single-trial analysis. To further investigate the relationship between the feedback-locked P300 and an upcoming decision to exploit or explore, I computed a single-trial EEG analysis. A P300 score was generated for each participant and trial using the same procedure as in my feedback-locked ERP analysis (i.e., I averaged the post-feedback voltage from 308-420 ms at electrode POz). I then calculated, for each participant, a regression line relating the trial-by-trial P300 score to the softmax probability of the upcoming trial decision. If exploration is associated with greater P300 scores, then I ought to see a negative relationship between P300 magnitude and softmax probability.

(Recall that I defined exploration as a decision with a less-than-maximal softmax probability; thus, less-likely decisions ought to be preceded by larger P300s.)

Inferential statistics. The existence of a reward positivity, defined as a difference score, was tested using a single-sample t-test (Holroyd & Krigolson, 2007; Rodríguez-Fornells, Kurzbuch, & Münte, 2002). Between decision conditions (pre-exploitation, pre-exploration) the reward positivity's were compared using a paired samples t-test. The feedback-locked P300 was analyzed using a 2 (feedback: win, loss) by 2 (decision type: pre-exploit, pre-explore) repeated-measures ANOVA. The choice-locked N200 scores were compared using a paired-samples t-test, as were the choice-locked P300 scores. Finally, participant slopes in my single-trial analysis were compared against zero with a single-sample t-test. For all t-tests, I computed Cohen's *d* according to:

$$d = \frac{M_{\text{diff}}}{s_{\text{diff}}}$$

where M_{diff} was the difference score mean and s_{diff} was the difference score standard deviation (or in the case of the reward positivity, the mean and standard deviation of the ERP score itself; see Cumming, 2014). For the ANOVA, I computed two different effect-size measures: η_p^2 and η_g^2 (Lakens, 2013; Olejnik & Algina, 2003).

Results

Modelling data. My greedy model generated an average negative log-likelihood of 787, 95% CI [602, 973]. Using softmax for action selection resulted in an improved model fit - a negative log-likelihood of 368, 95% CI [323, 413], $t(17) = -5.94$, $p < .001$,

Cohen's d = -1.40. The average tuned softmax model parameters were as follows: τ = 0.07, 95% CI [-0.04 0.17] and α = 0.14, 95% CI [-0.02, 0.30].

Behavioural data. The mean accuracy (all trials) was 78%, 95% CI [75, 82]. The mean proportion of explorations (all trials) was 20%, 95% CI [17, 24]. Mean accuracy was correlated with mean proportion of explorations, $r(16) = -0.79$, $p < .001$. Explorations and exploitations did not differ in response time exploitations: 376 ms, 95% CI [327, 424], explorations: 371 ms, 95% CI [324, 418]), $t(17) = 1.44$, $p = .16$, Cohen's d = 0.29. See Figure 9 for behavioural results.

Electroencephalographic data.

Reward positivity. Single-sample t-tests revealed reward positivities prior to decisions to exploit (3.25 μV, 95% CI [1.89 4.60], $t(17) = 9.74$, $p < .001$, Cohen's d = 2.30) and decisions to explore (3.27 μV, 95% CI [2.56 3.98], $t(17) = 5.04$, $p < .001$, Cohen's d = 1.19. A paired-samples t-test comparing the pre-explore reward positivity to the pre-exploit reward positivity revealed no effect of decision type: $t(17) = -0.05$, p = .96, Cohen's d = -0.01 (Figure 10).

Feedback-locked P300. Mean P300 was greater for wins than losses prior to both decisions to exploit (pre-exploit loss: 6.76 μV, 95% CI [4.99, 8.54]; pre-exploit win: 10.90 μV, 95% CI [8.91, 12.90]) and decisions to explore (pre-explore loss: 8.55 μV, 95% CI [6.48, 10.61]; pre-explore win: 12.51 μV, 95% CI [10.20, 14.81]). A 2X2 ANOVA with feedback (loss, win) and decision (exploit, explore) as repeated measures revealed main effects of feedback, $F(1,18) = 47.28$, $p < .001$, $\eta_p^2 = .736$, $\eta_g^2 = .205$, and, importantly, decision type, $F(1,17) = 18.43$, $p < .001$, $\eta_p^2 = .520$, $\eta_g^2 = .043$. There was

also no significant interaction between feedback and decision, $F(1,17) = 0.14$, $p = .86$, $\eta_p^2 = .008$, $\eta_g^2 < .001$ (Figures 12 and 13).

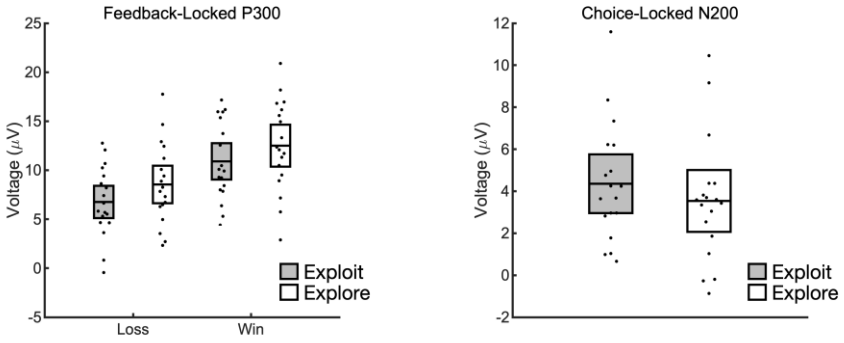


Figure 13. Summary of results. There was a main effect of upcoming decision type on the feedback-locked P300 (left) and choice-locked N200 (right). Error bars show the 95% confidence intervals.

Choice-locked N200. The choice-locked N200 magnitude was 4.36 μV, 95% CI [2.92, 5.79] on exploitation trials and 3.53 μV, 96% CI [2.02, 5.04] on exploration trials. A paired-sampled t-test indicated a statistically-significant difference, $t(17) = -3.24$, $p = .005$, Cohen's $d = -0.76$ (Figure 12).

Choice-locked P300. The bandit-locked P300 magnitude was 5.06 μV, 95% CI [3.76, 6.36] on exploitation trial and 5.26 μV, 96% CI [3.84, 6.68]. A paired-sampled t-test indicated no statistically-significant difference, $t(17) = 0.73$, $p = .47$, Cohen's $d = 0.17$ (Figure 12).

Single-trial analysis. The mean slope of a regression line relating P300 magnitude to softmax probability was -0.059, 95% CI [-0.076 -0.042]. In other words, the P300 dropped by 0.059 μV for every percent of softmax probability. A single-sample t-

test showed the slope to be significantly different from zero, $t(17) = -7.31$, $p < .001$, Cohen's $d = -1.72$ (Figure 14).

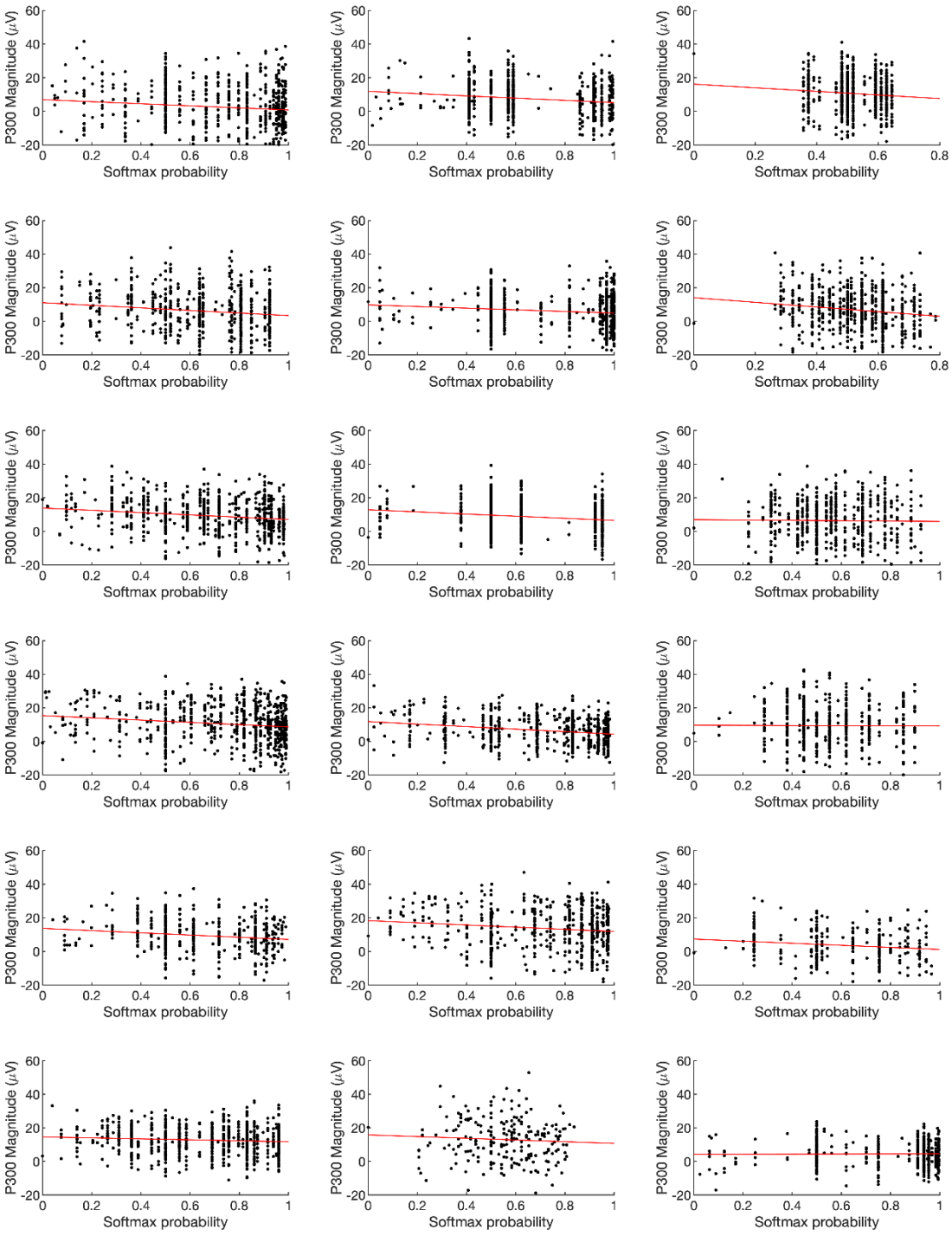


Figure 14. Relationship between each participant's trial-to-trial P300 and the model-generated likelihood (softmax) of the upcoming decision.

Discussion

Our results suggest the involvement of two neural systems when transitioning from an exploitative to an exploratory mode of decision-making. First, feedback-locked phasic activity of the LC-NE system is associated with decisions to explore. Second, exploratory decisions may elicit enhanced response conflict, processed within ACC. These two neural systems – phasic LC-NE activity and conflict-related ACC activity – are indexed by enhancements to the P300 and N200 ERP components, respectively.

Behaviourally, my participants learned to pick the optimal bandit in a two-armed bandit task. A model, fit to each individual's behaviour, determined which responses were exploitations and which were explorations – that is, on which trials the higher-valued bandit was chosen, and on which trials the alternative was chosen. My task was stationary – outcome contingencies never changed within a block – and more exploration was associated with poorer performance (see correlation in Figure 9). In general, exploration rate is driven by a combination of individual differences (e.g. Frank et al., 2009) and context. Sequential decision problems, such as bandit tasks, may be stationary or non-stationary, and may have any number of choices/actions. A full review of the decision-making literature is beyond the scope of this study, but it is worth mentioning a couple of relevant examples. Jepma and Niewenhuis (2011) used a four-armed bandit with continuously drifting average rewards, similar to Daw et al. (2006). They reported a mean exploration rate of 31%. Blanchard and Gershman (2018) used a two-armed bandit with only occasional reversals (5% of trials). Their participants' average exploration rate dropped from 85% to 12% over a 50-trial block. The point of these examples is to

illustrate that exploration rate can vary greatly across experiments, and that my mean exploration rate – which dropped from 47% to 13% over a 20-trial block – is in line with previous bandit studies (Figure 9).

Neural response to feedback. By categorizing choices as either explorations or exploitations, I was able to examine the neural response to feedback preceding each decision type. I observed an ERP difference at a scalp location and time range consistent with the P300, an ERP component that, like the N200, is thought to relate to a phasic release of norepinephrine (Nieuwenhuis et al., 2005). The neurotransmitter norepinephrine has featured heavily in several theories of decision making. Relevant here is the view that norepinephrine indexes a neural interrupt – a signal that one's current model of the world might be erroneous, potentially requiring a strategy switch (Bouret & Sara, 2005; Dayan & Yu, 2006; Yu & Dayan, 2005). For example, imagine a participant in the current study trying to win as many points as possible by selecting which of two slot machines to play (the two-armed bandit task). The participant is told that one of the bandits yields a greater average reward than the other. Initially, the participant continues to select one of the bandits because the payouts seem high. At some point, however, the participant decides explore the other option. I argue that this switch – from deciding to exploit one option, to deciding to explore the other – is one example of the neural interrupt discussed by (Bouret & Sara, 2005; Dayan & Yu, 2006; Yu & Dayan, 2005). Supporting this assertion is my observation that feedback preceding decisions to explore

(and less-likely decisions, in general) elicited an enhanced P300 compared to feedback preceding decisions to exploit.

Our other feedback-related hypothesis involved the reward positivity, an ERP component thought to index the phasic release of dopamine (Holroyd & Coles, 2002). Based on previous research linking tonic dopamine and exploration (and other work suggesting that tonic dopamine effects the reward positivity) I hypothesized an effect of decision type (exploit/explore) on the reward positivity. However, although I observed a robust reward positivity for both exploitations and explorations, I found no statistically-significant effect of decision type (Figure 10). One possible confound here is component overlap. Because the P300 and reward positivity components overlap in time, examining the reward positivity is problematic when P300 effects are also present (such as in the present study – see Figures 11 and 12). P300 contamination is usually due to frequency effects (e.g., when losses are less frequent than wins) but it's possible that other events affecting the P300 – such as the neural processes associated with exploration – may have hindered my reward positivity investigation. I suggest that to properly examine the role of the reward positivity in the explore/exploit trade-off would require a task for which the P300 is unaffected by feedback valence (as it is here). See Krigolson (2017) for a discussion of component overlap and other methodological considerations.

Neural response to bandits. Following feedback, participants were presented with the choice stimuli again, i.e. the bandits. I observed that the choice-locked N200 ERP component was greater for explorations compared to exploitations. Note, however, that this analysis was exploratory – the observed N200 effect was only apparent after an examination of the difference waveform, and appeared to overlap with a P200

component. Furthermore, I observed the predicted enhancement of the bandit-locked N200 prior to explorations, but no enhancement of the bandit-locked P300. These components often co-occur and are referred to as the N200/P300 (or N2/P3) complex (Duncan-Johnson & Donchin, 1977). According the modified LC-P3 theory, both the N200 and the P300 depend on phasic norepinephrine (Hong et al., 2014; Mückschel et al., 2017; Warren & Holroyd, 2012; Warren et al., 2011). In particular, the modified LC-P3 theory suggests that phasic bursts of norepinephrine have two effects: an initial cortical enhancement between 200 and 300 ms, and a later cortical impairment between 300 and 600 ms. In other words, phasic norepinephrine enhances both the N200 and the P300, but through different mechanisms (N200: abundance, P300: depletion). To be consistent with the modified LC-P3 theory, I must conclude that my bandit stimuli did not elicit a greater phasic release of norepinephrine prior to decisions to explore compared to decisions to exploit. If they had, I would have observed an exploration-dependent enhancement of both the N200 and the P300. I thus left with the following question: what could elicit an enhancement of the N200 but not the P300?

To answer this question, I turned to the cognitive control and conflict monitoring literature. Cognitive control is a set of processes that enable humans to flexibly adapt to new situations and goals. According to the conflict-monitoring hypothesis, the need for cognitive control is triggered via the detection of information processing conflict (Botvinick, Braver, Barch, Carter, & Cohen, 2001). For example, incongruent stimuli in the Stroop task activate two competing responses – reading the word and naming the colour – thus eliciting response conflict and a need for control (Botvinick et al., 2001; Stroop, 1935). In the brain, conflict is processed within the ACC, which generates a

conflict-dependent N200; incongruent stimuli in a flanker task elicit an enhanced N200 relative to congruent stimuli (Yeung, Botvinick, & Cohen, 2004). Tasks that elicit a conflict-dependent N200 tend to also elicit a P300, but Enriquez-Geppert, Konrad, Pantev, and Huster (2010) showed that the N200 mostly indexes conflict, while the P300 mostly indexes motor inhibition. Thus, an N200 effect in the absence of a P300 effect is possible provided that there is response conflict but not motor inhibition.

I speculate that my exploration trials prompted response conflict because of the simultaneous activation of two responses: the computationally valuable exploitative option, and the computationally less valuable exploratory option. Here, exploitations represented the prepotent response and, like go trials in a go/no-go task, elicited low response conflict. Thus, the bandit-locked N200 was enhanced for explorations (high conflict) relative to exploitations (low conflict). I observed no such enhancement of the bandit-locked P300, however. As Enriquez-Geppert et al. (2010) showed, this pattern of results is possible for tasks that elicit response conflict but not motor inhibition. Since motor inhibition is presumably most relevant around the time of the response, this seems a reasonable characterization of my bandit-locked results; my participants were not cued to respond until around one second after the appearance of the bandits. Thus, the appearance of my bandits impacted the N200 but not the P300.

A response-conflict interpretation of my bandit-locked N200 result aligns with work suggesting that the ACC (the neural generator of the conflict-dependent N200) is involved with decisions to explore or exploit only insofar as it is more active during difficult choices. Shenhav, Straccia, Cohen, and Botvinick (2014) pointed out that foraging experiments tend to confound foraging value – the value associated with

exploration – with choice difficulty (i.e., conflict). As the value of switching approaches the value of staying, and exploration becomes more likely, choice difficulty increases. When foraging value and choice difficulty are dissociated, ACC activity tends to track the latter (Shenhav et al., 2014). It is therefore problematic to conclude that the ACC has a special role in foraging beyond the processing of choice difficulty (e.g., in tracking foraging value: Kolling, Behrens, Mars, & Rushworth, 2012). The enhanced N200 I observed just prior to decisions to explore is consistent with the view that the ACC processes choice difficulty during explore/exploit decisions. It may also be consistent with the view that the ACC processes foraging value (Kolling et al., 2012), as I did not dissociate foraging value and choice difficulty. However, a foraging-value account of my N200 data does not seem as promising as a conflict-monitoring account given the amount of literature linking the ACC-generated N200 to response conflict (Baker & Holroyd, 2011; Enriquez-Geppert et al., 2010; Nieuwenhuis, Yeung, Wildenberg, & Ridderinkhof, 2003; Yeung et al., 2004).

