D I S S E R T A T I O N

zur Erlangung des akademischen Grades
Doktorin der Naturwissenschaften (Dr. rer. nat.)

Computational and Neural Mechanisms of Human Exploration-Exploitation Behavior

vorgelegt von
Liliana Polanski, M.Sc.

am Fachbereich Erziehungswissenschaft und Psychologie
der Freien Universität Berlin

Berlin, 2024

Erstgutachter: Prof. Dr. Ulman Lindenberger
Zweitgutachterin: Prof. Dr. Gesa Schaadt

Tag der Disputation: 28.11.2024

Für Roman.
Ohne dich hätte ich es nicht geschafft.

To Roman.
I wouldn’t have made it without you.

Acknowledgements

First of all, I would like to thank my supervisor, Douglas Garrett, for giving me an opportunity to develop my own project and experience the kind of research work that made me want to do a PhD in the first place. Thank you, Doug, for all the support and guidance, without which I would have never made it through; for countless hours you spent helping me (learn how to) make sense of things; for believing in me and helping me grow.

I would also like to thank my colleagues from the LNDG, MPI, COMP2PSYCH, and other institutions for their invaluable support and advice, for sharing their expertise, their scripts, and simply taking the time to discuss my work to help make it better.

I would like to thank Michael Krause for his help with working on the cluster and finding bugs in my code. Micha, I am afraid to think how much longer I would have needed for this PhD without your help (much, much longer!).

Big thank you to the IT team, the scanner team, and everyone who was involved in the data collection. Special thank you to Sabrina and Gabi from Telefonstudio – the lab dataset wouldn't have existed without your help.

Special thank you to Tobias Hauser for insisting that I don't use the task from Daw et al. (though it was a challenge at the time, now the ExploreExploit task is the basis of this thesis), to Julian Kosciessa for advising me to log everything during the experiment (Julian, you have no idea how vital this turned out!), and to Lennart Wittkuhn for sharing his expertise and scripts to help me prepare the online experiment.

Thank you to the members of the thesis committee for taking the time to evaluate this thesis.

This work wouldn't have been possible without countless people in and outside of academia, who shared their knowledge, their work materials, and their time with me; who supported me, gave me advice, listened to my ideas and worries, and let me literally and metaphorically cry on their shoulder. To all of you, I am eternally grateful.

Summary

For both living organisms and artificial agents, exploration-exploitation decisions are ubiquitous and vital. They (co-)determine human behavior in all areas of life, from the smallest everyday decisions like grocery shopping to life-changing choices such as choice of a partner or a career. Understanding how exploration-exploitation decisions are made is therefore crucial to understanding – and potentially being able to influence – how and why humans behave the way they do. This dissertation contributes to exploration-exploitation research by examining behavioral, computational, neural, and physiological mechanisms behind exploration-exploitation decision-making. After a brief introduction in which I outline open questions and how this dissertation addresses them (Chapter 1), I proceed to investigate behavioral and computational signatures of exploring and exploiting (Chapter 2). To this end, I use a newly designed task which captures naturally paced exploration-exploitation behavior, while allowing participants to directly indicate whether they explore or exploit in a model-independent way. Having tested participants both in the lab and online, I demonstrate that this task reliably captures key behavioral characteristics of exploration-exploitation behavior and is well suited to its further use in combination with neurophysiological methods. Using computational modeling, I further probe the underlying decision-making processes and their relationship to behavior. The best-fitting computational model highlights different roles that reward and uncertainty estimates play in exploration and exploitation, as well as differences in how quickly acquired information about reward and uncertainty becomes obsolete. I then use the task and computational model presented in Chapter 2 to investigate neural (Chapter 3) and eye tracking (Chapter 4) mechanisms behind exploration-exploitation decision-making. In Chapter 3, I show that uncertainty-driven BOLD signal variability could function as a neural mechanism that allows to adapt exploration-exploitation behavior to a rapidly changing environment. In Chapter 4, I demonstrate that gaze patterns during the decision-making period predict the trial type (exploration or exploitation) and that patterns with different number of dwell locations provide complementary insights into the decision-making behind exploration-exploitation choices. Lastly, Chapter 5 provides a summary of contributions this dissertation makes to the field of exploration-exploitation research, discusses limitations and presents suggestions for future studies. All in all, this dissertation presents a comprehensive investigation of the underlying mechanisms of human exploration-exploitation behavior.