Conclusions. By examining ERPs to two events – feedback and choice stimuli - I demonstrate the contribution of three neural systems to the explore-exploit dilemma. First, phasic activity of the LC-NE system, as indexed by a feedback-locked P300, plays a critical role in triggering a switch from exploitative to explorative decision making. Conversely, phasic midbrain dopamine does not appear to play this same role; the reward positivity, a dopamine-driven RL signal, did not predict decision type. Finally, the period just prior to a decision to explore appears to involve response conflict; the bandit-locked N200, a neural conflict signal originating in ACC, was enhanced prior to exploratory decisions.

Chapter 4: Experiment 3

Abstract

Decision making often involves striking a balance between exploiting previous experience and exploring the unknown. Much is still unknown about the neural basis of this trade-off, although recent progress has benefited from the combination of computational modelling and neuroimaging. The electroencephalographic (EEG) correlates of exploration have been underexplored, however, despite the usefulness of this methodology in studying other aspects of decision making. Here I replicated and extended previous work showing an enhancement of the human event-related potential (ERP, or average event-locked EEG) following decisions to explore. In particular, my participants completed three versions of a multi-armed bandit task, across which I varied the total number of response options (4, 9, or 16). I then assessed the fit of three computational models to my participants' choices, and established support for both a reinforcement learning approach and a heuristic approach called win-stay lose-shift (WSLS). Using my computational models, I classified trials as either exploitations or explorations, and examined both behavioural and neural data related to each decision type. Decisions to explore were less frequent, slower, and less rewarding than decisions to exploit. As the number of choices increased across tasks, so did the rate of exploration. This allowed us to examine the effect of exploration rate on the P300, an ERP component previously linked to decisions to explore. My results show that the exploration-related P300 diminishes with increasing exploration rate. I interpret the exploration-related P300 as a neural interrupt signal that helps shift away from one mode of decision-making (exploration) to another (exploitation).

Feedback Processing Depends on Rate of Exploration

Retailers know that offering more options can help consumers attain their purchase goals. On the other hand, offering more options can lead to *choice overload*: greater cognitive demand, and consumer dissatisfaction (Chernev, Böckenholt, & Goodman, 2015). When choosing multiple times from among several options, consumers are faced with an additional problem: to go with "the usual" or try something new. The explore-exploit dilemma (Gittins & Jones, 1974) describes the trade-off between making a value-maximizing choice (exploiting) and trying to learn something about the world (exploring).

Numerous behavioural and neuroimaging studies have helped uncover how humans and other animals manage the explore-exploit dilemma (Cohen et al., 2007; Daw et al., 2006; Frank et al., 2009; Hayden, Pearson, & Platt, 2011; Kolling et al., 2012). Previous electroencephalographic research examining the explore-exploit dilemma, though somewhat sparse, suggests that decisions to explore are related to phasic changes in norepinephrine (NE), as measured by changes in the electroencephalogram (Experiments 1 and 2). Specifically, I showed in Experiment 1 that decisions to explore are associated with an enhanced P300, a component of the human event-related brain potential (ERP) thought to be related to the release of NE from locus coeruleus (the LC-NE P300 theory: Nieuwenhuis et al., 2005; Vazey, Moorman, & Aston-Jones, 2018). According to the LC-NE theory of the P300, this component is tied specifically to phasic NE activity (as opposed to tonic, or baseline activity). The functional significance of this signal is to facilitate a response to the outcome of an internal decision process, such as during target detection. For example, monkeys show enhanced phasic NE activity to

rewarding stimuli during stimulus-response learning. During a reversal phase (when reward contingencies shift), they show less phasic NE activity to old targets and greater NE activity to new targets (Aston-Jones et al., 1997).

In humans, reward reversals are associated with an enhanced P300 (H. W. Chase, Swainson, Durham, Benham, & Cools, 2010; von Borries, Verkes, Bulten, Cools, & de Bruijn, 2013). Broadly speaking, the P300 is affected by both external factors, such as stimulus frequency, and internal factors related to the updating of working memory (Polich, 2007). These two aspects of the P300 have led to the identification of two subcomponents: the P3a and the P3b. Relevant here is the P3b subcomponent, often referred to simply as the P300, as this is the subcomponent that is thought to be tied to the LC-NE system (Nieuwenhuis et al., 2005; Polich, 2007). I previously used the LC-NE theory of the P300 to predict an exploration-related enhancement of this component. I had participants complete the Balloon Analogue Risk Task (BART: Lejuez et al., 2002) while EEG data were recorded. Responses in the BART were classified as either exploitations (fast pumps) or explorations (slow pumps). Successful pumps resulted in a balloon inflation, which elicited a P300 that was larger for explorations than exploitations, a result I attributed to exploration-related NE activity (Experiment 1). However, the P300 is a fairly ubiquitous ERP component that is affected by several factors, some of which were discussed above.

Relevant here is the fact that it is possible for the magnitude of the feedback-locked P300 to depend on the likelihood of a certain decision *type*. For example, feedback following high-risk decisions tends to elicit larger P300s compared to feedback following low-risk decisions. This effect is modulated by the relative proportion of risky

to non-risky decisions – if fewer risks are taken, the risk-related P300 is enhanced (Zheng, Li, Wang, Wu, & Liu, 2015). Likewise, high sensation seekers (who tend to take more risks) show a reduced P300 to feedback following risky choices compared to low sensation seekers (Zheng & Liu, 2015). Finally, when deciding whether or not to buy several items, a counter-intuitive decision – buying an over-priced item, or not buying an under-priced item – elicits a larger P300 compared to an intuitive decisions (Gajewski, Drizinsky, Zülch, & Falkenstein, 2016). In other words, the P300 is enhanced for less common decision types: risky compared to non-risky, and counter-intuitive compared to intuitive. Here I ask whether the same is true of exploratory decisions – that is, does the exploration-related P300 depend on the rate of exploration?

To answer this question, I examined behavioural and neural responses across three decision making tasks. Participants received rewards after choosing one of several options. Rewards were probabilistic and changed over time, encouraging exploration (Daw et al., 2006). Importantly, participants were shown either 4, 9, or 16 options, i.e., a 4-armed, 9-armed, or 16-armed bandit. My goal with this manipulation was to encourage exploration by offering more options. My P300 hypothesis had two parts. First, I predicted an enhanced signal for feedback following decisions to explore relative to decisions to exploit, in line with previous work (Experiments 1 and 2). Expanding on this result, I also anticipated that this enhancement would be modulated by task. As decisions to explore became more frequent, i.e., as the number of options increased, the difference between the exploratory P300 and the exploitative P300 was predicted to diminish or even disappear completely. Finally, I conducted an unplanned analysis of the P200, an

ERP component linked to risk processing (Kiat, Straley, & Cheadle, 2016; Schuermann, Endrass, & Kathmann, 2012).

Methods

Participants. Thirty-five university-aged participants (13 male, 3 left-handed, M_{age} = 22, 95% CI [20, 25] with no known neurological impairments and with normal or corrected-to-normal vision took part in the experiment. Data from five participants were excluded from the analysis due excessive EEG noise and/or too few trials of a certain type (exploit, explore). All of the participants were volunteers who received credit in an undergraduate course for their participation. Additionally, participants were paid a performance-dependent bonus of up to $5. The participants provided informed consent approved by the Human Research Ethics Board at the University of Victoria.

Apparatus and procedure. Participants were seated 60 cm in front of a 22-inch LCD display (75 Hz, 2 ms response rate, 1680 by 1050 pixels, LG W2242TQ-GF, Seoul, South Korea). Visual stimuli were presented using the Psychophysics Toolbox Extension (Brainard, 1997; Pelli, 1997) for MATLAB (Version 8.2, Mathworks, Natick, USA). Participants were given written and verbal instructions to minimize head and eye movements throughout the experiment (Appendix A).

Participants played three versions of a multi-armed bandit task (Daw et al., 2006) – a 4-armed, 9-armed, and 16-armed bandit – in pseudorandom order. There were 300 trials per task, for a total of 900 trials across all three tasks. Participants were told that they would be playing three slot machine games, and that in each game their goal was to maximize their point total by selecting the slot machine most likely to yield a large reward (a point amount that ranged from 1-100). The drifting nature of the payouts was

explained, using two examples (see Appendix A for participant instructions). Additionally, participants were told that their average point amount would be converted to a proportion of $5, which they would receive at the end of the experiment.

Bandits were represented by coloured squares presented against a black background. The locations of the squares never changed within each task. On each trial, participants had up to two seconds to select a bandit using the mouse. Upon selecting a bandit, the mouse cursor was occluded, and the bandit choice was highlighted with a white border. After a variable delay (400-600 ms), feedback was presented within the chosen square for one second, at which time the highlighting on the chosen square was removed and the mouse cursor shown, signalling the start of another trial. The coloured squares together made a square that was six degrees of visual angle on a side, a size that was maintained across each task. Likewise, the feedback in each task was one degree of visual angle high. Thus, although total stimulus size and feedback size were held constant, the size of the squares themselves varied (more choices meant smaller squares). See Figure 15 for a trial overview.

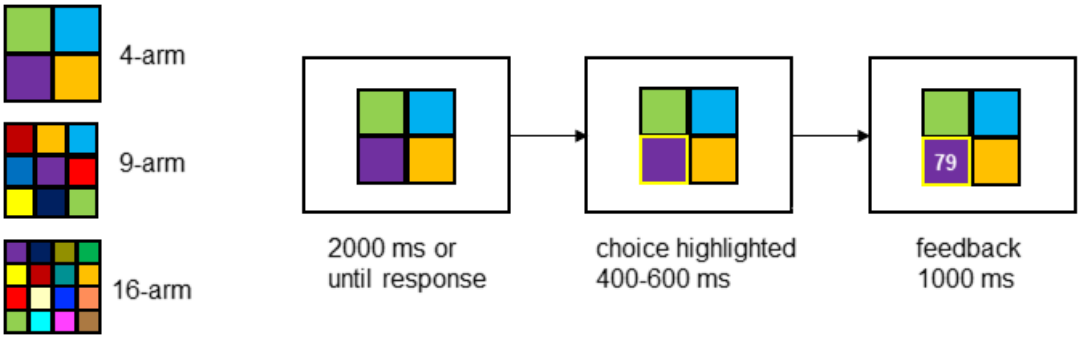


Figure 15. Task, with timing details. Participants played either a 4-armed, 9-armed, or 16-armed bandit (one block of each).

Data collection. Sixty-three channels of EEG data, referenced to channel AFz, were recorded using Brain Vision Recorder (Version 1.21.0004, Brain Products GmbH, Munich, Germany). Sixty-one electrodes were placed in a fitted cap according to the 10-20 system. Additionally, two electrodes were affixed to the mastoids (left and right). Conductive gel was applied to ensure that electrode impedances were below 20 kΩ prior to recording, and the EEG data were sampled at 500 Hz and amplified (actiCHamp, Brain Products GmbH, Munich, Germany).

Computational models. I used computational modelling to classify trials as either exploitations or explorations. Three models were implemented in MATLAB and evaluated based on their ability to account for my participants' decisions. In particular, I first examined two reinforcement learning models, differing only in their method of action selection: *greedy* and *softmax*. I then considered a "win-stay, lose-shift" heuristic (*WSLS*).

Greedy. A greedy model maintains values for all n actions (or bandits), which are updated on every trial: $v_t(1), \ldots, v_t(n)$. It then selects the highest-valued action. I

implemented *near*-greedy action selection, which tends to go with the highest-valued action most of the time, but selects any other action with probability ε (R. S. Sutton & Barto, 2018). Thus, for an *n*-armed bandit, the probability of selecting stimulus *i* on trial *t* (that is, the likelihood of making an action a_i) was computed using:

$$P_t(a_i) = \begin{cases} 1 - \varepsilon \; if \underset{x}{\text{argmax}}\, v_t(x) \\ \dfrac{\varepsilon}{n-1} \; otherwise \end{cases}$$

On each trial, following feedback R_t, a prediction error δ_t was generated for the selected stimulus *s* according to:

$$\delta_t = R_t - v_t\,(s)$$

The value of the chosen bandit *s* was then updated using the following learning rule:

$$v_{t+1}(s) = \; v_t(s) + \; \alpha\delta_t$$

in which prediction errors were scaled by a learning rate, α. The value of the unselected stimulus was unchanged.

The parameters (ε, α) were tuned for each participant and task using the MATLAB function fmincon (Optimization Toolbox, Release 2018a, Mathworks, Natick). To be clear, this was done for each model, participant, and task. Specifically, I constructed an objective function (the function to be minimized) as the negative log-likelihood of a participant's set of responses in a particular task (one-step-ahead prediction: Ahn et al., 2008). Log-likelihood was computed as:

$$\sum_t \log(P_t(a_s))$$

where $P_t(a_s)$ was the probability associated with the selected bandit *s* on trial *t*.

Thus, model tuning produced three values for each participant: a final log-likelihood, a final ε, and a final α. I then classified trials as exploitations or explorations using the action selection stage of the final model. Trials on which the participant made the greedy response (the highest-value response) were classified as exploitations. All other trials were classified as explorations.

Softmax. Next, I considered a model that allowed for more nuance in terms of choice preference. My softmax model differed from my greedy model in one aspect only. Instead of making near-greedy choices, it computed action probabilities according to the softmax equation:

$$P_t(a_i) = \frac{e^{v_t(i)/\tau}}{e^{v_t(1)/\tau} + e^{v_t(2)/\tau}}$$

where τ (temperature) determined the degree of bias towards choosing high-valued stimuli (greater bias for lower τ). Model tuning was done as before to produce a log-likelihood (as defined for the greedy model), τ, and α. Here, exploitation occurred when a participant chose the action with the greatest associated softmax probability; all other actions were explorations.

Win-stay, lose-shift. Win-stay, lose-shift is a decision-making strategy that only depends on the previous trial's feedback. If the previous trial's action resulted in a win, it is likely to be repeated in the current trial; if the previous trial's action resulted in a loss, it is likely to be avoided in the current trial. In particular, winning actions are repeated with probability *P(stay|win)*, and losing actions avoided with probability *P(shift|loss)*. Since the present experiment involved only gains (points from 1-100) I defined a "win" as a point amount greater than or equal to 50 (the long-run average of each bandit), and a

"loss" otherwise. Two free parameters were thus tuned for each task, *P(stay|win)* and *P(shift|loss)*. These probabilities determined the log-likelihood of each trial type within an *n*-armed bandit:

$$LL_{win-stay} = log(P(stay|win))$$

$$LL_{win-shift} = log(\frac{1 - P(stay|win)}{1 - n})$$

$$LL_{lose-shift} = log(P(shift|loss))$$

$$LL_{lose-stay} = log(\frac{1 - P(shift|loss)}{1 - n})$$

Model tuning involved minimizing the negative sum of these log-likelihoods, as before. Exploration here was defined as following the counter-intuitive strategy – that is, sticking with a choice following a loss, or switching to another action following a win.

Data analysis.

Modelling data. I first compared each model to a baseline model (Gureckis & Love, 2009). The reason why I used a baseline model was to attempt to control for the fact that actions (e.g., explorations) may appear less probable as more choices are added. For example, the probability of exploring in the greedy model was computed as ε/(*n*-1) where *n* was the size of the action space (the number of bandits). Thus, a particular exploratory option would occur with ε/15 likelihood in the 16-armed bandit, but only ε/3 likelihood in the 4-armed bandit.

To define a baseline model for each task I considered the probabilities with which participants selected particular bandits. Baseline models were tuned for each participant and task – an *n*-armed bandit model therefore had *n*-1 free parameters, representing the probabilities associated with each of the choices, $P_1, P_2, \ldots, P_{n-1}$ (the probability of

selecting the last bandit was one minus the sum of the other probabilities). These probabilities determined trial log-likelihoods, e.g. if bandit *i* was chosen then the log-likelihood for that trial was $\log(P_i)$. The baseline model was tuned by minimizing the negative sum of these log-likelihoods. For all models (including the baseline) I then converted each the overall log-likelihood (*LL*), found via model tuning, to an Akaike's information criterion (AIC) as per:

$$AIC = -2LL + 2(number\ of\ free\ parameters)$$

Next, for each participant, task, and model, I computed the relative fit *RF* compared to baseline for each model *m*:

$$RF_m = AIC_b - AIC_m$$

Large *RF* values indicate a better model fit compared to the baseline model. I used model RF to determine the best-fitting model for each of my participants and tasks. Because I evaluated my models in this way, a range of best-fitting models was possible, both across and within individuals. The best-fitting model for each participant and task was used to classify trials as either exploitations or explorations for the purpose of my behavioural and EEG analyses. As my results indicate, however, one model usually came out ahead of the others; furthermore, the models more-or-less agreed on the identity (exploit/explore) of trial. To quantify model agreement, I computed the proportion of trials, for each participant, that the two models gave the same classification.

Behavioural data. For each participant and task, I computed the mean number of trial of each type (exploit/explore). For each participant, task (4-armed, 9-armed, 16-armed), and decision type (exploit, explore), I computed the mean response time and reward. Trials with a response time of more than two seconds were excluded from all

analyses (model tuning, behavioural, and EEG). Finally, for each response I computed the number of times that particular response option had been chosen. I then computed the mean number of prior responses for each decision type and task.

Electroencephalographic data. EEG data were down-sampled to 250 Hz, filtered through a (0.1 Hz – 30 Hz pass band) phase shift-free Butterworth filter (60 Hz notch), and rereferenced to the average of the two mastoid channels. Next, ocular artifacts were removed using independent component analysis. Subsequent to this, 800 ms epochs of EEG data were constructed from 200 ms prior to 800 ms following feedback onset. All trials were then baseline corrected using a 200 ms pre-feedback window. Finally, trials in which the change in voltage in any channel exceeded 10 μV per sampling point or the change in voltage across the epoch was greater than 100 μV were discarded. On average, I removed 25% of epochs (95% CI [19, 31]).

Examination of the grand-grand waveform. To avoid biasing my analysis in favour of a statistically-significant difference between my conditions of interest, I defined the P300 by first examining the "grand-grand" average waveform (Kappenman & Luck, 2016). For each participant, I averaged across all EEG epochs (regardless of condition), then averaged across participants. Next, I identified the time/locations at which the most positive-going deflection occurred (460 ms post-stimulus at electrode CP3). To capture the apparent P300 deflection, I identified the times at which 85% of the maximum voltage was reached (352 to 592 ms post feedback), which formed my analysis window (see dashed line in Figure 16).

Effect of current/next trial decision. In previous work (Experiments 1 and 2), I analyzed the effect of the upcoming decision type (exploit/explore) on the P300 and

reported an exploration-related enhancement. Here, I was interested in whether this effect was driven by the current-trial decision or the next-trial decision. To determine these effects in this experiment, I combined trials across tasks (4-armed, 9-armed, 16-armed), and examined the effect of current/next trial decision on the P300, defined in the same way as in my main analysis (the mean voltage from 352-592 ms post feedback at electrode CP3). Four trial groupings (and four waveforms) were created for each participant based the current and next trial type: exploit-exploit, exploit-explore, explore-exploit, and explore-explore. See Figure 16 for each condition's trial count. I then conducted a 2 (current trial type: explore/exploit) by 2 (next trial type: explore/exploit) repeated measures ANOVA. The P300 effect appeared to be driven mostly by the current-trial decision (see Results). For my main analysis, described below, I therefore decided to focus on the current-trial decision in instead of next-trial decision.

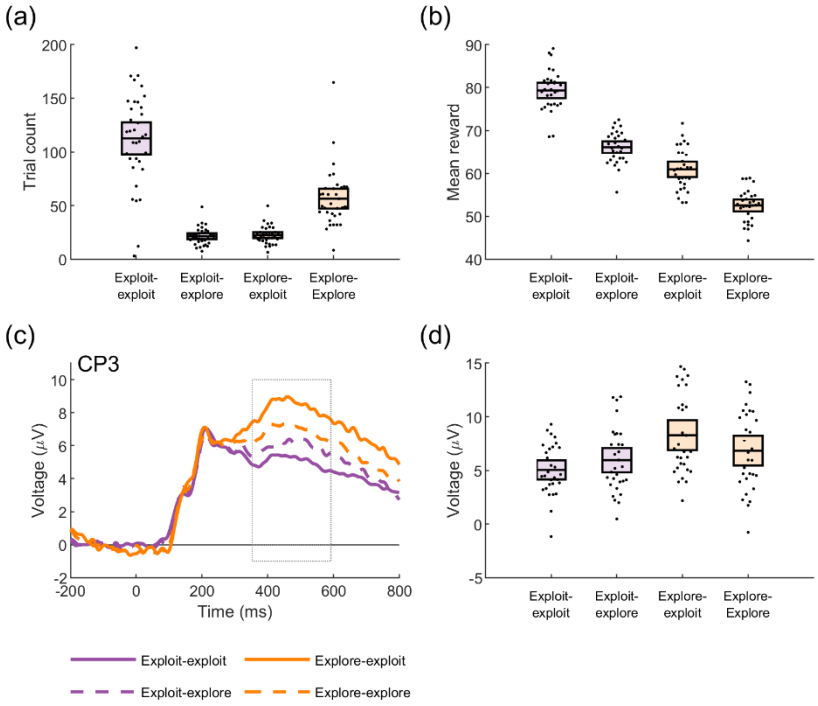


Figure 16. Behavioural and EEG data by current/next trial type. (a) Participants were more likely to exploit following exploitations, and more likely to explore following explorations. (b) Mean rewards, by current/next trial type. (c) Feedback-locked waveforms for each current/next trial type. The dotted line shows the region of analysis. (d) P300 scores by current/next trial type.

Feedback-locked P300. To quantify the P300 for my main analysis, I averaged the feedback-locked EEG data for each channel, participant, task, and current-trial decision type (exploit/explore). For each waveform, the P300 was then defined as the mean voltage from 352-592 ms post feedback at electrode CP3, as described earlier. See Figure 17.

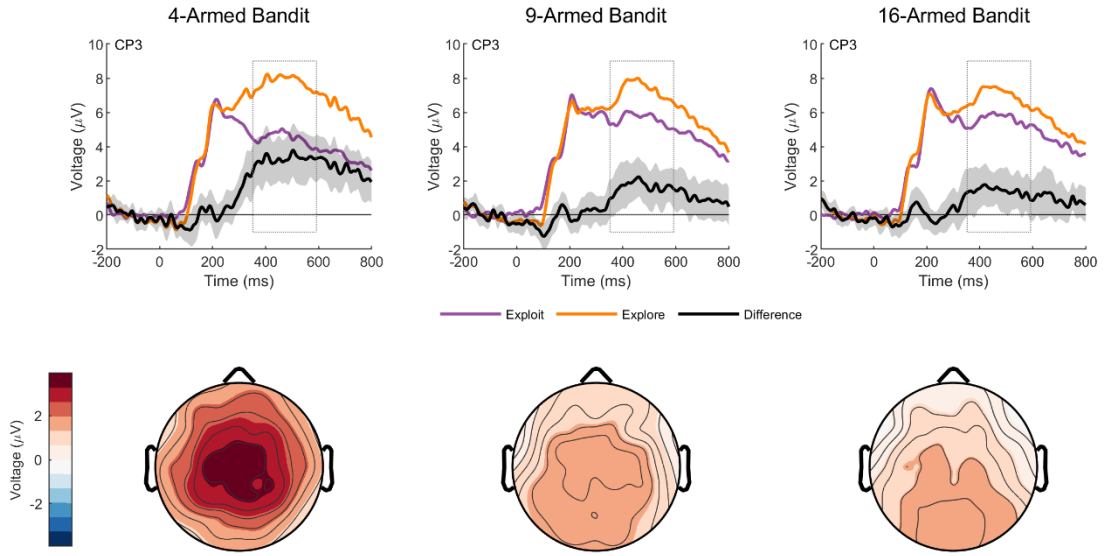


Figure 17. Feedback-locked P300 responses following decisions to either exploit or explore, with difference waveforms and scalp topographies. The shaded region on the difference waveforms shows the 95% confidence intervals. The dotted line shows the region of analysis (i.e., the range of scores that were averaged to obtain the magnitude of the P300 for each condition).

Feedback-locked P200. Upon examining the neural response to feedback (Figure 18) I noted a prominent difference at electrode FCz in the time range of a second ERP component called the P200, or P2. Post-hoc, I decided to analyze this component using the same ANOVA structure as with the P300 (see below). A time window of 40 ms centered around 160 ms post feedback was chosen in order to capture as much of the observed difference, across all tasks, as possible.

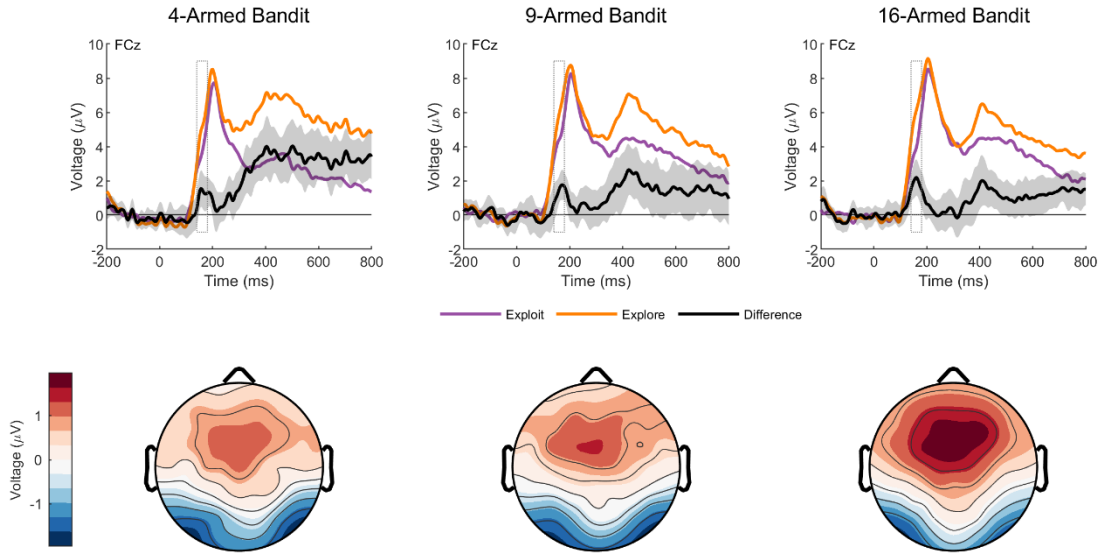


Figure 18. Feedback-locked P200 responses following decisions to either exploit or explore, with difference waveforms and scalp topographies. The shaded region on the difference waveforms shows the 95% confidence intervals. The dotted line shows the region of analysis (i.e., the range of scores that were averaged to obtain the magnitude of the P200 for each condition).

Inferential statistics. The effects of current/next trial decision on trial count, mean reward, and the P300 were analyzed using a 2 (current-trial decision: exploit/explore) by 2 (next-trial decision: exploit/explore) ANOVA. The effect of task on the proportion of trials classified as explorations was analyzed in a one-way ANOVA (task: 4-armed, 9-armed, 16-armed). All other behavioural measures (response time and reward), the feedback-locked P300, and the feedback-locked P200 were analyzed using a 2 (decision: exploit, explore) by 3 (task type: 4-armed, 9-armed, 16-armed) repeated-measures ANOVA. Alpha was assumed to be .05 for all statistical tests. Mauchly's Test

of Sphericity was used to check for violations of sphericity, when appropriate. Two different effect-size measures were computed: η_p^2 and η_g^2 (Olejnik & Algina, 2003).

Results

Modelling data. The WSLS model provided the best relative fit for most participant, across all three tasks (Table 1). However, I noted that the softmax and WSLS models tended to make similar predictions. On average, there was 95.1% classification overlap in the 4-armed bandit (95% CI [93.8, 96.4]), 95.1% overlap in the 9-armed bandit (95% CI [94.0, 96.2]), and 95.8% overlap in the 16-armed bandit (95% CI [94.9, 96.8]). See Figure 19 for mean relative fits.

Table 1

Best-fitting model participant counts (N = 30)

	4-Armed Bandit	9-Armed Bandit	16-Armed Bandit
Greedy	0	0	0
Softmax	8	1	1
WSLS	22	29	29

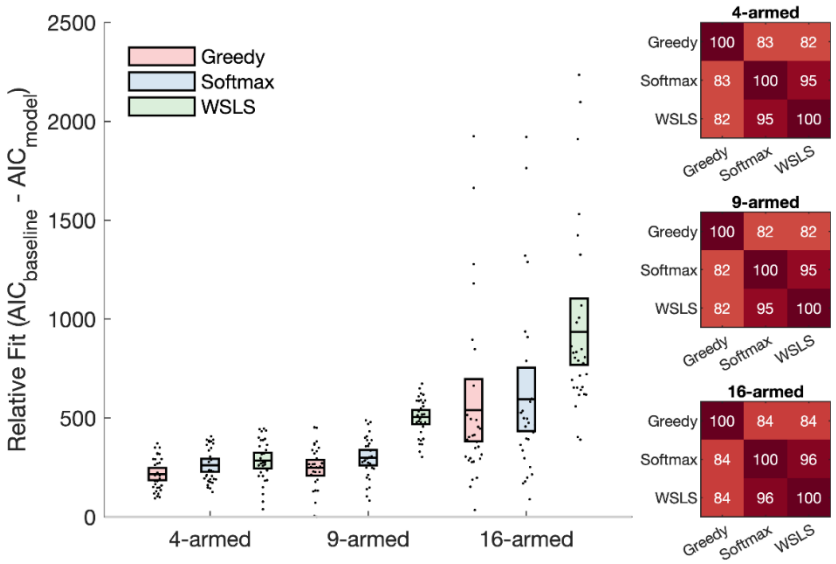


Figure 19. Comparison of model fits. Scores indicate improvement in AIC relative to a baseline model ($AIC_{baseline} - AIC_{model}$); greater scores mean more of an improvement (left). The models were used to classify trials as either exploitations or explorations, and model classification overlapped across all trials – see mean percentage of equivalent classifications (right). Error bars show 95% confidence intervals.

Behavioural data.

Effect of current/next trial decision. I used each participant's best-fitting model to classify trials as exploitations or explorations. After collapsing across task to examine the effect of current/next trial type on trial counts, I noted an effect of current-trial type (more exploitations), $F(1,29) = 35$, $p < .001$, $\eta_p^2 = 0.55$, $\eta_g^2 = 0.28$ and next-trial type (more exploitations)), $F(1,29) = 16$, $p < .001$, $\eta_p^2 = 0.36$, $\eta_g^2 = 0.11$. Importantly, there was a current-trial by next-trial interaction; decision types tended to repeat, $F(1,29) = 50$, $p < .001$, $\eta_p^2 = 0.63$, $\eta_g^2 = 0.28$. I also examined the effect of current/next trial type on reward and found an effect of both current-trial type (larger rewards for exploitations),

$F(1,29) = 359$, $p < .001$, $\eta_p^2 = 0.93$, $\eta_g^2 = 0.78$, and next-trial type (larger rewards for exploitations), $F(1,29) = 311$, $p < .001$, $\eta_p^2 = 0.91$, $\eta_g^2 = 0.62$. There was a smaller interaction effect between current-trial type and next-trial type – it appeared to take a greater points difference to switch from exploitation to exploration than it took to switch from exploration to exploitation, $F(1,29) = 23$, $p < .001$, $\eta_p^2 = 0.44$, $\eta_g^2 = 0.08$. See Table 2 for exact values, and Figure 16.

Exploration rate. A one-way ANOVA revealed that the proportion of trials classified as explorations differed by task, which was the desired effect of my task-level manipulation, $F(2, 58) = 30.22$, $p < .001$, $\eta_p^2 = .510$, $\eta_g^2 = .285$.

Response time. Mauchly's Test of Sphericity revealed no violation of the assumption of sphericity for either task (Mauchly's $W = 0.90$, $\chi^2(2) = 2.9$, $p = .23$), or the task by decision type interaction (Mauchly's $W = 0.99$, $\chi^2(2) = 0.39$, $p = .82$). A 2X3 ANOVA with decision type (exploit, explore) and task (4 arms, 9 arms, 16 arms) as repeated measures showed a main effect of decision type (explorations slower than exploitations), $F(1, 29) = 28.87$, $p < .001$, $\eta_p^2 = .499$, $\eta_g^2 = .064$, but not task, $F(2, 58) = 1.82$, $p = .17$, $\eta_p^2 = .059$, $\eta_g^2 = .009$. There was a slight interaction between decision type and task – the exploration-related slowing was more pronounced in the larger bandits, $F(2, 58) = 5.69$, $p = .006$, , $\eta_p^2 = .16$, $\eta_g^2 = .003$ (Figure 20).

Table 2

Effects of current/next trial decision

	Exploit (current)				Explore (current)			
	Exploit (next)		Explore (next)		Exploit (next)		Explore (next)	
Measure	*M*	95% CI	*M*	95% CI	*M*	95% CI	*M*	95% CI
Trial count	113	[98, 128]	21	[19, 24]	22	[20, 25]	56	[47, 66]
Reward (points)	79	[78, 81]	66	[65, 67]	61	[59, 63]	53	[51, 54]
P300 (μV)	5.0	[4.2, 5.9]	6.0	[4.8, 7.1]	8.3	[6.9, 9.7]	6.8	[5.5, 8.2]

Reward. Mauchly's Test of Sphericity revealed no violation of the assumption of sphericity for either task (Mauchly's $W = 0.95$, $\chi^2(2) = 1.2$, $p = .51$), or the task by decision type interaction (Mauchly's $W = 0.95$, $\chi^2(2) = 1.5$, $p = .47$). Exploration tended to result in a smaller reward compared to exploitation – a similar 2X3 ANOVA with reward as the dependent variable showed a main effect of decision type, $F(1, 29) = 1321$, $p < .001$, $\eta_p^2 = .979$, $\eta_g^2 = .809$. The effect of task was also significant, $F(2, 58) = 22.41$, $p < .001$, , $\eta_p^2 = .436$, $\eta_g^2 = .309$, as was the interaction, $F(2, 58) = 10.19$, $p < .001$, $\eta_p^2 = .260$, $\eta_g^2 = .043$. See Figure 20 and Table 3 for the means and 95% confidence intervals.

Prior responses. Mauchly's Test of Sphericity revealed a violation of the assumption of sphericity for task (Mauchly's $W = 0.74$, $\chi^2(2) = 8.6$, $p = .01$), but not the task by decision type interaction (Mauchly's $W = 0.97$, $\chi^2(2) = 0.8$, $p = .7$). A Greenhouse-Geisser ε of 0.79 was used to correct the degrees of freedom for the task effect. Exploration tended to result in fewer prior responses, $F(1, 29) = 185$, $p < .001$, $\eta_p^2 = .86$, $\eta_g^2 = .53$. The effect of task was also significant, $F(1.6, 46) = 155$, $p < .001$, $\eta_p^2 =$

.84, η_g^2 = .64, There was no task by decision-type interaction, $F(2, 58) = 1.0$, $p = .4$, η_p^2 = .03, η_g^2 = .01. See Table 3 and Figure 21 for means and 95% confidence intervals.

Table 3

Behavioural summary (means and 95% confidence intervals)

	4-Armed Bandit		9-Armed Bandit		16-Armed Bandit	
	M	95% CI	*M*	95% CI	*M*	95% CI
Exploration rate (%)	25	[22, 29]	39	[34, 44]	42	[37, 46]
Exploitative RT (s)	300	[247, 352]	297	[259, 335]	314	[271, 357]
Exploratory RT (s)	348	[285, 412]	379	[328, 429]	396	[341, 450]
Exploitative reward (points)	67	[64, 69]	74	[72, 76]	77	[76, 79]
Exploratory reward (points)	49	[47, 51]	53	[51, 54]	54	[53, 56]
Exploitative prior (trials)	55	[51, 59]	35	[31, 38]	28	[24, 31]
Exploratory prior (trials)	35	[32, 39]	18	[17, 19]	12	[11, 14]

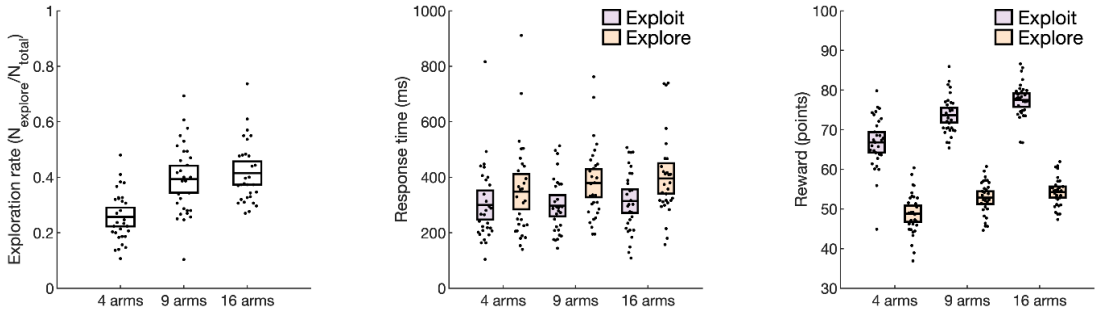


Figure 20. Behavioural data. The number of explorations, as determined by a computational model, increased with task size (left). Decisions to explore took slightly longer than decisions to exploit (middle). Exploration tended to result in fewer rewards, while having more options tended to result in greater rewards (right). Error bars show 95% confidence intervals.