Zusammenfassung

Sowohl für lebende Organismen als auch für Computeralgorithmen sind Exploration-Exploitation Entscheidungen allgegenwertig und von größter Bedeutung. Diese Entscheidungen können menschliches Verhalten in allen Bereichen des Lebens (mit-)bestimmen, von den kleinsten alltäglichen Aufgaben wie zum Beispiel das Einkaufen im Supermarkt bis hin zu den lebenswichtigen Entscheidungen wie die Wahl eines Lebenspartners oder eines Berufes. Daher ist es von größter Wichtigkeit zu untersuchen, wie Exploration-Exploitation Entscheidungen getroffen werden, um zu verstehen, wie und warum Menschen sich auf eine bestimmte Art und Weise verhalten (und um dieses Verhalten potentiell beeinflussen zu können). Diese Doktorarbeit trägt zu der Exploration-Exploitation Forschung bei, indem sie verhaltensrelevante, mathematische, neurologische und physiologische Mechanismen hinter der Exploration-Exploitation Entscheidungsfindung untersucht. Nach einer kurzen Einleitung (Kapitel 1), in der ich die offenen Forschungsfragen darlege und wie diese Dissertation sie behandelt, untersuche ich die verhaltensrelevanten und mathematischen Eigenschaften von Exploration-Exploitation Entscheidungen im Kapitel 2. Dafür benutze ich eine neu entwickelte Aufgabe, die selbstbestimmtes Exploration-Exploitation Verhalten der Probanden erfasst und es ihnen erlaubt direkt anzugeben, ob sie eine Exploration oder Exploitation Entscheidung getroffen haben (und somit eine valide Unterscheidung zwischen Exploration und Exploitation gewährleistet). Nachdem ich Probanden sowohl im Labor als auch online untersucht habe, zeige ich, dass diese Aufgabe die Verhaltenscharakteristiken von Exploration-Exploitation zuverlässig widerspiegelt und sich gut für die Kombination mit neurophysiologischen Methoden eignet. Mittels mathematischer Modellierung untersuche ich dann die zugrundeliegenden Entscheidungsprozesse und ihren Zusammenhang mit dem gezeigten Verhalten. Die Modellierung betont die unterschiedlichen Rollen, die die Schätzung der Belohnung und der Unsicherheit für Exploration-Exploitation Entscheidungen spielen, und die Unterschiede in der Geschwindigkeit, mit welcher gelernte Informationen über Belohnung und Unsicherheit veralteten. Daraufhin benutze ich die gleiche Aufgabe und das mathematische Modell aus dem Kapitel 2, um die neuronale (Kapitel 3) und die Eye-tracking (Kapitel 4) Mechanismen der Exploration-Exploitation Entscheidungsfindung zu untersuchen. Im dritten Kapitel zeige ich, dass die von der Unsicherheit gesteuerte Variabilität des BOLD-Signals einen neuronalen Mechanismus darstellen könnte, der eine Anpassung des Exploration-Exploitation Verhaltens an die sich immer verändernde Umwelt ermöglicht. Im vierten Kapitel zeige ich, dass die Blickmuster, die während der Entscheidungsphase stattfinden, den Trial-typ (also Exploration oder Exploitation) vorhersagen und dass die Anzahl der angeschauten Optionen in den Blickmustern ergänzende Einsichten in den Entscheidungsfindungsprozess gewähren. Schließlich fasst Kapitel 5 die Beiträge dieser Dissertation zu dem Exploration-Exploitation Forschungsgebiet zusammen und weist mögliche Richtungen für die zukünftige Forschung auf. Insgesamt stellt diese Dissertation eine umfassende Untersuchung der zugrundeliegenden Mechanismen des menschlichen Exploration-Exploitation Verhaltens dar.

Table of Contents

1. General Introduction

1.1 The exploration-exploitation dilemma

One day, a little grey mouse, who usually went to the pantry for a yummy and nutritious portion of bread or flour, is attracted by a delicious smell of chocolate from a dark corner of the kitchen. Now she must decide; should she go to the pantry, as she has been doing safely for months, or should she explore the dark corner, which could promise a treat, but could also be a trap? Choosing between a familiar rewarding option (exploitation) and an unfamiliar option that could be more or less rewarding (exploration) is vital for both human and non-human animals (Hills et al., 2015), and has been transferred to the world of artificial agents in the context of reinforcement learning (Sutton and Barto, 1998). Though the solution to the exploration-exploitation dilemma has proven to be elusive, the key to successfully navigating it lies in adaptively switching between exploration and exploitation modes (Sutton and Barto, 1998; Cohen et al., 2007; Bond et al., 2021). Which mode is more adaptive at a given time point depends on multiple factors that influence the costs and gains of exploring vs. exploiting. Structured, stable, well-known environments that consist of options with low reward conflict favor exploitation, whereas volatile environments, high option similarity, and uncertainty about the environment invite exploration (Cohen et al., 2007; Doya, 2008; Mehlhorn et al., 2015; Bond et al., 2021).

Exploration and exploitation modes can be differentiated based on what can be achieved with each of these actions and the costs incurred by forgoing the other. While the purpose of exploitative actions is receiving reward, information gathering in order to reduce uncertainty about the environment is often the focus of exploratory decisions (Sutton and Barto, 1998; Wilson et al., 2014; Blanchard and Gershman, 2018). Further, costs of exploration include forgoing a reliably satisfying reward obtained by exploiting a currently preferred option and the risk of choosing a less rewarding option or even incurring a penalty. Conversely, the costs of exploitation are comprised of not learning about potentially better alternatives and – in non-stationary environments – the possibility that previously learned information has become obsolete (Cohen et al., 2007; Doya, 2008).

1.2 Open questions in exploration-exploitation research and how this dissertation addresses them

Increasing research interest in the exploration-exploitation trade-off over the past two decades has yielded a better understanding of the phenomenon. Kalman filter (Daw et al., 2006), UCB and Thompson sampling (Gershman, 2018), Markov Decision Process (MDP) (Averbeck, 2015; Schwartenbeck et al., 2019) have been used to computationally model exploration-exploitation decision-making. Different types of exploration have been described: random and directed exploration based on different types of

uncertainty (Wilson et al., 2014), undirected exploration based on value differences (Fan et al., 2023), and heuristic-driven exploration (Dubois et al., 2021; Dubois and Hauser, 2022). Research showed how exploration is influenced by reward volatility (Speekenbrink and Konstantinidis, 2015; Piray and Daw, 2021), by the interaction of uncertainty and time pressure (Wu et al., 2019), and how exploration-exploitation behavior may be guided by generalization of learned information (Schulz et al., 2020). However, the neural mechanisms of exploration-exploitation behavior remain poorly understood, and I will make the case that poor task design is at the heart of this lack of understanding.

1.2.1 Optimizing assessment via the ExploreExploit task

Task design in exploration-exploitation research is not trivial, especially for studies focusing on neural correlates. **Table 1-1** provides an overview of all known task fMRI studies on exploration-exploitation behavior in healthy participants. One immediately notices a variety of employed paradigms (though most are a variant of a multi-armed bandit task) and the different ways in which exploration and exploitation were operationalized. Exploitation has been most often defined as either choosing the highest-paying bandit according to a computational model (e.g. Daw et al., 2006; Cockburn et al., 2022) or according to the reward structure (e.g. Muller et al., 2019). Another way to define exploitation trials was by selecting trials on which the same option was chosen several times in a row (e.g. Boorman et al., 2009; Muller et al., 2019). On the other hand, exploration has been operationalized as not choosing the highest-paying option (Daw et al., 2006), switching to another alternative (Boorman et al., 2009), choosing a novel option (Hogeveen et al., 2022), or relating choice to uncertainty estimates (Tomov et al., 2020).

The variety of construct definitions is closely related to the need to categorize trials into exploration vs. exploitation in a valid manner. Most studies employed computational modeling to estimate the expected value (and often uncertainty) for the options and used these estimates alone (e.g. Daw et al., 2006) or in combination with some option characteristics (e.g. novelty (Hogeveen et al., 2022)) or response patterns (e.g. choosing the same bandit multiple times in a row (Boorman et al., 2009)). These methods, however, provide only an approximation of participants' intentions. In contrast to these studies, Blanchard and Gershman (2018) capitalized on the distinction of reward as the focus of exploitative decisions vs. gaining information as the primary goal of explorative choices. They separated feedback (only information on exploration trials vs. only reward on exploitation trials) and response buttons for exploration and exploitation responses (Tversky and Edwards, 1966; Navarro et al., 2016), thus allowing participants to directly indicate which type of response they gave on each trial.

In this dissertation, I pair the strategy of letting participants directly indicate the trial type (as used by Blanchard and Gershman (2018)) with a multi-armed bandit task with nonstationary reward structure (as used by e.g. Daw et al. (2006)) to create the ExploreExploit task. Such non-stationary reward structure possesses great flexibility allowing one to manipulate various aspects of the individual bandits' rewards (e.g. magnitude, volatility) and the relationships between them (discriminability). Moreover, the number of bandits can be easily changed (increased or decreased) to meet the demands of the research

Table 1-1. Summary of task fMRI studies in exploration-exploitation domain.

First author (Year)	N sbj	N trials	Task	Trial categorization method	Operationalization of exploration & exploitation	Contrast: Brain regions
Addicott et al. (2014)	22	200	6-armed bandit with non-stationary reward structure	Model (Kalman filter + softmax)	Exploitation: choosing the option with the highest expected value predicted by the model, Exploration: choosing the option not with the highest expected value	(1) explore – exploit: bilateral superior parietal lobule, intraparietal sulcus, precuneus, supramarginal gyrus, lateral occipital cortex, bilateral superior, middle, and precentral frontal gyrus, left superior, middle, and precentral frontal gyrus, bilateral middle frontal gyrus, frontal pole, paracingulate gyrus, cerebellum, pallidum, thalamus, putamen, caudate, insula (2) exploit – explore: superior, middle temporal gyrus, planum temporale, angular gyrus
Badre et al. (2012)	15	400	Clock task: find best RT in a time interval based on reinforcement (points): reward contingencies varied probability/magnitude trade-offs over time: (1) IEV - increasing EV, (2) DEV - decreasing EV, (3) CEV - constant EV when probability decreases & magnitude increases, (4) CEVR - constant EV when probability increases & magnitude decreases	RT & model (temporal difference learning: tracking PE depending on RT difference)	Exploitation: (1) RTs incrementally adjust to the direction of the highest perceived value throughout the block (i.e. increase in IEV/decrease in DEV) (2) difference in means of the estimated belief distributions: slow minus fast RTs (Directed) exploration: (1) RT swings: deviations from incremental adjusting to EV in the direction of a more uncertain option, i.e. the one with larger SD of the estimated belief distribution (2) relative uncertainty = difference in SDs of the estimated belief distributions: slow minus fast RTs	(1) relative uncertainty: all positive: R rostrolateral PFC, R dlPFC, R superior parietal lobule (SPL), R intraparietal sulcus (IPS), bilateral occipital cortex, R operculum, bilateral cerebellum