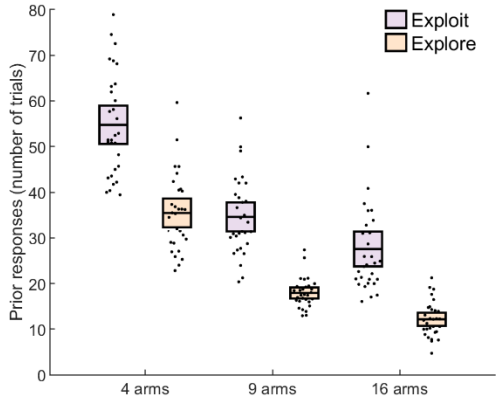


Figure 21. Mean number of prior responses, for each decision type and task. Explorations tended to be preceded by fewer responses of the same choice. Particular response options were sampled less often in the larger decision space.

Electroencephalographic data.

Effect of current/next trial decision. There was a significant effect of current-trial type ($F(1,29) = 29$, $p < .001$, $\eta_p^2 = 0.50$, $\eta_g^2 = 0.09$), but not next-trial type ($F(1,29) = 2.5$, $p = .1$, $\eta_p^2 = 0.08$, $\eta_g^2 = 0.00$). There was a significant current-trial by next-trial interaction, $F(1,29) = 25$, $p < .001$, $\eta_p^2 = 0.46$, $\eta_g^2 = 0.03$. Specifically, the effect of exploring on the current trial appeared to be modulated by next-trial decision type (greater when the next trial was an exploitation). See Figure 16.

Feedback-locked P300. Mauchly's Test of Sphericity revealed no violation of the assumption of sphericity for either task (Mauchly's $W = 1.00$, $\chi^2(2) = 0.0$, $p = .99$), or the task by decision type interaction (Mauchly's $W = 0.88$, $\chi^2(2) = 3.5$, $p = .17$). My two (decision type: exploit, explore) by three (task: 4 arms, 9 arms, 16 arms) ANOVA revealed a main effect of decision type on the P300 component, $F(1,29) = 25$, $p < .001$,

$\eta_p^2 = 0.46$, $\eta_g^2 = 0.09$. There was no effect of task, $F(2, 58) = 0.4$, $p = .7$, $\eta_p^2 = 0.01$, $\eta_g^2 = 0.00$. There was, however, a significant interaction between decision type and task (the difference between the exploit response and the explore response was smaller for the larger tasks), $F(2, 58) = 11$, $p < .001$, $\eta_p^2 = 0.27$, $\eta_g^2 = 0.01$. See Figure 17 for waveforms and scalp topographies, and Table 4 for P300 means and 95% confidence intervals.

Table 4

ERP summary (means and 95% confidence intervals)

	4-Armed Bandit		9-Armed Bandit		16-Armed Bandit	
	M	95% CI	*M*	95% CI	*M*	95% CI
Exploit P300 (μV)	4.5	[3.5, 5.5]	5.6	[4.5, 6.7]	5.6	[4.6, 6.6]
Explore P300 (μV)	7.8	[6.1, 9.4]	7.3	[5.8, 8.8]	7.0	[5.6, 8.3]
Exploit P200 (μV)	3.84	[2.75, 4.93]	4.44	[3.31, 5.58]	4.19	[3.01, 5.37]
Explore P200 (μV)	5.03	[3.47, 6.60]	5.81	[4.47, 7.15]	6.05	[4.62, 7.49]

The difference in P300 score (explore minus exploit) correlated negatively with likelihood of exploration in the 4-armed bandit, $r(28) = -0.56$, $p = .001$. I found no evidence for a relationship in either the 9-armed case, $r(28) = -0.09$, $p = .6$, or the 16-armed case, $r(28) = -0.26$, $p = .2$ (Figure 22).

Feedback-locked P200. Mauchly's Test of Sphericity revealed no violation of the assumption of sphericity for either task (Mauchly's $W = 0.98$, $\chi^2(2) = 0.53$, $p = .77$), or the task by decision type interaction (Mauchly's $W = 0.91$, $\chi^2(2) = 2.5$, $p = .28$). There was main effect of decision type, $F(1, 29) = 17.52$, $p < .001$, $\eta_p^2 = .377$, $\eta_g^2 = .044$. There was also a main effect of task (P200 scores became more positive overall across task),

$F(2,58) = 4.12$, $p = .02$, $\eta_p^2 = .124$, $\eta_g^2 = .008$. The interaction was not significant, $F(1, 29) = 1.74$, $p = .18$, $\eta_p^2 = .057$, $\eta_g^2 = .002$. See Figure 18 for waveforms and scalp topographies, and Table 4 for P200 means and 95% confidence intervals.

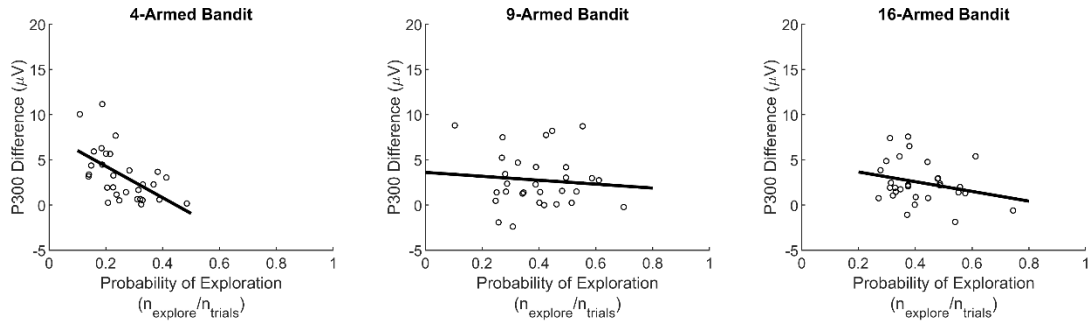


Figure 22. Correlations between P300 difference (exploit minus explore) and likelihood of exploring. Participants who rarely explore have a larger P300 difference (relationship significant in the 4-armed case only).

Discussion

In the present experiment, I showed an exploration-dependent enhancement of the P300, an ERP component though to reflect a neural interrupt signal. By examining how this signal responded across three bandit tasks, I found support for the hypothesis that the exploration-related P300 depends on the rate of exploration.

I first examined which of three models was more effective in accounting for my participants' trial-to-trial decisions. Consistent with previous work (Daw et al., 2006) I found that, between two RL models, a softmax method of action selection resulted in a better model fit compared to a greedy method. However, a WSLS model yielded the superior fit in most cases (especially for the 16-armed bandit). Others have made similar

observations. For example, when choosing from one of four decks of cards in the Iowa gambling task (IGT), around half of participants will use a WSLS (as opposed to RL) strategy (Worthy, Hawthorne, & Otto, 2013). Older adults tend to make responses more in line with a WSLS model compared to younger adults, a result attributed to either cognitive decline or wisdom (Worthy & Maddox, 2012). Supporting the cognitive decline hypothesis, people are more likely to employ a lose-shift strategy under increasing cognitive load (Ivan, Banks, Goodfellow, & Gruber, 2018). This is in line with my observation that the WSLS model advantage appeared to depend on the number of bandits (Figure 19), if I assume that more bandits meant greater cognitive load[2].

I then used my computational models to classify participant decisions as exploitations or explorations and examined the neural response to feedback, controlling for current- and upcoming-trial decision type. These effects and their interaction were an open question, as previously I had only examined the role of upcoming-trial decision type (Experiments 1 and 2). Here I showed that – for this task, at least – modulation of the feedback-locked P300 amplitude is best explained by the decision made on the current trial (enhanced for exploration). I also noted an interesting interaction between current- and next-trial decision type; the feedback-locked P300 is even greater when a participant is about to switch from exploration to exploitation. This interaction will be discussed in more detail later on.

Next, I made several observations related to my main question of how the exploration-related P300 is affected by task size. First, exploration rate increased as more

---

[2] On the other hand, others have shown that increasing working memory load tends to shift people *away* from a WSLS strategy and towards an RL strategy (Otto, Taylor, & Markman, 2011).

response options were presented. This observation confirmed my main manipulation; I was able to induce more exploratory behaviour by offering up more response options. Second, decisions to explore were slower and less rewarding than decisions to exploit. Finally, I observed an enhanced P300 to feedback following to decisions to explore, an effect that diminished across my tasks as the rate of exploration decreased. This result supports the hypothesis that the amplitude of the exploration-related P300 depends in part on the likelihood of exploration. Further support for this hypothesis is seen in the significant relationship between the P300 difference score and exploration rate in the 4-armed bandit; the more likely a participant was to explore, the smaller their P300 effect.

Why would the exploration-related P300 depend on rate of exploration? The answer could relate to the observation that phasic NE activity is greater for infrequent target stimuli compared to frequent target stimuli (G. Aston-Jones, Rajkowski, Kubiak, & Alexinsky, 1994). This observation aligns with the model of Dayan and Yu (2006), in which phasic NE magnitude depends inversely on the prior probability of a target; rare targets elicit larger signals. Recall that the purpose of this signal is to trigger a shift away from the default distractor state, which Dayan and Yu (2006) defined as the more frequent state. Accordingly, in the current study I might define exploitation as the default state because it was the more frequent decision type compared to exploration. Participants who explore less frequently should therefore exhibit increased phasic NE (and a larger P300 effect; see correlation in Figure 22). Likewise, tasks in which exploration is more common should elicit reduced phasic NE (and a smaller P300 effect; see Figure 17).

This interpretation of my results is complicated by two factors. First, recall that when examining the effect of current/next trial decision, I observed that the largest P300 amplitude occurred for exploratory feedback when the upcoming trial was an exploitation. The reason why this is an issue is that it implies that this signal is particularly sensitive to shifting away from exploration, towards exploitation. This does not align with the view that exploitation is the default strategy in the current task and may even suggest the opposite. Second, the relationship between my P300 effect and exploration rate was observed in the 4-armed bandit only, in which exploration was less frequent than exploitation for all participants. This relationship did not hold in the 9-armed and 16-armed cases, when exploration became the dominant strategy for some participants (Figure 22). In other words, people who explored more than they exploited did not show a reverse effect – a larger P300 for exploitative feedback.

I therefore have partial support only for the hypothesis that the purpose of the neural interrupt signal observed here is to halt one mode of decision making in favour of another. Alternatively, this signal may reflect not the interruption of ongoing decision-making processes, but rather the disruption of one's model of the external environment. According to my models, participant decisions were driven by the value of various response options. I suggest here that feedback that did not align with a chosen option's representation elicited a neural interrupt signal. To align this hypothesis with the model of Dayan and Yu (2006), I might specify that the magnitude of this signal was proportional to the likelihood that the chosen option was the best option, and inversely proportional to the likelihood of sampling that chosen option. The appeal of a description like this is that exploratory feedback will tend to elicit greater NE signals since

exploratory responses have lower priors compared to exploitative responses (Figure 21). In other words, exploratory feedback might elicit larger P300s not because exploration is rare compared to exploitation, but because exploratory outcomes tend to be more uncertain. Further work would be needed to confirm this, e.g. using a task in which exploration is more frequent than exploitation.

Examining the feedback-locked response revealed the exploration-related enhancement of another positive ERP component, the P200. The P200 (or P2) peaks earlier than the P300, usually around 200 ms post stimulus for visual stimuli (Schuermann et al., 2012). Though not part of my original hypothesis, I feel that the presence of this signal suggests an interesting direction for future work. In particular, the P200 has been linked to risk processing; feedback following high-risk decisions elicits a larger P200 compared to feedback following low-risk decisions (Kiat et al., 2016; Schuermann et al., 2012). Similarly, an unexpected outcome elicits a larger P200 compared to an expected outcome (Polezzi, Lotto, Daum, Sartori, & Rumiati, 2008). Interestingly, the P300 is also enhanced for outcomes following risky choices (Polezzi, Sartori, Rumiati, Vidotto, & Daum, 2010; Schuermann et al., 2012). Thus, my two exploration-related effects (P200 and P300) look suspiciously like risk-related effects, according to previous literature. This is perhaps unsurprising, as exploration is often risky. But is all exploration risky? Alternatively, are all risky decisions a form of exploration? Although my experiment was not designed to answer these questions, future work might.

Here I have shown further support for the role of the LC-NE system in the explore-exploit dilemma. The P300, an ERP index of phasic LC-NE activity, is enhanced

for exploratory feedback relative to exploitative feedback. Additionally, I have shown that this effect depends on exploration rate – the more frequent that one explores, the smaller the difference between the exploratory P300 and the exploitative P300. It remains to be seen whether this effect could be further reduced, abolished, or even reversed (i.e., a larger P300 following exploitative feedback).

Chapter 5: Experiment 4

Abstract

Decision-making is typically studied by presenting participants with a small set of options. However, real-world behaviour, like foraging, often occurs in continuous environments. The degree to which human decision-making in discrete tasks generalizes to continuous tasks is questionable. For example, successful foraging comprises both exploration (learning about the environment) and exploitation (taking advantage of what is known). Although progress has been made in understanding the neural processes related to this trade-off in discrete tasks, it is currently unknown how, or whether, the same processes are involved in continuous tasks. To address this, I recorded electroencephalographic data while participants "dug for gold" by selecting locations on a map. Participants were cued beforehand that the map contained either a single patch of gold, or many patches of gold. I then used a computational model to classify participant responses as either exploitations, which were driven by previous reward locations and amounts, or explorations. My participants were able to adjust their strategy based on reward distribution, exploring more in multi-patch environments and less in single-patch environments. I observed an enhancement of the feedback-locked P300, a neural signal previously linked to exploration in discrete tasks, which suggests the presence of a general neural system for managing the explore-exploit trade-off.

Feedback Processing Is Enhanced Following Exploration in Continuous Environments

There is a growing body of literature on how animals – including humans – manage the trade-off between exploiting prior experience and exploring new options. The explore-exploit trade-off is affected by several factors, including environmental volatility (Behrens et al., 2007), stress (Lenow, Constantino, Daw, & Phelps, 2017), the total number of remaining decisions (Wilson, Geana, White, Ludvig, & Cohen, 2014), and reward distribution (Constantino & Daw, 2015). The effect of reward distribution is of particular interest because it cuts across multiple species. For instance, snail communities are affected by the patchiness, or spatial clustering, of available food (J. M. Chase, Wilson, & Richards, 2001). Highly patchy environments, which are more heterogenous, are dominated by snail species that tend to explore ("grazers"). Conversely, less patchy environments, which are more homogenous, are dominated by snail species that tend to exploit ("diggers").

Unlike snails, humans are flexible decision-makers. For example, Constantino and Daw (2015) observed that individuals will tailor their patch-leaving decisions to the current reward distribution; thus, our decision-making is adaptable to the (external) environment. However, individual (internal) factors also play a role. For instance, patch-leaving strategies in a simulated fishing game show considerable inter-subject variability (Hutchinson, Wilke, & Todd, 2008). There, participants were asked to make a series of decisions – to either fish or switch ponds – and were told that the number of fish in each pond might vary. Due to response variability, and contrary to the authors' predictions, patch-leaving decisions were unaffected by reward distribution (Hutchinson et al., 2008).

It is therefore unclear in what way humans are able to use reward distribution knowledge, if at all.

Also somewhat unclear is the neural basis of exploration, although it has been studied using a variety of neuroimaging techniques, including electroencephalography (EEG). Early work suggests that a machine learning classifier can use the EEG at frontal and parietal sites to accurately predict whether an individual will explore or exploit (Bourdaud et al., 2008; Tzovara et al., 2012). Similarly, I have identified a parietal component of the event-related potential called the P300 that is associated with decisions to explore (Experiments 1 and 2). In particular, feedback following decisions to explore elicits an enhanced P300 relative to feedback following decisions to exploit. I interpreted this signal as indicative of a phasic release of the neuromodulator norepinephrine from locus coeruleus (Nieuwenhuis et al., 2005). I argued that this signal reflected the interruption of the dominant mode of decision making (exploitation) in favour of something new (exploration)[3].

However, exploitation was the more frequent strategy in these previously-used tasks. Thus, it was unclear whether the enhanced neural feedback processing associated with exploration was due to exploration per se, or to the fact that exploration was rare. The frequency of exploration relative to exploitation is of interest because rare events are known to elicit larger P300s compared to frequent events (Polich, 2007). To test these hypotheses, I had participants search for sparse but spatially-correlated rewards on a continuous two-dimensional map**.** By "sparse" I mean that the majority of possible

---

[3] Phasic norepinephrine as a neural interrupt signal is usually discussed in the context of target detection, not decision making generally (Dayan & Yu, 2006; Yu & Dayan, 2005).

responses resulted in little or no reward; my hope in designing the task this way was to encourage more exploration than exploitation. If the exploration-dependent P300 enhancement I observed previously was due to the relative infrequency of exploration, then I ought to observe an *exploitation*-dependent P300 enhancement in the current study. On the other hand, if my previous results replicate, I would conclude that the effect is not due to frequency, but rather to some other property of exploration.

Additionally, I manipulated the distribution of rewards across blocks. Participants were cued that they would encounter either many reward patches (a multi-patch environment) or one reward patch only (a single-patch environment). The purpose of this manipulation was to further test the hypothesis that the exploration-dependent P300 effect is due to the frequency of exploration relative to exploitation. Stimulus frequency is known to modulate P300 amplitude such that the more infrequent a stimulus is, the larger the P300 that is elicited (Duncan-Johnson & Donchin, 1977). My hope was that, like snails, my participants would explore more in highly patchy environments, and exploit more in less patchy environments. If a greater exploration-dependent P300 was observed when exploration was less frequent (i.e., in the single-patch environment), this would lend support to a frequency hypothesis. Cues were used to help encourage these behaviours, since previous research on our ability to adapt to different reward distributions is mixed (Constantino & Daw, 2015; Hutchinson et al., 2008).

Methods

Participants. Twenty-four university-aged participants (3 male, all right-handed, M_{age} = 21.0, 95% CI [19.3, 22.7] with no known neurological impairments and with normal or corrected-to-normal vision took part in the experiment. All of the participants

were volunteers who received credit in an undergraduate course for their participation. Additionally, participants were paid a performance-dependent bonus of up, M_{bonus} = $10.11, 95% CI [8.89, 11.35]. The participants provided informed consent approved by the Human Research Ethics Board at the University of Victoria.