Blanchard & Gershman (2018) *	18	200-250	Observe or bet: 2-armed bandit with 80% red or blue bias, 5% probability of change	Different response buttons for observe (explore) and bet blue or bet red (exploit)	Exploitation: receive reward, but no information Exploration: receive information, but no reward	(1) explore – exploit: ROI: bilateral insula, dorsal ACC; full-brain: bilateral thalamus
Boorman et al. (2009)	18	120	2-armed bandit: random reward magnitude 1-100 points, probabilities for each option changed 0-1 in a random walk	Choosing same vs different option, model (Bayesian reinforcement-learning algorithm with Markovian-fashion "predictor" (+ forgetting: probabilities move towards 0.5) and "selector" with sigmoidal probability distribution)	Exploitation: (1) stay trials: choose same option, (2) Relative chosen value provides evidence in favor of the current decision (chosen – unchosen subjective action value) Exploration: (1) switch trials: choose different option (2) Relative unchosen probability drives switching to the alternative action (log unchosen – log chosen action probability)	(1) relative unchosen probability: FPC, dlPFC, mid-IPS (2) relative chosen value: vmPFC
Chakroun et al. (2020)	31	300	4-armed bandit with non-stationary reward structure	Model (Kalman filter + softmax + directed exploration + perseveration)	Exploitation: choosing the option with the highest expected value predicted by the model, Directed exploration: out of the options not with the highest expected value, choosing the option with the highest exploration bonus Random exploration: choosing one of the remaining 2 options	(1) explore – exploit: middle frontal gyrus (FPC), intraparietal sulcus, precuneus, precentral gyrus, postcentral gyrus, supplementary motor cortex, dorsal ACC, cerebellum, thalamus, calcarine cortex, anterior insula, pallidum, vermis, supramarginal gyrus, anterior orbital gyrus, posterior cingulate cortex, caudate, lingual gyrus (2) exploit – explore: angular gyrus, posterior cingulate cortex, precuneus, postcentral gyrus, cerebellum, rostral ACC, superior temporal gyrus, lateral orbital gyrus, central operculum, middle temporal gyrus, superior & inferior frontal gyrus, medial frontal cortex (vmPFC), hippocampus is in the figure but not in the table

Cockburn et al. (2022)	32	~ 400	2-armed bandit (5 options available in each block, but only 2 could be played on each trial): 3 visually familiar bandits (from previous block), 2 novel bandits; 2 bandits available on each trial counterbalanced value (probability of reward), uncertainty (which bandit was sampled less) & novelty (familiar vs novel)	Model (fmUCB: UCB with familiarity-modulated uncertainty bonus + forgetting)	Exploitation: choosing the option with higher expected value according to the model, Exploration: (1) novelty (choosing option with fewer previous exposures) (2) uncertainty (choosing option that was sampled fewer times in the current block)	(1) expected value: selected option (positive): vmPFC, bilateral accumbens (2) expected value: selected vs rejected (sel < rej): FPC, paracingulate, bilateral insula (3) expected value: selected & rejected (positive): vmPFC (4) uncertainty bias: selected option (positive): subcallosal cortex, vmPFC, middle temporal gyrus (5) uncertainty bias: selected & rejected: positive: medial PFC; negative: bilateral lateral occipital cortex, superior parietal lobule
Daw et al. (2006)	14	300	4-armed bandit with non-stationary reward structure	Model (Kalman filter + softmax)	Exploitation: choosing the option with the highest expected value predicted by the model, Exploration: choosing the option not with the highest expected value	(1) explore – exploit: bilateral frontopolar cortex, bilateral anterior intraparietal sulcus (bordering on postcentral gyrus) (2) exploit – explore: n.s.
Dombrovski et al. (2020) *	70	400	Clock task: find best RT in a time interval based on reinforcement (points): reward contingencies varied probability/magnitude trade-offs over time	RT & model (StrategiC Exploration/Exploitation of Temporal Instrumental Contingencies; SCEPTIC)	Exploitation: convergence on global maximum (choosing RT with highest estimated value) Exploration: larger RT swings (RT autocorrelation)	(1) regression model: participants with stronger anterior hippocampus activity related to the global value maximum chose RTs near it more often (more likely to exploit) (2) regression model: participants with stronger PE-related activity in posterior hippocampus had weaker RT autocorrelation (more likely to explore)
Hogeveen et al. (2022) *	37	224	Novelty-bandit task: 3-armed bandit with low (p=0.2), medium (p=0.5), or high (p=0.8) reward probability, novel choice option introduced every 5-12 (M=6) trials	Choosing novel option & model (partially observable Markov decision process; POMDP)	Exploration: (1) choose novel option in 2 & 6 trials following insertion, (2) choices with high FEV/positive BONUS (more uncertain) Exploitation: (1) choose best alternative in 2 & 6 trials following insertion,	(1) BONUS/exploration: positive encoding: dlPFC, vlPFC, vmFPC, OFC, rostral ACC, posterior parietal regions (LIP), inferior temporal regions, visual regions (fusiform face complex), caudate, putamen, nACC, amygdala; negative encoding: lateral frontopolar cortex (FPC), posterior cingulate cortex (2) IEV/exploitation:

					(2) choices with highest IEV & low FEV/negative BONUS	positive encoding: bilateral vmPFC, posterior cingulate cortex, bilateral somatomotor regions, anterior temporal regions, visual cortex, nACC, amygdala, negative encoding: dlPFC, dorsomedial PFC, ACC, LIP, anterior insula
Laureiro-Martinez et al. (2014) *	G1:24, G2:26	300	4-armed bandit with non-stationary reward structure	Model (Kalman filter + softmax)	Exploitation: choosing the option with the highest expected value predicted by the model, Exploration: choosing the option not with the highest expected value	(1) explore – exploit: precuneus, inferior & superior parietal lobule, supramarginal gyrus, superior & middle frontal gyrus, SMA, middle cingulate cortex, frontopolar cortex, IFG (p. Triangularis), insula, LC (2) exploit – explore: superior frontal gyrus, anterior & posterior cingulate cortex, midorbital gyrus, middle temporal gyrus, IFG (p. Triangularis & p. Orbitalis), hippocampus, anterior insula/vmPFC, paracentral lobule, postcentral gyrus
Muller et al. (2019)	19	800	4-armed bandit: 1 option 70-90%, 3 options 20% reward probability, high reward moved (M=20, SD=5 trials)	Switch heuristic (a streak of choosing highest-paying option)	Exploitation: choosing high-reward option in a sequence of consecutive trials Exploration: choose low-reward options	(1) exploit – explore: more active in exploit: medial OFC, hippocampus; more active in explore: fronto-parietal action network, dorsal ACC, preSMA
Tardiff et al. (2021) *,**	34	320	2-armed bandit 1 option always pays 10 points less than the other On each trial: P(flip) = 0.05 (low volatility), P(flip) = 0.2 (high volatility)	Choosing lower-paying option	Exploration: choosing lower-paying option based on previously observed outcome (e.g., choosing the right option after having observed that the left option was now worth 120 points and prior choice of the left option had yielded 110 points)	(1) average integration in the brain in the "peri-explore" period appears to increase leading up to exploration, peak around the explore choice, and fall thereafter. (2) integration of each brain system with the rest of the brain in the "peri-explore" period was significant in the dorsal attention, default, frontoparietal, and limbic systems

Tomov et al. (2020)	31	320	2-armed bandit: safe option: smaller stable reward; risky option: more variable rewards	Model (UCB + Thompson sampling; each trial has exploitation, directed and random exploration elements)	Exploitation: value difference (V) Directed exploration: relative uncertainty (RU) Random exploration: total uncertainty (TU)	(1) relative uncertainty (RU): all negative: middle and inferior occipital gyrus, cerebellum, precentral gyrus, dorsolateral superior frontal gyrus, SMA, middle and posterior cingulate gyrus, middle frontal gyrus (2) total uncertainty (TU): positive: inferior parietal gyrus, middle occipital gyrus, precentral gyrus, insula, thalamus, middle frontal gyrus; negative: dorsolateral & medial orbital superior frontal gyrus, precuneus, gyrus rectus (3) value difference: n.s.

Notes. N trials (number of trials) includes only trials done in the scanner. N sbj – number of subjects, G – group, UCB – upper confidence bound, EV – expected value, PE – (reward) prediction error, RT – reaction time, SD – standard deviation. Brain region abbreviations: L – left, R – right, ACC – anterior cingulate cortex, dlPFC – dorsolateral prefrontal cortex, IFG – inferior frontal gyrus, IPS – intraparietal sulcus, FPC – frontopolar cortex, LC – locus coeruleus, LIP – lateral intraparietal area, nACC – nucleus accumbens, OFC – orbitofrontal cortex, PFC – prefrontal cortex, SMA – supplementary motor area, vmPFC – ventromedial prefrontal cortex. A reference list of publications included in this table can be found in **Table 1-S1**.

* Studies based on ROI or a combination of ROI and full-brain analyses.

** functional connectivity study.

question or to adapt to a different study population (e.g. children or older adults). This non-stationarity invites participants to explore more than they would in the face of deterministic rewards, thus increasing the number of observed exploration trials (Muller et al., 2019; Chakroun et al., 2020). A high number of repetitions has been shown to be particularly important for neuroimaging analyses (Nee, 2019). Finally, determining the trial type based on a computational model may make further use of model parameters (which were used to produce the reward estimates on which trial categorization was based in the first place) for unpacking behavioral and neural differences between exploration and exploitation problematic. A task design that does not rely on computational modeling for trial categorization allows to fully utilize computational modeling as a way to gain insights into the decision-making process behind exploration-exploitation choices. Overall, the ExploreExploit task combines multiple useful features of previously employed tasks while avoiding their drawbacks, which makes it a versatile and adaptable tool to examine behavioral, computational, neural and physiological mechanisms of exploration-exploitation decision-making.

1.2.2 Can brain signal variability provide a new lens on the exploration-exploitation dilemma?

With some exceptions (Tardiff et al. (2021): functional connectivity analysis), task fMRI studies in exploration-exploitation domain (**Table 1-1**) focused on contrasting BOLD signal to map differences in brain activity in certain brain regions to exploration or exploitation trials, or other elements of the decision-making process (such as value or uncertainty estimates). With more neuroimaging research, it became apparent that the results varied vastly, often encompassing large portions of the brain (**Table 1-1**), including frontal, temporal, sensorimotor, parietal and occipital regions, as well as the cingulate cortex, cerebellum, and multiple subcortical structures, such as nucleus accumbens (ventral striatum), caudate and putamen (dorsal striatum), amygdala and thalamus. In light of a relatively small number of studies, varying methodologies, and how distributed topographical results are in the existing neuroimaging exploration-exploitation literature, more research is needed to achieve a better understanding of the neural mechanisms involved in exploration-exploitation decision-making.

This dissertation adopts a different approach to examining neural mechanisms behind exploration and exploitation. Brain signal variability has been shown to potentially provide a neural mechanism that supports flexible adaptation of behavior to changes in the environment (Armbruster-Genç et al., 2016; Garrett et al., 2020; Waschke et al., 2021), while reflecting changes in various types of uncertainty in the environment (Garrett et al., 2014; Kosciessa et al., 2021; Waschke et al., 2021). Furthermore, brain regions showing most prominent variability effects did not always coincide with the regions showing strongest effects based on the mean signal (Garrett et al., 2013), allowing topographic profiles of variability analyses to provide a new angle on the neural basis of cognitive functions.

This dissertation is the first to analyze brain signal variability in the context of exploration-exploitation research. Specifically, I examine how uncertainty-driven brain signal variability might serve as a neural mechanism that helps to adapt exploration-exploitation behavior to a changing environment. In contrast

to previous studies (Garrett et al., 2014, 2015; Kosciessa et al., 2021) that operationalized parametric uncertainty levels based on task design, I *quantify* uncertainty estimates via computational modeling. Overall, the existing neuroimaging literature showed higher levels of uncertainty to be associated with higher levels of brain signal variability, potentially supporting flexible adaptation to a changing environment (Garrett et al., 2014, 2020; Kosciessa et al., 2021). I capitalize on the ability to estimate *three different types of uncertainty* from my computational model. I investigate which uncertainty type has the strongest influence on brain signal variability in the context of exploration-exploitation behavior by contrasting trials in which uncertainty of each type is modulated. This dissertation further extends our understanding of neural underpinnings of exploration-exploitation behavior by presenting topographic profiles of variability-based effects, which hint to cognitive functions that might be involved in the underlying cognitive processes.

1.2.3 Gaze analysis as a real-time observable measure of exploration-exploitation decision-making

Existing exploration-exploitation studies that utilize eye tracking in the context of value-based decision-making have focused exclusively on pupillometry (Jepma and Nieuwenhuis, 2011; Hayes and Petrov, 2015; Pajkossy et al., 2017; Muller et al., 2019; Bond et al., 2021; Tardiff et al., 2021; Fan et al., 2023). A strong interest in the relationship between pupil diameter and exploration-exploitation behavior was fueled by research showing a connection between arousal produced by the locus-coeruleus-noradrenaline (LC-NA) system (of which pupil diameter can serve as a proxy measure (Joshi et al., 2016)) and maintaining focus vs. switching (Rajkowski et al., 1994; Aston-Jones and Cohen, 2005; Bouret and Sara, 2005; Cohen et al., 2007; Sara and Bouret, 2012). Furthermore, past explore-exploit studies that categorize trials based on a computational model brings the same imprecision to eye tracking analyses as in the case of neuroimaging studies. Contrasting eye tracking signals between model-defined exploration and exploitation trials entails the same imprecisions as in the case of neural data because trial categorization in such cases is inherently dependent on the type of the model (Blanchard and Gershman, 2018).