Apparatus and procedure. Participants were seated 60 cm in front of a 22-inch LCD display (75 Hz, 2 ms response rate, 1680 by 1050 pixels, LG W2242TQ-GF, Seoul, South Korea). Visual stimuli were presented using the Psychophysics Toolbox Extension (Brainard, 1997; Pelli, 1997) for MATLAB (Version 8.3, Mathworks, Natick, USA). Participants were given written and verbal instructions to minimize head and eye movements throughout the experiment.

Participants played 40 rounds of "gold rush", a mining simulator in which the goal was to find as much gold as possible. Each round consisted of 20 trials. Participants used a mouse to select a map location at which to dig for gold. Participants were then shown a number from 1-100, representing the amount of gold they had found. The total amount of gold found was tracked for each round, and at the end of the experiment the participant was paid for their best round at a conversion rate of $0.01 per point.

Prior to beginning the experiment, participants were shown on-screen instructions indicating that the distribution of gold was spatially correlated. The distribution of gold was fixed within a round but changed between rounds. Two types of reward distribution were possible: single-patch, and multi-patch. In single-patch maps, rewards were concentrated at one map location. Multi-patch maps contained rewards concentrated at between four and six "peaks". The maximum reward at each peak was randomly chosen from a uniform distribution from 50 to 100. Peak locations were also randomly chosen

(uniform distribution). The distribution of rewards around each peak was Gaussian, computed using the MATLAB function mvnpdf (Statistics and Machine Learning Toolbox, Release 2014a, Mathworks, Natick). The Gaussian reward distributions were circular (i.e., identity covariance matrix). See Appendix B for participant instructions, and for examples of each map type.

Prior to each round, participants were shown a cue indicating whether the upcoming map was single-patch or multi-patch. The meaning of these cues was explained in the on-screen instructions (Appendix B). Participants completed two practice rounds – one for each reward distribution type. On each trial, participants were shown the outline of the map (the dig boundary), a centrally-presented fixation cross, and an 'x' at each previous dig location. After each practice round, participants were shown the underlying reward distribution, with their choices overlaid. During the experiment, participants were never shown the underlying reward distribution. See Figure 23 for a block/trial overview.

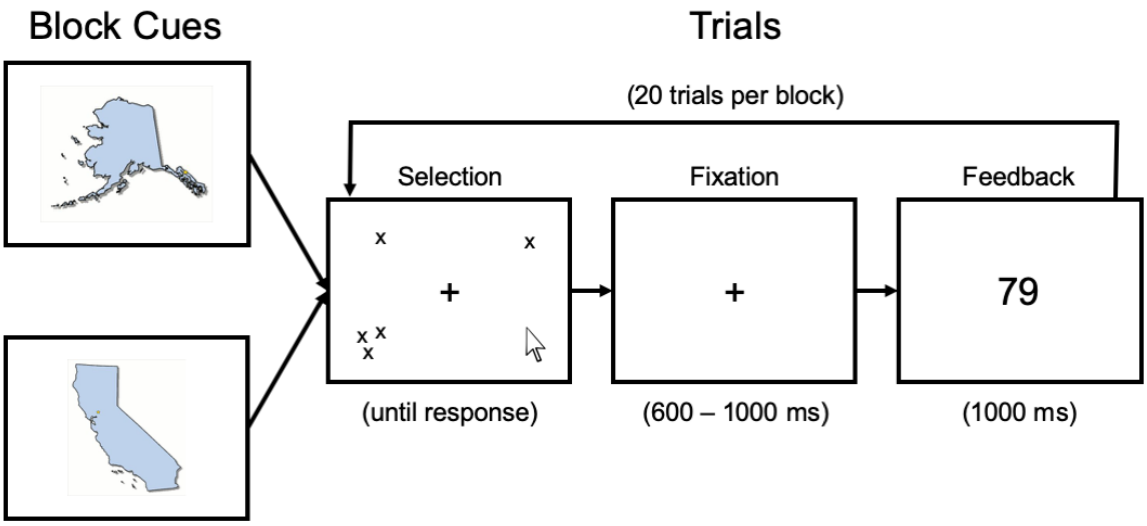


Figure 23. Task with timing details. Blocks started with a cue indicating the type of reward distribution (single-patch or multi-patch). Participants chose a dig location and were rewarded with an amount of gold from 1-100. Previous dig locations were marked on the map.

Data collection. Sixty-three channels of EEG data, referenced to channel AFz, were recorded using Brain Vision Recorder (Version 1.20, Brain Products GmbH, Munich, Germany). Sixty-one electrodes were placed in a fitted cap according to the 10-20 system. Additionally, two electrodes were attached to the left and right mastoids. Conductive gel was used to ensure that electrode impedances were below 20 kΩ prior to recording, and the EEG data were sampled at 500 Hz and amplified (actiCHamp, Brain Products GmbH, Munich, Germany).

Computational models. Several computational models were implemented in MATLAB and evaluated based on how well they accounted for my participants' chosen dig locations. The goal of this modelling was to classify trials as either exploitations or explorations. In general, exploitations were defined as trials for which a participant

responded in a value-maximizing way, e.g. choosing the location with the best-known reward. All other responses were explorations. Although several models were tested, only the best-fitting model was used to classify trials for my ERP analysis.

Participant decisions were classified as either exploitations or explorations using computational models. Several models were evaluated for their ability to account for my participants' decisions. First, each model was fit to each participant's data using the MATLAB function fmincon (Optimization Toolbox, Release 2018a, Mathworks, Natick). This function works by searching for parameters that minimize a specified objective function. In my case, the objective function was the negative log-likelihood of a participant's responses, given a particular model. Specifically, each model maintained a probability P_t associated with every possible action a on trial t (i.e., each location on the map). In practice, to reduce the computational complexity of my model-fitting procedure, I further discretized my 800 by 800 pixel maps to an 80 by 80 grid (6400 possible actions).

A good fit meant that the model was assigning high probabilities to a participant's actions (i.e., the chosen map locations). The trial-to-trial probabilities were combined according to the log-likelihood function:

$$-LL = -\sum_{t} \log(P_t(a_s))$$

where a_s was the selected action. I then computed the mean -*LL* across participants; the best-fitting model, defined as the one the one with the smallest mean –*LL*, was later used to classify trials as exploitations or explorations.

All of my models generated a probability associated with each map location. For all but one of my models (the win-stay, lose-shift model) this was done by first generating a value for each map location. The values *v* were then converted to action probabilities for each a_i and trial *t* according to the softmax equation:

$$P_t(a_i) = \frac{e^{v_t(i)/\tau}}{\sum_j e^{v_t(j)/\tau}}$$

where *i* was the index of the chosen action, *j* indexed over all possible actions, and τ (temperature) determined the degree of bias towards choosing high-valued locations. Next, I will describe how the values were computed for each function-approximation model.

Nearest-neighbours. This model computed a value for each map location using the nearest-neighbours approach (i.e., the value at a point was equal to the value of the closest previously-chosen point). The values were updated following feedback. In particular, I used MATLAB's griddata function with the "nearest" method. See Figure 24

for an illustration of how action probabilities were represented after sampling from an example reward distribution.

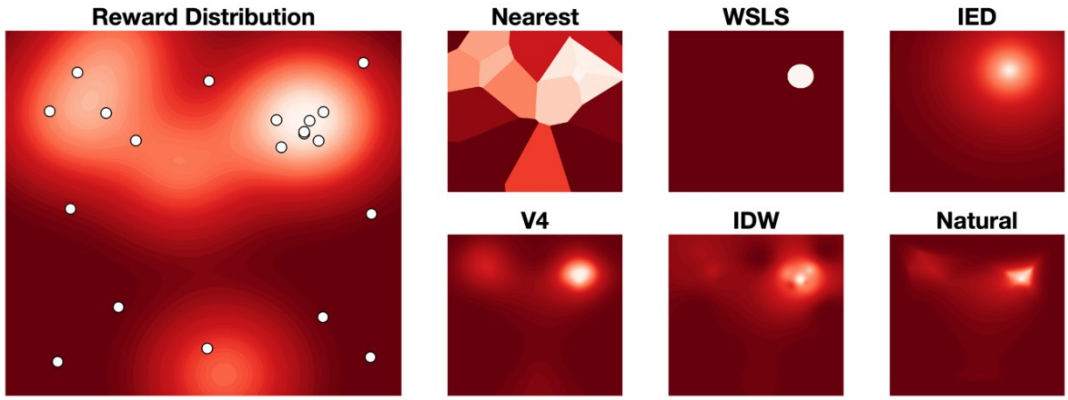


Figure 24. Sample responses and model representations. The patchiness of the underlying reward distribution was high in this case (left). The participant's responses are shown as white dots. Participant responses were modelled several ways – the final action probabilities are shown on the right (lighter areas were more likely to be chosen, according to the model).

Inverse Euclidean distance. Here, the value associated with each map location was defined as the inverse Euclidean distance from the previously-chosen location (i.e., a bias towards making the same action as before).

Spline interpolation. This model attempted to estimate the underlying reward distribution using the history of feedback. I again used MATLAB's griddata function, but with the "V4" method.

Inverse distance weighting. This model was similar to the inverse Euclidean distance model; action values were determined based on the inverse distance to each previously-chosen reward, weighted by the values of the previously-chosen rewards.

Natural neighbours. This model estimated the underlying reward distribution using the natural-neighbours method (MATLAB's griddata function with the "nearest" option), which provides a smoother interpolation compared to nearest-neighbours.

Win-stay, lose-shift. As mentioned, I also tested a win-stay, lose-shift model. Generally speaking, these models implement a simple heuristic that tends to repeat an action following a win but switch to a different action following a loss. My win-stay, lose-shift model assigned a single action probability ε to a radius *r* around the best-chosen option (the win-stay probability), and 1- ε to the rest of the map (C. M. Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018). See Figure 24.

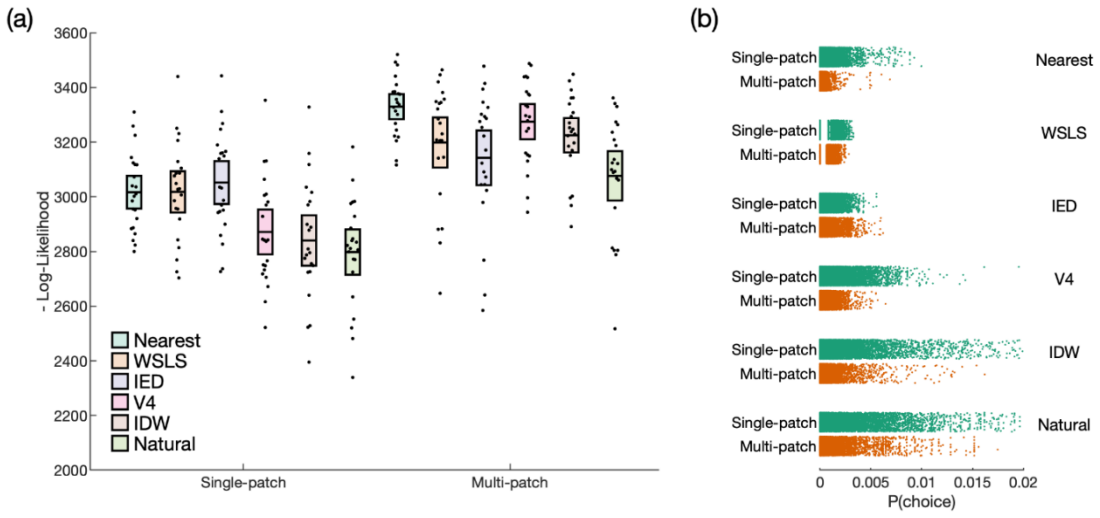


Figure 25. Model fit results. Most models maintained a value associated with each map location: nearest-neighbours, inverse Euclidean distance: IED, spline interpolation (V4), inverse distance weighting: IDW, and natural neighbours. The win-stay, lose-shift model (WSLS) tended to choose map locations close to the location of greatest previous reward. (a) The models provided comparable fits (lower is better). The natural-neighbours model provided the best mean fit in each environment. (b) Softmax probabilities – these are the model-generated likelihoods for each trial (all participants). Models that yielded better fits tended to generate greater trial-by-trial likelihoods.

Data analysis.

Modelling data.

Model tuning. All of my function-approximation models had a single tunable parameter: the softmax temperature, τ. The win-stay, lose-shift model had two tunable parameters: the win-stay probability, ε, and the affected radius, *r*. The model-fitting procedure described earlier (minimization via MATLAB's fmincon) yielded, for each participant and model, final *–LL* values, and final model parameters. A comparison of the mean *–LL* scores revealed that the natural-neighbours method provided the best fit for my

participants' data, regardless of block type (Figure 25a). This was the model I used to classify trials as exploitations or explorations.

All of my function-approximation models thus had a single tunable parameter: the softmax temperature, τ. The win-stay, lose-shift model had two tunable parameters: the win-stay probability, ε, and the affected radius, *r*. The model-fitting procedure described earlier (minimization via MATLAB's fmincon) yielded, for each participant and model, final *–LL* values, and final model parameters. A comparison of the mean *–LL* scores revealed that the natural-neighbours method provided the best fit for my participants' data, regardless of block type (Figure 25a). This was the model I used to classify trials as exploitations or explorations.

Trial classification. Previously, I classified a trial as an exploitation if the participant's choice matched the model's value-driven choice – i.e., the most likely action, according to the model. All other actions were considered to be explorations. Here, however, the action space was quite large (an 800 by 800 grid), so rather than focus on single actions I expanded my definition of exploitation to include a *range* of likely actions. This was possible because the model-generated action probabilities were continuous (see Figure 25b). For each participant and block patch type (single/multi) I computed the mean action probability. Trials with greater-than-average action probabilities were defined as exploitations; all other trials were explorations.

Behavioural data. For each participant and environment (single-patch/multi-patch) I computed the mean number of trials of each decision type (exploit/explore). This was also done on a trial-by-trial basis (i.e., for trial 2, 3, … 20). I then computed, for each

participant, environment (single-patch/multi-patch), and decision type (exploit/explore) the mean response time, displacement from previous response, and reward.

Electroencephalographic data. EEG data were downsampled to 250 Hz, filtered through a (0.1 Hz – 30 Hz pass band) phase shift-free Butterworth filter (60 Hz notch), and re-referenced to the average of the two mastoid channels. Next, ocular artifacts were removed using independent component analysis (ICA). In particular, ICA was used to identify components associated with eye movements. These components were then removed when the data were subsequently reconstructed. Subsequent to this, 800 ms epochs of EEG data were constructed from 200 ms prior to 800 ms following feedback onset. All trials were then baseline corrected using a 200 ms pre-feedback window. Finally, trials in which the change in voltage in any channel exceeded 10 μV per sampling point or the change in voltage across the epoch was greater than 100 μV were discarded. On average, I removed 28% of epochs (95% CI [23, 34]).

Examination of the grand-grand waveform. To avoid biasing my analysis in favour of a statistically-significant difference between my conditions of interest, I defined the P300 by first examining the "grand-grand" average waveform (Kappenman & Luck, 2016). For each participant, I averaged across all EEG epochs (regardless of condition), then averaged across participants. Next, I identified the time/locations at which the most positive-going deflection occurred (388 ms post-stimulus at electrode P4). To capture the apparent P300 deflection, I then identified the times at which 75% of the maximum voltage was reached (288 to 544 ms post feedback), which formed my analysis window (see dashed line in Figure 26c).

Effect of current/next trial decision. Previously, I analyzed the effect of the upcoming decision type (exploit/explore) on the P300 and reported an exploration-related enhancement (Experiments 1 and 2). Here, I was interested in the possibility that this neural signal may be affected by both the current and the upcoming trial type. I therefore combined trials across environments (single-patch/multi-patch) and examined the effect of current/next trial decision on the P300, defined in the same way as in my main analysis (the mean voltage from 288-544 ms post feedback at electrode P4). Four trial groupings (and four waveforms) were created for each participant based the current and next trial type: exploit-exploit, exploit-explore, explore-exploit, and explore-explore. See Figure 26a for each condition's trial count. I then conducted a 2 (current trial type: explore/exploit) by 2 (next trial type: explore/exploit) repeated measures ANOVA. The P300 effect here appeared to be driven mostly by the current trial type (see Results). For my main analysis, described below, I therefore decided to focus on the current-trial decision instead of the next-trial decision.

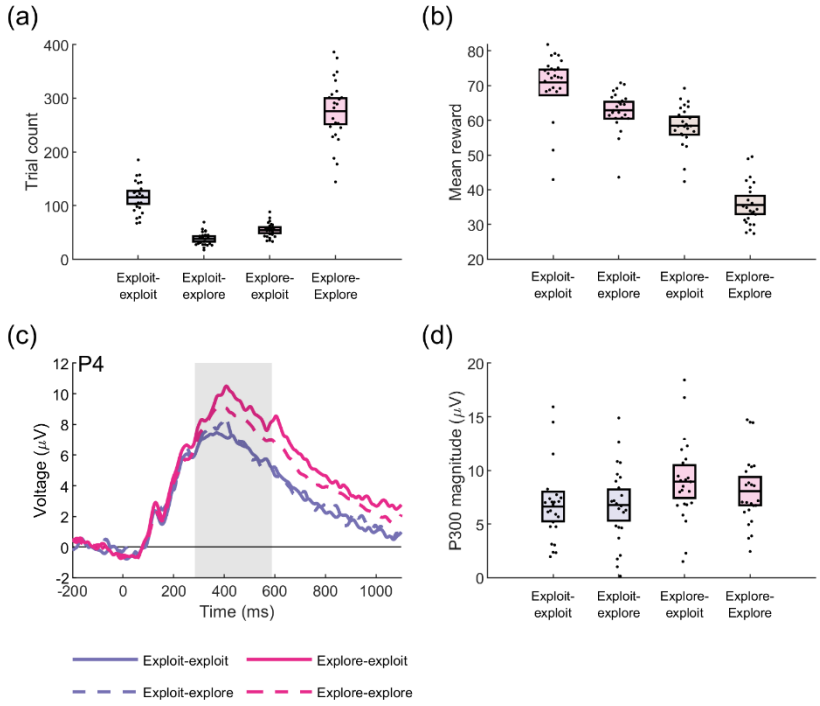


Figure 26. Behavioural and EEG data by current/next trial type. (a) Participants were more likely to exploit following exploitations, and more likely to explore following explorations. (b) Mean reward by current/next trial. (c) Feedback-locked waveforms for each current/next trial type. The shaded area shows the region of analysis. (d) P300 scores by current/next trial type.