Beyond pupillometry, my thesis establishes *how we look* (gaze) as a unique signature of the underlying dynamics of exploration-exploitation decision-making when the definition of trial type is directly observed from participant behavior. Eye tracking (gaze tracking) is a powerful physiological measure that can provide real-time insights into the decision-making process (Huddleston et al., 2018; Spering, 2022). While computational modeling reflects the latent level of the decision-making process that leads to a button press signifying a response, gaze analysis (specifically, scan path analysis, which represents spatial information of fixations in temporal order (Jacob and Karn, 2003)) has proven itself a useful observed measure for capturing the real-time development of a decision-making process (Polonio et al., 2015; Byrne et al., 2023). For example, studies utilizing economic games showed that observed gaze patterns predicted participants' choices (Polonio et al., 2015) and differentiated between optimal and sub-optimal choice strategies even before a response was given (Byrne et al., 2023). Moreover, value-based decision-making studies with many-alternative (minimum 6 items) choice sets point to a tendency

of participants to shift their gaze between items in a repeated fashion, as if comparing them (Russo and Rosen, 1975; Thomas et al., 2021), landing further support to the utility of gaze analyses for understanding the evolution of a decision-making process. Gaze analyses, therefore, may provide an observed marker for a real-time readout of how a decision to explore or exploit is made before a response in the form of a button press has been given. These insights go beyond pupillometry analyses, which focus on the link between exploration-exploitation behavior and LC-NA system.

Equipped with a task that allows the categorization of exploration and exploitation trials based on observed behavior rather than computational parameters, this dissertation is the first to investigate how gaze patterns could shed light on the real-time dynamics of exploration-exploitation decision-making. I utilize expected value and uncertainty estimates from the computational model to investigate how these features might drive gaze patterns during the decision-making period (before a response is made) while a decision to explore vs. to exploit is being made. To better understand how gaze behavior may reflect the underlying decision-making process, I investigate what patterns with different numbers of dwell locations reveal about the focus of the decision-making process on corresponding trials (see below).

1.3 Overview of the dissertation

The aim of this dissertation is to investigate behavioral, computational, neural and physiological mechanisms underlying human exploration-exploitation behavior. The data used in this dissertation was collected in two studies of young adults: (1) a lab study during which they performed the ExploreExploit task in the MRI scanner with concurrent eye tracking, and (2) a behavioral replication study, during which young adults performed the ExploreExploit task online. Computational modeling was applied to both behavioral data sets and further used to support fMRI and eye tracking analyses. Methodological details pertaining to the task design, experimental setup and computational modeling procedures, though relevant for the entire dissertation, are described in detail in Chapter 2. The Chapters of this dissertation are written in such a way that they can be read as self-contained manuscripts. Materials contained in this dissertation are currently being prepared for publication.

In Chapter 2, I present newly designed ExploreExploit task – a multi-armed bandit task with non-stationary reward structure, which allows participants to freely indicate whether they are exploring or exploiting on each trial. Analyses of the behavioral data from the lab study and from the online replication study show that this task captures key features of exploration-exploitation behavior and produces stable (replicable) results across experimental contexts. Among others, the ExploreExploit task allows one to examine such behavioral features as performance markers (optimal choice percentage, switch percentage), how much participants engage in each action throughout the task (exploration/exploitation percentage), how long they spent continuously in each mode (continuous exploration/exploitation sequences), how reward (exploring and exploiting a bandit with highest-middle-lowest reward) and uncertainty (number of trials before the same bandit is explored again) drive exploration and exploitation, and how reward observed during exploration influences action taken on the next trial. Next, I present a computational model that best reflects the behavioral data from the ExploreExploit task. This model

sheds further light on the role that different aspects of the decision-making process (maximizing reward, reducing uncertainty, learning and forgetting information) play in exploration-exploitation behavior and how they are related to observed task performance.

Capitalizing on the behavioral and computational results presented in Chapter 2, Chapter 3 takes the investigation of the underlying mechanisms of exploration-exploitation behavior further into the neuroimaging domain. In this chapter, I examine whether uncertainty-driven brain signal variability could provide a neural mechanism for adapting exploration-exploitation behavior to a changing environment. To this end, BOLD signal variability in the fMRI data was analyzed using multivariate partial least squares (PLS) analyses. The results show that BOLD signal variability indeed changed in the direction of uncertainty change and was most strongly related to posterior estimation uncertainty (uncertainty that results from a change in knowledge about the options' values after a choice has been made). The results also indicate that changes in BOLD signal variability were associated with task performance and that higher levels of BOLD signal variability might be beneficial for more flexible behavior. Topographically, these effects spanned a broad network of brain regions, including those involved in supporting behavioral flexibility and uncertainty processing. The results highlight the importance of these cognitive functions for adapting exploration-exploitation behavior to a dynamic, changing environment.