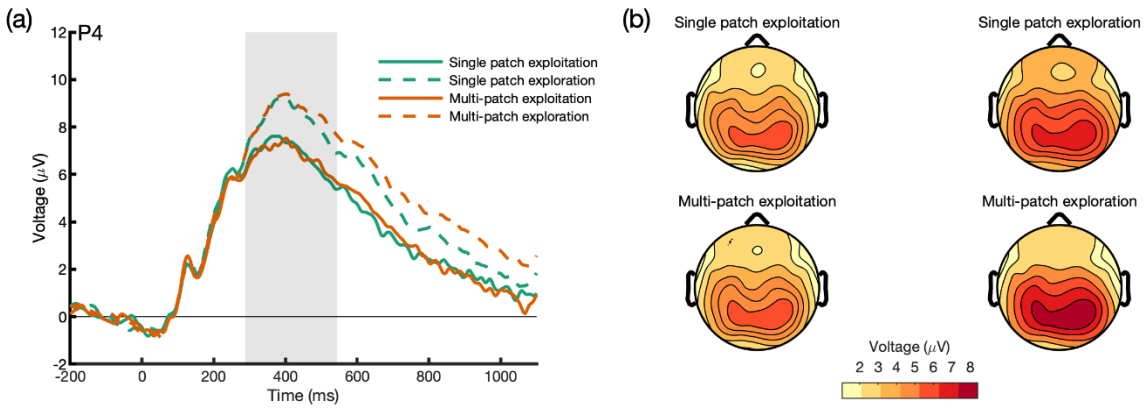


Figure 27. Feedback-locked waveforms (left) and scalp topographies (right).

Feedback-Locked P300. Conditional waveforms were created by averaging the feedback-locked EEG for each participant, environment (single-patch/multi-patch), and current-trial decision type (exploit/explore). Finally, a P300 was computed as the mean voltage within my analysis window (288 to 544 ms post feedback) at electrode P4, for each participant, block patch type (single/multi), and decision type (exploit/explore). See Figure 27 for the resulting waveforms and scalp topographies.

Feedback-locked late potential. Upon examining the feedback-locked waveforms (Figure 27), I noted that the effect of decision type (exploit/explore) on feedback processing appeared to be sustained well beyond the usual P300 time range. To investigate this difference, I averaged my waveforms across environment (single-patch/multi-patch) and constructed a difference wave (explore minus exploit) to define a second analysis window. As before, I located the time/location of the maximum voltage – of the difference wave, this time – and computed the interval within which 75% of this value was reached. This yielded a later time range, at a more central location: 440-804 ms post feedback at electrode POz (Figure 28).

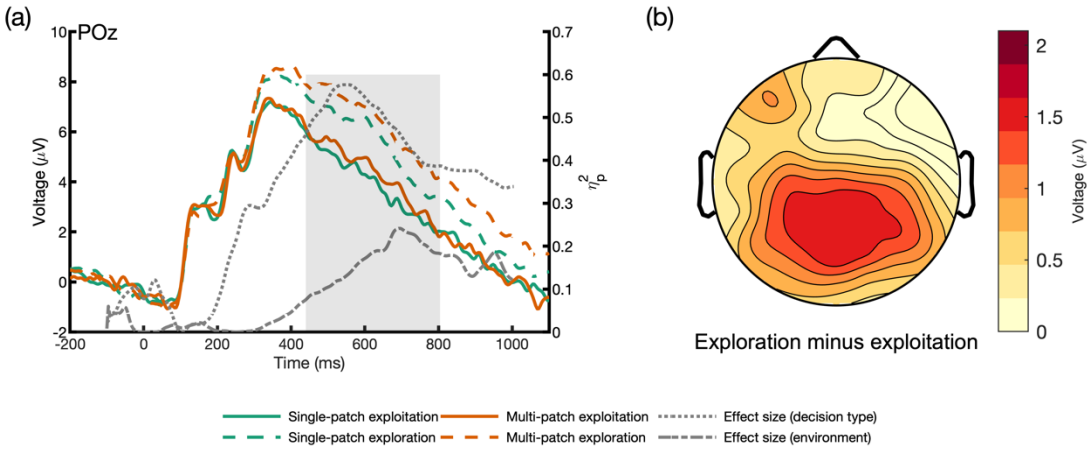


Figure 28. Feedback-locked waveforms (left) and scalp topography of the explore-minus-exploit difference scores (right). The grey lines show the decision/environment effect sizes computed on a 200-ms sliding window.

Inferential statistics. The effect of environment (single-patch/multi-patch) on exploration rate was determined using a paired-samples t-test. Cohen's *d* was computed according to:

$$d = \frac{M_{\text{diff}}}{s_{\text{diff}}}$$

where M_{diff} was the difference score mean and s_{diff} was the difference score standard deviation (Cumming, 2014). To determine whether exploration rate changed within a block, a linear model relating exploration rate to trial number (2, 3, … 20) was fit to each participant's data using the MATLAB function polyfit. The effect of trial number on the model slopes was assessed using a single-sample t-test (and Cohen's *d* computed by dividing the slope mean by the slope standard deviation). Next, my behavioural scores (mean response time, mean displacement, and mean reward) and ERP scores (P300, late potential) were subjected to a 2 (decision: exploit, explore) by 2 (environment: single-

patch, multi-patch) repeated-measures ANOVA. Two different effect-size measures were computed: η_p^2 and η_g^2 (Olejnik & Algina, 2003). To help illustrate how my effects of interest (decision, environment) changed over time, I computed η_p^2 for each on a 200 ms sliding window (grey lines in Figure 28).

Results

Modelling data. Participants explored slightly more in the multi-patch environment (69%, 95% CI [67, 71]) compared to the single-patch environment (66%, 95% CI [64, 68]), $t(23) = 3.1$, $p = .005$, Cohen's $d = 0.64$. I also noted that participants tended to explore less as they discovered the location of the rewards – the slope of the relationship between exploration rate and trial number was non-zero in both the single-patch environment, $t(23) = -29$, $p < .001$, Cohen's $d = -6.0$, and the multi-patch environment, $t(23) = -14$, $p < .001$, Cohen's $d = -4.1$. See Figure 29.

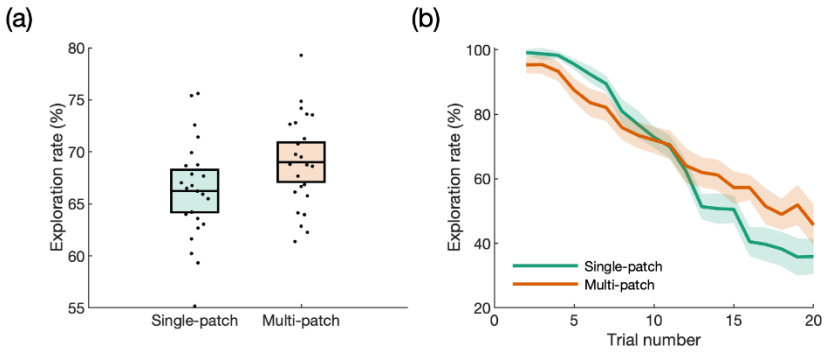


Figure 29. Trial classification. (a) Overall, participants explored more in the multi-patch environment. (b) The exploration rate decreased throughout a block as participants learned the reward locations. Error bars/shaded regions show 95% confidence intervals.

Behavioural data.

Effect of current/next trial decision. After collapsing across task to examine the effect of current/next trial type on trial counts, I observed no effect of current-trial type, $F(1,23) = 0.6$, $p = .5$, $\eta_p^2 = 0.02$, $\eta_g^2 = 0.00$ or next-trial type, $F(1,23) = 3.3$, $p = .08$, $\eta_p^2 = 0.12$, $\eta_g^2 = 0.01$. There was also no current-trial by next-trial interaction, $F(1,23) = 0.8$, $p = .4$, $\eta_p^2 = 0.03$, $\eta_g^2 = 0.00$. I also examined the effect of current/next trial type on reward and found an effect of both current-trial type (larger rewards for exploitations), $F(1,23) = 328$, $p < .001$, $\eta_p^2 = 0.93$, $\eta_g^2 = 0.79$, and next-trial type (larger for exploitations), $F(1,23) = 179$, $p < .001$, $\eta_p^2 = 0.88$, $\eta_g^2 = 0.70$. There was an interaction effect between current-trial type and next-trial type – it appeared to take less of a points difference to switch from exploitation to exploration than it took to switch from exploration to exploitation, $F(1,23) = 109$, $p < .001$, $\eta_p^2 = 0.83$, $\eta_g^2 = 0.34$. See Table 1 for exact values, and Figure 26.

Table 5

Effects of current/next trial decision

	Exploit (current)				Explore (current)			
	Exploit (next)		Explore (next)		Exploit (next)		Explore (next)	
Measure	*M*	95% CI	*M*	95% CI	*M*	95% CI	*M*	95% CI
Trial count	116	[103, 128]	38	[33, 43]	54	[49, 60]	277	[253, 301]
Reward (points)	72	[69, 75]	64	[62, 65]	59	[57, 61]	35	[33, 38]
P300 (μV)	6.6	[5.3, 8.0]	6.9	[5.3, 8.4]	9.0	[7.4, 10.6]	8.0	[6.7, 9.3]

Response time. Response times were affected by decision type, $F(1,23) = 13$, $p = .002$, $\eta_p^2 = 0.36$, $\eta_g^2 = 0.04$. There was no effect of environment, $F(1,23) = 0.7$, $p = .4$, $\eta_p^2 = 0.03$, $\eta_g^2 = 0.00$, and no decision by environment interaction, $F(1,23) = 1.1$, $p = .3$, $\eta_p^2 = 0.05$, $\eta_g^2 = 0.00$. See Table 2 and Figure 30 for mean response times.

Displacement. Decision type also affected displacement from previous choice – explorations covered a greater distance, $F(1,23) = 379$, $p < .001$, $\eta_p^2 = 0.94$, $\eta_g^2 = 0.82$. There was a smaller effect of environment (greater displacements in the multi-patch environment), $F(1,23) = 6.1$, $p = .02$, $\eta_p^2 = 0.21$, $\eta_g^2 = 0.04$. No interaction was detected, $F(1,23) = 0.2$, $p = .69$, $\eta_p^2 = 0.01$, $\eta_g^2 = 0.00$. See Table 1 and Figure 30 for mean displacements.

Reward. Exploitations resulted in greater point gains, on average, compared to explorations, $F(1,23) = 1077$, $p < .001$, $\eta_p^2 = 0.98$, $\eta_g^2 = 0.93$. The single-patch environment yielded more rewards compared to the multi-patch environment, $F(1,23) = 53$, $p < .001$, $\eta_p^2 = 0.70$, $\eta_g^2 = 0.37$. Finally, there was an interaction between decision and environment on reward; the points-advantage of exploiting over exploring appeared to be greatest in the single-patch environment, $F(1,23) = 7.5$, $p = .01$, $\eta_p^2 = 0.25$, $\eta_g^2 = 0.05$. See Table 2 and Figure 30.

Table 6

Behavioural means, with 95% confidence intervals

	Single-patch				Multi-patch			
	Exploit		Explore		Exploit		Explore	
Measure	*M*	95% CI	*M*	95% CI	*M*	95% CI	*M*	95% CI
Response time (ms)	426	[377, 476]	482	[429, 535]	427	[378, 476]	468	[409, 528]
Displacement (mm)	6.9	[5.5, 8.3]	48	[43, 52]	10	[8.5, 12]	52	[45, 59]
Reward (points)	75	[74, 76]	41	[38, 43]	66	[64, 68]	36	[34, 38]

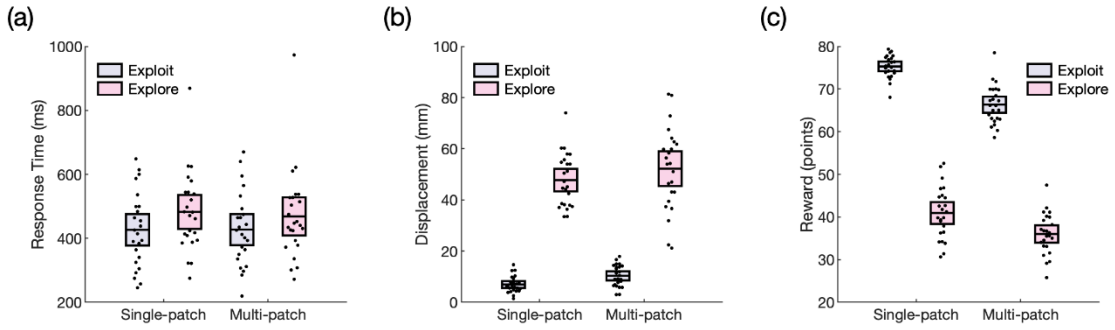


Figure 30. Behavioural results. Explorations were (a) slower and (b) farther from the previous choice. (c) Exploitation resulted in a greater average point gain.

Electroencephalographic data.

Effect of current/next trial decision. There was an effect of current-trial type ($F(1,23) = 7.5$, $p = .01$, $\eta_p^2 = 0.25$, $\eta_g^2 = 0.07$), but not next-trial type ($F(1,23) = 0.0$, $p = .9$, $\eta_p^2 = 0.00$, $\eta_g^2 = 0.00$). There was a current-trial by next-trial interaction, $F(1,23) = 12$, $p = .002$, $\eta_p^2 = 0.33$, $\eta_g^2 = 0.13$. Specifically, the effect of exploring on the current

trial appeared to be modulated by next-trial decision type (greater when the next trial was an exploitation). See Table 1 and Figure 26.

P300. There was an effect of decision type on the feedback-locked P300 (enhanced for explorations), $F(1,23) = 25$, $p < .001$, $\eta_p^2 = 0.52$, $\eta_g^2 = 0.05$. There was no effect of environment, $F(1,23) = 0.19$, $p = .66$, $\eta_p^2 = 0.01$, $\eta_g^2 = 0.00$, and no interaction, $F(1,23) = 0.80$, $p = .38$, $\eta_p^2 = 0.03$, $\eta_g^2 = 0.00$. See Table 3 for condition means.

Table 7

ERP scores, with 95% confidence intervals

	Single-patch				Multi-patch			
	Exploit		Explore		Exploit		Explore	
Measure	*M*	95% CI	*M*	95% CI	*M*	95% CI	*M*	95% CI
P300 (μV)	6.7	[5.1, 8.4]	8.2	[6.7, 9.8]	6.8	[5.2, 8.3]	8.5	[7.1, 10]
Late potential (μV)	3.9	[2.5, 5.3]	5.6	[4.1, 7.1]	4.3	[2.9, 5.7]	6.4	[4.8, 8.0]

Late potential. The late potential was affected by decision type, $F(1,23) = 25$, $p < .001$, $\eta_p^2 = 0.52$, $\eta_g^2 = 0.08$. There was also a small effect of environment, $F(1,23) = 5.4$, $p = .03$, $\eta_p^2 = 0.19$ $\eta_g^2 = 0.01$. No interaction was detected, $F(1,23) = 0.78$, $p = .39$, $\eta_p^2 = 0.03$, $\eta_g^2 = 0.00$. See Table 3.

Discussion

In this experiment, I observed an enhanced P300 for feedback following decisions to explore, a result previously observed when exploration was rare compared to exploitation (Experiments 1 and 2). Here I showed that exploration enhances the feedback-locked P300 even when exploration is frequent. Furthermore, the exploration-

related P300 appears to be unaffected by reward distribution knowledge, even though such knowledge affects exploration rate and choice behaviour.

I began by testing how well several models could account for my participants' trial-to-trial decisions. Although model comparison was not my main goal, the model-fitting procedure itself was important because this was how I classified trials as exploitations or explorations. By choosing the best of several models, I gained confidence in my trial classification. I discovered that a model that approximated the underlying value function provided the best fit for my data (in particular, using the natural-neighbours approach). This discovery is in line with work by Wu and colleagues (2018) who, after comparing many different models, found evidence that humans rely on function approximation to find spatially correlated rewards in a large decision space (an 11-by-11 grid). I have shown here that this finding holds true in a more continuous space (an 800-by-800 grid).

After classifying participant decisions as exploitations or explorations, I confirmed two critical features of my experiment. First, and in contrast to earlier work, explorations were more common than exploitations — a feature that allowed us to test whether or not frequent exploration would elicit a P300 enhancement (discussed below). Second, I verified that my between-block manipulation had worked; participants explored more when they were shown a multi-patch cue compared to when they were shown a single-patch cue[4]. In other words, my participants' decisions to explore were influenced by the reward distribution. This observation is in line with some (Constantino & Daw,

---

[4] Though significant, this effect was small, and we noted considerable inter-participant variability in exploration rate (Figure 29a).

2015), but not all previous work (Hutchinson et al., 2008). Furthermore, explorations were slower compared to exploitations, in line with other studies (Beharelle et al., 2015; Experiment 1). This is somewhat unsurprising here, given that explorations covered a greater distance than exploitations. Unlike these previous studies, my participants were told the upcoming reward distribution, and adjusted their strategy accordingly. It remains to be seen whether my task would elicit these adaptations if the nature of the reward distribution was initially unknown. Others (Wu et al., 2018) have found that humans assume smooth, spatially-correlated reward distributions (which ours are). But would naïve participants assume the presence of one reward patch or many reward patches?

In line with Experiments 1 and 2, my examination of the neural response to feedback revealed an exploration-related P300. In particular, the feedback-locked P300 was greater following exploration than following exploitation. This is noteworthy because, unlike previous work examining the exploration-related P300, exploration in my experiment was the more common decision type. Thus, these results rule out the possibility that the exploration-related P300 is driven entirely by the frequency of exploration relative to exploitation. Furthermore, an exploratory analysis identified that the observed exploration-related P300 likely peaked around 600 ms post feedback – later than what is usually associated with the P300, but not unheard-of (Polich, 2007). In the remaining discussion, my goal will be to identify the likely contributors to this neural signal.

One potential contributor relates to the *late positive potential* (LPP), a P300-like ERP component linked to motivational relevance (Olofsson, Nordin, Sequeira, & Polich, 2008; Schupp et al., 2000). The LPP peaks between 400 and 800 ms post stimulus, at

central or parietal locations, and is typically evoked by emotional images; pleasant and unpleasant images elicit larger LPPs compared to neutral images (Schupp et al., 2000). The LPP is thought to reflect the engagement of a general motivational system in the brain – general because it is sensitive not only to emotional images, but also to task-relevant features (Bradley, 2009). For example, the LPP is enhanced if participants are asked to count emotional but not neutral images (Ferrari, Codispoti, Cardinale, & Bradley, 2008). Relevant here, the LPP is still present after repeated viewings of the same stimulus (Codispoti, Ferrari, & Bradley, 2007). This allows for the possibility of an LPP for feedback following exploration, even though exploration is frequent, provided that such exploratory feedback is more motivating than exploitative feedback. Although I did not manipulate or test for level of motivation here, this approach might prove promising in the future as previous research has shown a link between overall task involvement and the feedback-locked P300 (Yeung, Holroyd, & Cohen, 2005).

Here I have shown that exploration in continuous environments is followed by enhanced feedback processing, even when exploration is the dominant strategy. I suggest that this effect is driven mainly by the neural processes required to switch from exploration to exploitation (a neural interrupt signal). These neural processes are general; they operate across different task types (discrete and continuous) and exploration rates (rare and common).

Chapter 6: General Discussion

The main goal of this research was to understand how the brain represents and implements the trade-off between exploring the world and exploiting previous learning. My work suggests that transitioning from one decision type to another is a significant neural event that is triggered in part by external feedback. In particular, I observed an enhanced neural signal both *prior to* exploration (Experiments 1 and 2) and *following* exploration (Experiments 3 and 4). The reason for this discrepancy was that Experiments 1 and 2 were based on earlier studies that aimed to use EEG and machine learning to predict whether participants *would* explore or exploit (Bourdaud et al., 2008; Tzovara et al., 2012). Thus, I chose to focus on the previous trial's neural response (as others had done).

In Experiments 3 and 4 I decided to test whether there was more to the story. I discovered that the feedback-locked P300 was mostly driven by the decision made just prior to feedback (enhanced for exploration), but that our past decisions interact with our future decisions; the effect is greatest when exploration is followed by exploitation. Furthermore, decisions tended to cluster by type – explorations are more likely to be followed by explorations, and exploitations are more likely to be followed by exploitations (Figures 16a and 26a). This result suggests that the effects seen in Experiments 1 and 2 (and the machine learning predictions made in related work) may have been driven by the current trial in addition to the previous trial.

The results from Experiments 1–4 suggest that neural interruption, as indexed by P300 enhancement, provides a general mechanism by which we switch from exploring our environment to exploiting what we have learned. But what, exactly, triggers the shift

from exploration to exploitation? According to Dayan and Yu (2006), neural interruption occurs when the likelihood of ongoing events is incompatible with our internal model of the world. For example, we expect the weather report to be incorrect some proportion of the time. However, if it is incorrect enough times in a row, we might become suspicious that our weather app is set to the wrong city. A neural interrupt, signalled by a phasic release of NE, prompts us to change our behaviour (in this case, update our weather app). I argue here that the same neural system – neural interruption – controls transitions from exploration to exploitation. Feedback value no doubt plays a role; when we find a large reward while exploring, we tend to switch to an exploitative mode (Experiments 3 and 4). Or, if an option is no longer yielding the reward that it used to, we may choose to explore (Experiment 3).

However, since exploitation tends to be the more rewarding strategy overall, feedback value alone is not enough to explain my results. I also note that although rate of exploration plays a role (Experiment 3), the exploration-related P300 is still present even when exploration is the dominant strategy (Experiment 4). The signal is present in both a risk-taking task (Experiment 1) and a feedback-based learning task (Experiments 2–4). It is elicited in both discrete tasks (Experiments 1–3) and a continuous task (Experiment 4). In other words, there is something special about exploration per se, and in particular about the transition from exploration to exploitation.

If the exploration-related P300 reflects a neural-interrupt signal, then it is not (as I had originally thought) due to the frequency of exploration relative to exploitation (Experiments 1 and 2). In other words, if I am to maintain a neural-interrupt explanation for my results, I can no longer define as default whichever decision type (exploit/explore)

is more frequent. Instead, I must consider that exploration may always be the default strategy that requires neural interrupt – and concomitant P300 enhancement – in order to switch to an exploitative mode of decision making. The claim that exploration is always the default strategy is difficult to test because it is not clear what exploration and exploitation mean outside of a laboratory/task context. However, it has been suggested that mind-wandering may be a form of exploration, and goal-directed thinking a form of exploitation (Sripada, 2018). Although rates of mind-wandering vary from a third (M. J. Kane et al., 2007) to half of our daily lives (Killingsworth & Gilbert, 2010), mind-wandering may be related to the brain's default state (the default mode network, or DMN: Raichle, 2015). Thus, there are theoretical reasons to suspect that switching from exploration to exploitation always requires neural interruption, regardless of task.

One issue with this explanation is that it may not be possible to characterize a decision-making event (switching from exploration to exploitation) by examining feedback processing only. In other words, examining feedback processing via the feedback-locked P300 may not tell us much about decision-making processes if those processes are not time-locked to an external event. Other techniques, such as examining time-frequency aspects of the EEG, may prove more promising. If the feedback-locked P300 does not relate directly to a decision to explore, then what explains my results? An alternative explanation relates to the fact that the P300 is sensitive to the information content of feedback. For example, Cockburn and Holroyd (2017) found that making feedback more informative results in ERP enhancements in the P300/LPP time range and scalp location. This observation is consistent with my results, assuming that exploratory

feedback in my tasks was more uncertain compared to exploitative feedback (see Figure 21 for evidence that this may be the case).

But is exploratory feedback always more uncertain than exploitative feedback? Future experiments might resolve this question if feedback uncertainty can be manipulated independently of whether a participant explores or exploits. This could be done using a forced-choice design (forced exploration/exploitation). Or, outcome uncertainty could be manipulated across blocks to determine the extent to which the feedback-locked P300 is driven by uncertainty or by exploration itself (e.g., via a comparison of effect sizes). If more of the P300 variance turned out to be explainable by uncertainty it would suggest that my observed effects were somewhat incidental (especially for Experiments 2–4, which focussed on feedback processing).

In summary, this work contributes to our understanding of the neural basis of the decision-making trade-off between exploration and exploitation. Across four experiments, exploration was consistently followed by enhanced feedback processing, especially when transitioning to an exploitative mode. These results suggest that neural interruption – the halting of one set of cognitive processes in favour of another – may play a critical role in resolving the explore-exploit dilemma generally. Prior evidence for this claim has been inconclusive due to the paucity of EEG studies in this area, and the low ecological validity of existing foraging tasks. Hopefully, this work has highlighted the value of (a) EEG, for examining the time-course of decision-making, (b) computational modelling, for informing EEG analyses, and (c) using ecologically-valid decision-making tasks, when possible.

Bibliography

Ahn, W.-Y., Busemeyer, J. R., Wagenmakers, E.-J., & Stout, J. C. (2008). Comparison of Decision Learning Models Using the Generalization Criterion Method. *Cognitive Science*, *32*(8), 1376–1402. https://doi.org/10.1080/03640210802352992

Aston-Jones, G., Rajkowski, J., Kubiak, P., & Alexinsky, T. (1994). Locus coeruleus neurons in monkey are selectively activated by attended cues in a vigilance task. *Journal of Neuroscience*, *14*(7), 4467–4480. https://doi.org/10/gf2zp6

Aston-Jones, G., & Bloom, F. E. (1981). Activity of norepinephrine-containing locus coeruleus neurons in behaving rats anticipates fluctuations in the sleep-waking cycle. *The Journal of Neuroscience*, *1*(8), 876–886.

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*, 403–450. https://doi.org/10/bztcf6

Aston-Jones, G., Rajkowski, J., & Kubiak, P. (1997). Conditioned responses of monkey locus coeruleus neurons anticipate acquisition of discriminative behavior in a vigilance task. *Neuroscience*, *80*(3), 697–715.

Baker, T. E., & Holroyd, C. B. (2011). Dissociated roles of the anterior cingulate cortex in reward and conflict processing as revealed by the feedback error-related negativity and N200. *Biological Psychology*, *87*(1), 25–34. https://doi.org/10.1016/j.biopsycho.2011.01.010

Beeler, J. A. (2012). Thorndike's Law 2.0: Dopamine and the Regulation of Thrift. *Frontiers in Neuroscience*, *6*. https://doi.org/10/gfkpgj

Beeler, J. A., Daw, N. D., Frazier, C. R. M., & Zhuang, X. (2010). Tonic Dopamine Modulates Exploitation of Reward Learning. *Frontiers in Behavioral Neuroscience*, *4*. https://doi.org/10/b6qhtm

Beharelle, A. R., Polanía, R., Hare, T. A., & Ruff, C. C. (2015). Transcranial Stimulation over Frontopolar Cortex Elucidates the Choice Attributes and Neural Mechanisms Used to Resolve Exploration–Exploitation Trade-Offs. *Journal of Neuroscience*, *35*(43), 14544–14556. https://doi.org/10.1523/JNEUROSCI.2322-15.2015

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. https://doi.org/10/ddsv2g

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

Berridge, C. W., & Waterhouse, B. D. (2003). The locus coeruleus–noradrenergic system: Modulation of behavioral state and state-dependent cognitive processes. *Brain Research Reviews*, *42*(1), 33–84.

Blanchard, T. C., & Gershman, S. J. (2018). Pure correlates of exploration and exploitation in the human brain. *Cognitive, Affective, & Behavioral Neuroscience*, *18*(1), 117–126. https://doi.org/10/gf3ksf

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652. https://doi.org/10/cbwggz

Bourdaud, N., Chavarriaga, R., Galan, F., & Millan, J. d R. (2008). Characterizing the EEG Correlates of Exploratory Behavior. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *16*(6), 549–556. https://doi.org/10.1109/TNSRE.2008.926712

Bouret, S., & Sara, S. J. (2005). Network reset: A simplified overarching theory of locus coeruleus noradrenaline function. *Trends in Neurosciences*, *28*(11), 574–582. https://doi.org/10.1016/j.tins.2005.09.002

Bowlby, J. (1988). *A secure base: Parent-child attachment and healthy human development*. New York: Basic Books.

Bradley, M. M. (2009). Natural selective attention: Orienting and emotion. *Psychophysiology*, *46*(1), 1–11. https://doi.org/10/fpbzmw

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Busemeyer, J. R., & Pleskac, T. J. (2009). Theoretical tools for understanding and aiding dynamic decision making. *Journal of Mathematical Psychology*, *53*(3), 126–138. https://doi.org/10/dnphmb

Calhoun, A. J., Tong, A., Pokala, N., Fitzpatrick, J. A. J., Sharpee, T. O., & Chalasani, S. H. (2015). Neural Mechanisms for Evaluating Environmental Variability in Caenorhabditis elegans. *Neuron*, *86*(2), 428–441. https://doi.org/10/f68w3w

Cavanagh, J. F., Figueroa, C. M., Cohen, M. X., & Frank, M. J. (2011). Frontal Theta Reflects Uncertainty and Unexpectedness during Exploration and Exploitation. *Cerebral Cortex*, bhr332. https://doi.org/10.1093/cercor/bhr332

Chase, H. W., Swainson, R., Durham, L., Benham, L., & Cools, R. (2010). Feedback-related Negativity Codes Prediction Error but Not Behavioral Adjustment during

Probabilistic Reversal Learning. *Journal of Cognitive Neuroscience*, *23*(4), 936–946. https://doi.org/10.1162/jocn.2010.21456

Chase, J. M., Wilson, W. G., & Richards, S. A. (2001). Foraging trade-offs and resource patchiness: Theory and experiments with a freshwater snail community. *Ecology Letters*, *4*(4), 304–312. https://doi.org/10/c4wz6j

Chernev, A., Böckenholt, U., & Goodman, J. (2015). Choice overload: A conceptual review and meta-analysis. *Journal of Consumer Psychology*, *25*(2), 333–358. https://doi.org/10/f68vfs

Clayton, E. C., Rajkowski, J., Cohen, J. D., & Aston-Jones, G. (2004). Phasic Activation of Monkey Locus Ceruleus Neurons by Simple Decisions in a Forced-Choice Task. *Journal of Neuroscience*, *24*(44), 9914–9920. https://doi.org/10.1523/JNEUROSCI.2446-04.2004

Cockburn, J., & Holroyd, C. B. (2017). Feedback information and the reward positivity. *International Journal of Psychophysiology*. https://doi.org/10/gfdtbg

Codispoti, M., Ferrari, V., & Bradley, M. M. (2007). Repetition and Event-related Potentials: Distinguishing Early and Late Processes in Affective Picture Perception. *Journal of Cognitive Neuroscience*, *19*(4), 577–586. https://doi.org/10/dgnmzn

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 933–942. https://doi.org/10/ff7k65

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

Crawley, M. J., & Krebs, J. R. (1992). Foraging Theory. In M. J. C. FIBiol FLS (Ed.), *Natural Enemies* (pp. 90–114). https://doi.org/10.1002/9781444314076.ch4

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. https://doi.org/10/frvfpv

Dayan, P., & Yu, A. J. (2006). Phasic norepinephrine: A neural interrupt signal for unexpected events. *Network: Computation in Neural Systems*, *17*(4), 335–350. https://doi.org/10.1080/09548980601004024

Donchin, E. (1981). Surprise!… Surprise? *Psychophysiology*, *18*(5), 493–513. https://doi.org/10/ff2cmm

Donchin, E., & Coles, M. G. (1988). Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences*, *11*(03), 357–374. https://doi.org/10/cws2jr

Duncan, C. C., Barry, R. J., Connolly, J. F., Fischer, C., Michie, P. T., Näätänen, R., … Van Petten, C. (2009). Event-related potentials in clinical research: Guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400. *Clinical Neurophysiology*, *120*(11), 1883–1908. https://doi.org/10/d3gz5h

Duncan-Johnson, C. C., & Donchin, E. (1977). On Quantifying Surprise: The Variation of Event-Related Potentials With Subjective Probability. *Psychophysiology*, *14*(5), 456–467. https://doi.org/10/c34pf5

Enriquez-Geppert, S., Konrad, C., Pantev, C., & Huster, R. J. (2010). Conflict and inhibition differentially affect the N200/P300 complex in a combined go/nogo and stop-signal task. *NeuroImage*, *51*(2), 877–887. https://doi.org/10/d98vjp

Feeney, B. C. (2004). A Secure Base: Responsive Support of Goal Strivings and Exploration in Adult Intimate Relationships. *Journal of Personality and Social Psychology*, *87*(5), 631–648. https://doi.org/10/dcxshf

Ferrari, V., Codispoti, M., Cardinale, R., & Bradley, M. M. (2008). Directed and Motivated Attention during Processing of Natural Scenes. *Journal of Cognitive Neuroscience*, *20*(10), 1753–1761. https://doi.org/10/bqcqfw

Foti, D., & Hajcak, G. (2012). Genetic variation in dopamine moderates neural response during reward anticipation and delivery: Evidence from event-related potentials. *Psychophysiology*, *49*(5), 617–626. https://doi.org/10/gfvd4j

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience*, *12*(8), 1062–1068. https://doi.org/10.1038/nn.2342

Gajewski, P. D., Drizinsky, J., Zülch, J., & Falkenstein, M. (2016). ERP Correlates of Simulated Purchase Decisions. *Frontiers in Neuroscience*, *10*. https://doi.org/10/gfv8cg

Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus

coeruleus function. *Cognitive, Affective & Behavioral Neuroscience*, *10*(2), 252–269. https://doi.org/10/frkp6k

Gittins, J., & Jones, D. (1974). A Dynamic Allocation Index for the Sequential Design of Experiments. In J. Gani (Ed.), *Progress in Statistics* (pp. 241–266). North-Holland.

Grace, A. A., Floresco, S. B., Goto, Y., & Lodge, D. J. (2007). Regulation of firing of dopaminergic neurons and control of goal-directed behaviors. *Trends in Neurosciences*, *30*(5), 220–227. https://doi.org/10.1016/j.tins.2007.03.003

Gratton, G., Coles, M. G. H., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology*, *55*(4), 468–484. https://doi.org/10/dtzns5

Grinband, J., Savitskaya, J., Wager, T. D., Teichert, T., Ferrera, V. P., & Hirsch, J. (2011). The dorsal medial frontal cortex is sensitive to time on task, not response conflict or error likelihood. *NeuroImage*, *57*(2), 303–311. https://doi.org/10/d935zb

Gureckis, T. M., & Love, B. C. (2009). Short Term Gains, Long Term Pains: How Cues About State Aid Learning in Dynamic Environments. *Cognition*, *113*(3), 293–313. https://doi.org/10/ct3fsx

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

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

Heekeren, H. R., Marrett, S., & Ungerleider, L. G. (2008). The neural systems that mediate human perceptual decision making. *Nature Reviews Neuroscience*, *9*(6), 467–479. https://doi.org/10/cpqw75

Hendricks, M. (2015). Neuroecology: Tuning Foraging Strategies to Environmental Variability. *Current Biology*, *25*(12), R498–R500. https://doi.org/10/gf3xpx

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679–709. https://doi.org/10.1037//0033-295X.109.4.679

Holroyd, C. B., & Krigolson, O. E. (2007). Reward prediction error signals associated with a modified time estimation task. *Psychophysiology*, *44*(6), 913–917. https://doi.org/10/fn4n7x

Holroyd, C. B., & McClure, S. M. (2015). Hierarchical control over effortful behavior by rodent medial frontal cortex: A computational model. *Psychological Review*, *122*(1), 54–83. https://doi.org/10.1037/a0038339

Holroyd, C. B., Pakzad-Vaezi, K. L., & Krigolson, O. E. (2008). The feedback correct-related positivity: Sensitivity of the event-related brain potential to unexpected positive feedback. *Psychophysiology*, *45*(5), 688–697. https://doi.org/10.1111/j.1469-8986.2008.00668.x

Holroyd, C. B., & Yeung, N. (2012). Motivation of extended behaviors by anterior cingulate cortex. *Trends in Cognitive Sciences*, *16*(2), 122–128. https://doi.org/10.1016/j.tics.2011.12.008

Hong, L., Walz, J. M., & Sajda, P. (2014). Your Eyes Give You Away: Prestimulus Changes in Pupil Diameter Correlate with Poststimulus Task-Related EEG Dynamics. *PLOS ONE*, *9*(3), e91321. https://doi.org/10/f52t35

Hutchinson, J. M. C., Wilke, A., & Todd, P. M. (2008). Patch leaving in humans: Can a generalist adapt its rules to dispersal of items across patches? *Animal Behaviour*, *75*(4), 1331–1349. https://doi.org/10/fmk7px

Ishii, S., Yoshida, W., & Yoshimoto, J. (2002). Control of exploitation–exploration meta-parameter in reinforcement learning. *Neural Networks*, *15*(4), 665–687.

Ivan, V. E., Banks, P. J., Goodfellow, K., & Gruber, A. J. (2018). Lose-Shift Responding in Humans Is Promoted by Increased Cognitive Load. *Frontiers in Integrative Neuroscience*, *12*. https://doi.org/10/gc77nv

Jepma, M., & Nieuwenhuis, S. (2011). Pupil diameter predicts changes in the exploration-exploitation trade-off: Evidence for the adaptive gain theory. *Journal of Cognitive Neuroscience*, *23*(7), 1587–1596. https://doi.org/10/dtsgbb

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, & Behavioral Neuroscience*, *17*(6), 1073–1083. https://doi.org/10/gfwv2h

Kane, M. J., Brown, L. H., McVay, J. C., Silvia, P. J., Myin-Germeys, I., & Kwapil, T. R. (2007). For whom the mind wanders, and when an experience-sampling study of

working memory and executive control in daily life. *Psychological Science*, *18*(7), 614–621. https://doi.org/10/cwt295

Kappenman, E. S., & Luck, S. J. (2016). Best Practices for Event-Related Potential Research in Clinical Populations. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *1*(2), 110–115. https://doi.org/10/gfz986

Kayser, A. S., Mitchell, J. M., Weinstein, D., & Frank, M. J. (2015). Dopamine, Locus of Control, and the Exploration-Exploitation Tradeoff. *Neuropsychopharmacology*, *40*(2), 454–462. https://doi.org/10/f6rtrm

Khamassi, M., Wilson, C. R. E., Rothé, M., Quilodran, R., Dominey, P. F., & Procyk, E. (2011). Meta-Learning, Cognitive Control, and Physiological Interactions between Medial and Lateral Prefrontal Cortex. In R. B. Mars, J. Sallet, M. F. S. Rushworth, & N. Yeung (Eds.), *Neural Basis of Motivational and Cognitive Control* (pp. 350–369). https://doi.org/10.7551/mitpress/9780262016438.003.0019

Kiat, J., Straley, E., & Cheadle, J. E. (2016). Escalating risk and the moderating effect of resistance to peer influence on the P200 and feedback-related negativity. *Social Cognitive and Affective Neuroscience*, *11*(3), 377–386. https://doi.org/10/f8h7gq

Killingsworth, M. A., & Gilbert, D. T. (2010). A Wandering Mind Is an Unhappy Mind. *Science*, *330*(6006), 932–932. https://doi.org/10/fw6xg5

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception*, *36*(14), 1.

Kolling, N., Behrens, T. E. J., Mars, R. B., & Rushworth, M. F. S. (2012). Neural Mechanisms of Foraging. *Science*, *336*(6077), 95–98. https://doi.org/10/f4n5jk

Krigolson, O. E. (2017). Event-related brain potentials and the study of reward processing: Methodological considerations. *International Journal of Psychophysiology*. https://doi.org/10.1016/j.ijpsycho.2017.11.007

Krigolson, O. E., Hassall, C. D., & Handy, T. C. (2014). How We Learn to Make Decisions: Rapid Propagation of Reinforcement Learning Prediction Errors in Humans. *Journal of Cognitive Neuroscience*, *26*(3), 635–644. https://doi.org/10/gd32z4

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00863

Laureiro-Martínez, D., Brusoni, S., & Zollo, M. (2010). The neuroscientific foundations of the exploration−exploitation dilemma. *Journal of Neuroscience, Psychology, and Economics*, *3*(2), 95–115. https://doi.org/10/cxc8bx

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., … Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, *8*(2), 75–84. https://doi.org/10.1037//1076-898X.8.2.75

Lenow, J. K., Constantino, S. M., Daw, N. D., & Phelps, E. A. (2017). Chronic and Acute Stress Promote Overexploitation in Serial Decision Making. *Journal of Neuroscience*, *37*(23), 5681–5689. https://doi.org/10/gbhw5g

Luck, S. J. (2014). *An introduction to the event-related potential technique* (Second edition). Cambridge, Massachusetts: The MIT Press.

March, J. G. (1996). Exploration and exploitation in organizational learning. *Organizational Learning*, 101–123.

Marco-Pallarés, J., Cucurell, D., Cunillera, T., Krämer, U. M., Càmara, E., Nager, W., … Rodriguez-Fornells, A. (2009). Genetic Variability in the Dopamine System (Dopamine Receptor D4, Catechol-O-Methyltransferase) Modulates Neurophysiological Responses to Gains and Losses. *Biological Psychiatry*, *66*(2), 154–161. https://doi.org/10/cgdhfq

Mars, R. B., Sallet, J., Rushworth, M. F. S., & Yeung, N. (2011). *Neural Basis of Motivational and Cognitive Control*. MIT Press.

McClure, S., Gilzenrat, M. S., & Cohen, J. D. (2006). An exploration-exploitation model based on norepinepherine and dopamine activity. *Advances in Neural Information Processing Systems*, *18*, 867.

Mobbs, D., Trimmer, P. C., Blumstein, D. T., & Dayan, P. (2018). Foraging for foundations in decision neuroscience: Insights from ethology. *Nature Reviews Neuroscience*, *19*(7), 419. https://doi.org/10.1038/S41583-018-0010-7

Mückschel, M., Chmielewski, W., Ziemssen, T., & Beste, C. (2017). The norepinephrine system shows information-content specific properties during cognitive control – Evidence from EEG and pupillary responses. *NeuroImage*, *149*, 44–52. https://doi.org/10/gf2bn9

Murphy, P. R., Robertson, I. H., Balsters, J. H., & O'connell, R. G. (2011). Pupillometry and P3 index the locus coeruleus–noradrenergic arousal function in humans. *Psychophysiology*, *48*(11), 1532–1543. https://doi.org/10/ftx2fg

Nieuwenhuis, S., Aston-Jones, G., & Cohen, J. D. (2005). Decision making, the P3, and the locus coeruleus—Norepinephrine system. *Psychological Bulletin*, *131*(4), 510–532. https://doi.org/10/b3mh34

Nieuwenhuis, S., De Geus, E. J., & Aston-Jones, G. (2011). The anatomical and functional relationship between the P3 and autonomic components of the orienting response: P3 and orienting response. *Psychophysiology*, *48*(2), 162–175. https://doi.org/10.1111/j.1469-8986.2010.01057.x

Nieuwenhuis, S., Yeung, N., Wildenberg, W. van den, & Ridderinkhof, K. R. (2003). Electrophysiological correlates of anterior cingulate function in a go/no-go task: Effects of response conflict and trial type frequency. *Cognitive, Affective, & Behavioral Neuroscience*, *3*(1), 17–26. https://doi.org/10.3758/CABN.3.1.17

Niv, Y., Daw, N. D., Joel, D., & Dayan, P. (2007). Tonic dopamine: Opportunity costs and the control of response vigor. *Psychopharmacology*, *191*(3), 507–520. https://doi.org/10/b6fhmh

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, *8*(4), 434–447. https://doi.org/10.1037/1082-989X.8.4.434

Olofsson, J. K., Nordin, S., Sequeira, H., & Polich, J. (2008). Affective picture processing: An integrative review of ERP findings. *Biological Psychology*, *77*(3), 247–265. https://doi.org/10/bbdfp6

Otto, A. R., Taylor, E. G., & Markman, A. B. (2011). There are at least two kinds of probability matching: Evidence from a secondary task. *Cognition*, *118*(2), 274–279. https://doi.org/10/fb74j5

Pascual-Marqui, R. D. (2002). Standardized low-resolution brain electromagnetic tomography (sLORETA): Technical details. *Methods and Findings in Experimental and Clinical Pharmacology*, *24 Suppl D*, 5–12.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442. https://doi.org/10.1163/156856897X00366

Pineda, J. A., Foote, S. L., & Neville, H. J. (1989). Effects of locus coeruleus lesions on auditory, long-latency, event- related potentials in monkey. *The Journal of Neuroscience*, *9*(1), 81–93.

Pleskac, T. J. (2008). Decision making and learning while taking sequential risks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(1), 167–185. https://doi.org/10/dgcss9

Pleskac, T. J., & Wershbale, A. (2014). Making assessments while taking repeated risks: A pattern of multiple response pathways. *Journal of Experimental Psychology: General*, *143*(1), 142–162. https://doi.org/10/gf3xp4

Polezzi, D., Lotto, L., Daum, I., Sartori, G., & Rumiati, R. (2008). Predicting outcomes of decisions in the brain. *Behavioural Brain Research*, *187*(1), 116–122. https://doi.org/10.1016/j.bbr.2007.09.001

Polezzi, D., Sartori, G., Rumiati, R., Vidotto, G., & Daum, I. (2010). Brain correlates of risky decision-making. *NeuroImage*, *49*(2), 1886–1894. https://doi.org/10/dzrkzt

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, *118*(10), 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019

Raichle, M. E. (2015). The Brain's Default Mode Network. *Annual Review of Neuroscience*, *38*(1), 433–447. https://doi.org/10/gdqcqz

Rodríguez-Fornells, A., Kurzbuch, A. R., & Münte, T. F. (2002). Time course of error detection and correction in humans: Neurophysiological evidence. *The Journal of Neuroscience*, *22*(22), 9990–9996. https://doi.org/10/gf3kwc

Schuermann, B., Endrass, T., & Kathmann, N. (2012). Neural correlates of feedback processing in decision-making under risk. *Frontiers in Human Neuroscience*, *6*. https://doi.org/10/gfv567

Schupp, H. T., Cuthbert, B. N., Bradley, M. M., Cacioppo, J. T., Ito, T., & Lang, P. J. (2000). Affective picture processing: The late positive potential is modulated by motivational relevance. *Psychophysiology*, *37*(2), 257–261. https://doi.org/10/cf9hfx

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. https://doi.org/10/f6fr89

Sripada, C. S. (2018). An Exploration/Exploitation Trade-off Between Mind-Wandering and Goal-Directed Thinking. *The Oxford Handbook of Spontaneous Thought*. https://doi.org/10/gf3sjr

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. https://doi.org/10.1037/h0054651

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (Second edition). Cambridge, Massachusetts: The MIT Press.

Sutton, S., Braren, M., Zubin, J., & John, E. R. (1965). Evoked-Potential Correlates of Stimulus Uncertainty. *Science*, *150*(3700), 1187–1188. https://doi.org/10/bgmjh4

Tzovara, A., Murray, M. M., Bourdaud, N., Chavarriaga, R., Millán, J. del R., & De Lucia, M. (2012). The timing of exploratory decision-making revealed by single-trial topographic EEG analyses. *NeuroImage*, *60*(4), 1959–1969. https://doi.org/10/f3w3s6

Usher, M., Cohen, J. D., Servan-Schreiber, D., Rajkowski, J., & Aston-Jones, G. (1999). The Role of Locus Coeruleus in the Regulation of Cognitive Performance. *Science*, *283*(5401), 549–554. https://doi.org/10/bbq9dq

Vazey, E. M., Moorman, D. E., & Aston-Jones, G. (2018). Phasic locus coeruleus activity regulates cortical encoding of salience information. *Proceedings of the National Academy of Sciences*, *115*(40), E9439–E9448. https://doi.org/10.1073/pnas.1803716115

von Borries, A. K. L., Verkes, R. J., Bulten, B. H., Cools, R., & de Bruijn, E. R. A. (2013). Feedback-related negativity codes outcome valence, but not outcome expectancy, during reversal learning. *Cognitive, Affective, & Behavioral Neuroscience*, *13*(4), 737–746. https://doi.org/10/gfv79t

Wallsten, T. S., Pleskac, T. J., & Lejuez, C. W. (2005). Modeling Behavior in a Clinically Diagnostic Sequential Risk-Taking Task. *Psychological Review*, *112*(4), 862–880. https://doi.org/10/dzpwtg

Warren, C. M., & Holroyd, C. B. (2012). The Impact of Deliberative Strategy Dissociates ERP Components Related to Conflict Processing vs. Reinforcement Learning. *Frontiers in Neuroscience*, *6*. https://doi.org/10.3389/fnins.2012.00043

Warren, C. M., Tanaka, J. W., & Holroyd, C. B. (2011). What can topology changes in the oddball N2 reveal about underlying processes? *NeuroReport*, *22*(17), 870. https://doi.org/10.1097/WNR.0b013e32834bbe1f

Warren, C. M., Wilson, R. C., van der Wee, N. J., Giltay, E. J., van Noorden, M. S., Cohen, J. D., & Nieuwenhuis, S. (2017). The effect of atomoxetine on random and directed exploration in humans. *PLOS ONE*, *12*(4), e0176034. https://doi.org/10/f94pq2

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074–2081. https://doi.org/10/f6tr8t

Worthy, D. A., Hawthorne, M. J., & Otto, A. R. (2013). Heterogeneity of strategy use in the Iowa gambling task: A comparison of win-stay/lose-shift and reinforcement learning models. *Psychonomic Bulletin & Review*, *20*(2), 364–371. https://doi.org/10/f4qfp6

Worthy, D. A., & Maddox, W. T. (2012). Age-Based Differences in Strategy Use in Choice Tasks. *Frontiers in Neuroscience*, *5*. https://doi.org/10/fztcdf

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*(12), 915–924. https://doi.org/10/gfjwtq

Wu, Y., & Zhou, X. (2009). The P300 and reward valence, magnitude, and expectancy in outcome evaluation. *Brain Research*, *1286*, 114–122. https://doi.org/10/fxczvj

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The Neural Basis of Error Detection: Conflict Monitoring and the Error-Related Negativity. *Psychological Review*, *111*(4), 931–959. https://doi.org/10/cm7fhs

Yeung, N., Cohen, J. D., & Botvinick, M. M. (2011). Errors of interpretation and modeling: A reply to Grinband et al. *NeuroImage*, *57*(2), 316–319. https://doi.org/10/ck9sbm

Yeung, N., Holroyd, C. B., & Cohen, J. D. (2005). ERP Correlates of Feedback and Reward Processing in the Presence and Absence of Response Choice. *Cerebral Cortex*, *15*(5), 535–544. https://doi.org/10.1093/cercor/bhh153

Yeung, N., & Sanfey, A. G. (2004). Independent Coding of Reward Magnitude and Valence in the Human Brain. *Journal of Neuroscience*, *24*(28), 6258–6264. https://doi.org/10/dbn7qt

Yu, A. J., & Dayan, P. (2005). Uncertainty, Neuromodulation, and Attention. *Neuron*, *46*(4), 681–692. https://doi.org/10.1016/j.neuron.2005.04.026

Zheng, Y., Li, Q., Wang, K., Wu, H., & Liu, X. (2015). Contextual valence modulates the neural dynamics of risk processing. *Psychophysiology*, *52*(7), 895–904. https://doi.org/10/f7gf3k

Zheng, Y., & Liu, X. (2015). Blunted neural responses to monetary risk in high sensation seekers. *Neuropsychologia*, *71*, 173–180. https://doi.org/10/f7dhp9

Appendix A

SLOT MACHINES

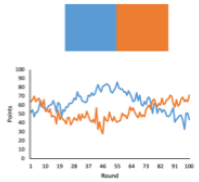
Your goal in this experiment is to win as many points as possible by playing several slot machines.
Each slot machine pays out a point amount ranging from 1 to 100.
Although payouts are somewhat random, the average payout differs for each slot machine.
Over time, the average payouts slowly change
In other words, the best choice may change over time

TASK DETAILS

The number of slot machines will vary (four, nine, or sixteen).
Use the mouse to pick a slot machine.
For each set of slot machines you will play 300 rounds

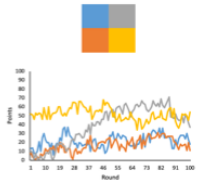
EXAMPLE ONE

Here is an example of two slot machines, represented by coloured squares.
Initially, the orange slot machine is better because it has a higher average payout (see graph).
Over time though, the blue slot machine becomes the better choice (then orange again at the end)



EXAMPLE TWO

Here is an example of four slot machines.
The yellow slot machine is the best choice in beginning, but it is overtaken by the grey slot machine.
Note that these are just examples and that during the actual experiment the colours and payouts will be randomly generated



EEG QUALITY

Please try to minimize eye and head movements. After choosing a slot machine, your choice will be highlighted in white. A point amount will then be briefly displayed within your chosen slot machine. Please remain fixated on the points until they disappear. If you are too slow to respond, the message "too slow" will be displayed and we will not count that round. You will be given several rest breaks. Please use these opportunities to rest your eyes, as needed

SUMMARY

Use the mouse to pick a slot machine. Points (1-100) will appear within your choice - wait for the points to disappear before looking away.
Each slot machine's average payout slowly changes across 300 rounds.

Figure A1. Instructions to participants in Experiment 3.

Appendix B

In this experiment you will use a map to dig for gold. Use the mouse to choose a place on the map at which to dig. You will then be shown the amount of gold you found at that dig site (ranging from 1-100 units). In each round you will have 20 chances to dig at one map location, after which you will be moved to a new map location.

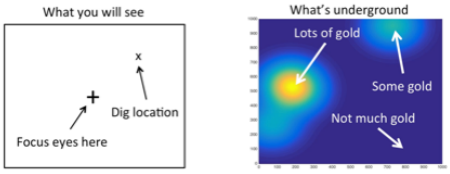
Throughout the experiment you will see a '+' in the middle of the display. Try to focus on the center of the '+' at all times (even while you are deciding where to dig). This will help minimize eye movements.

1

You will be digging for gold in two different states: Alaska and California. Before beginning each round, you will be shown a picture indicating in which state you are digging.

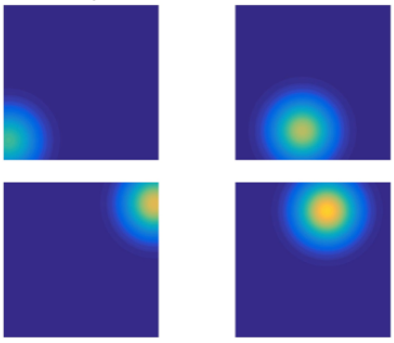
2

Although each map will appear blank to you, some areas of the map will have more gold, and some will have less. There may be "pockets" of gold concentrated in some areas of the map. Your goal is to find as much gold as possible.



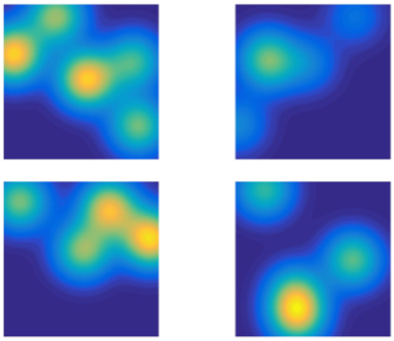
3

In Alaska, gold is usually found in only one area of the map. For example:



4

In California, gold may be found all over the map. For example:



5

There is always a "best" location to dig on each map. **The amount of gold at the best location may be different for each map.**

Each unit of gold you find is worth $0.01.

At the end of the experiment you will be paid for your best round.

You will now play two practice rounds. At the end of each practice round we will show you the underlying map. During the actual experiment, you will not be shown the underlying map.

6

Figure B1. Task instructions for Experiment 4.