In Chapter 4, I analyze eye tracking data to show how gaze patterns could provide an observed marker of the real-time decision-making process that leads to a subsequent button press indicating exploration or exploitation response. Using a series of logistic regression analyses, I examine whether gaze patterns during the decision-making period predict the trial type (exploration or exploitation). For analyses, fixations to the same bandit were grouped into "dwell locations" and gaze patterns included patterns with one, two or three dwell locations (bandits) during the decision-making period. Not only did gaze behavior reflect the computational model in how expected value and uncertainty drive exploration and exploitation responses, but gaze patterns indeed predicted the trial type. Furthermore, gaze patterns with different numbers of dwell locations (fixated bandits) might reflect different characteristics of – and thus deliver complimentary insights into – the decision-making process behind exploration-exploitation behavior. Fixating on just one option likely indicated a strong focus on that specific option, especially if it was the bandit with the highest expected value (in which case the probability of exploitation was nearly 80%). In patterns with two fixated bandits, the option fixated last was most likely an indicator of an upcoming exploration response, if it had the highest uncertainty, and exploitation response, if it had the highest expected value. Trials with three dwell locations revealed a complex relationship between *how* (type of the pattern) and *where* (expected value and uncertainty of the options) participants looked in determining the trial type. The results further show an association between gaze patterns and task performance: participants who performed worse, used typical ("correct") gaze strategy, but applied them to ("incorrect") options with lower expected values.

Chapter 5 summarizes contributions of this dissertation to the exploration-exploitation research field and outlines limitations and possible directions for future research.

1.4 References

Addicott, M. A., Pearson, J. M., Froeliger, B., Platt, M. L., & McClernon, F. J. (2014). Smoking automaticity and tolerance moderate brain activation during explore–exploit behavior. *Psychiatry Research: Neuroimaging*, *224*(3), 254–261. https://doi.org/10.1016/j.pscychresns.2014.10.014

Armbruster-Genç, D. J. N., Ueltzhöffer, K., & Fiebach, C. J. (2016). Brain Signal Variability Differentially Affects Cognitive Flexibility and Cognitive Stability. *The Journal of Neuroscience*, *36*(14), 3978–3987. https://doi.org/10.1523/jneurosci.2517-14.2016

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

Averbeck, B. B. (2015). Theory of Choice in Bandit, Information Sampling and Foraging Tasks. *PLOS Computational Biology*, *11*(3), e1004164. https://doi.org/10.1371/journal.pcbi.1004164

Badre, D., Doll, B. B., Long, N. M., & Frank, M. J. (2012). Rostrolateral Prefrontal Cortex and Individual Differences in Uncertainty-Driven Exploration. *Neuron*, *73*(3), 595–607. https://doi.org/10.1016/j.neuron.2011.12.025

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/10.3758/s13415-017-0556-2

Bond, K., Dunovan, K., Porter, A., Rubin, J. E., & Verstynen, T. (2021). Dynamic decision policy reconfiguration under outcome uncertainty. *ELife*, *10*. https://doi.org/10.7554/elife.65540

Boorman, E. D., Behrens, T. E. J., Woolrich, M. W., & Rushworth, M. F. S. (2009). How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action. *Neuron*, *62*(5), 733–743. https://doi.org/10.1016/j.neuron.2009.05.014

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

Byrne, S. A., Reynolds, A. P. F., Biliotti, C., Bargagli-Stoffi, F. J., Polonio, L., & Riccaboni, M. (2023). Predicting choice behaviour in economic games using gaze data encoded as scanpath images. *Scientific Reports*, *13*(1), 4722. https://doi.org/10.1038/s41598-023-31536-5

Chakroun, K., Mathar, D., Wiehler, A., Ganzer, F., & Peters, J. (2020). Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making. *ELife*, *9*, e51260. https://doi.org/10.7554/elife.51260

Cockburn, J., Man, V., Cunningham, W. A., & O'Doherty, J. P. (2022). Novelty and uncertainty regulate the balance between exploration and exploitation through distinct mechanisms in the human brain. *Neuron*, *110*(16), 2691-2702.e8. https://doi.org/10.1016/j.neuron.2022.05.025

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/10.1098/rstb.2007.2098

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. https://doi.org/10.1038/nature04766

Dombrovski, A. Y., Luna, B., & Hallquist, M. N. (2020). Differential reinforcement encoding along the hippocampal long axis helps resolve the explore–exploit dilemma. *Nature Communications*, *11*(1), 5407. https://doi.org/10.1038/s41467-020-18864-0

Doya, K. (2008). Modulators of decision making. *Nature Neuroscience*, *11*(4), 410–416. https://doi.org/10.1038/nn2077

Dubois, M., Habicht, J., Michely, J., Moran, R., Dolan, R. J., & Hauser, T. U. (2021). Human complex exploration strategies are enriched by noradrenaline-modulated heuristics. *ELife*, *10*, e59907. https://doi.org/10.7554/elife.59907

Dubois, M., & Hauser, T. U. (2022). Value-free random exploration is linked to impulsivity. *Nature Communications*, *13*(1), 4542. https://doi.org/10.1038/s41467-022-31918-9

Fan, H., Burke, T., Sambrano, D. C., Dial, E., Phelps, E. A., & Gershman, S. J. (2023). Pupil Size Encodes Uncertainty during Exploration. *Journal of Cognitive Neuroscience*, *35*(9), 1508–1520. https://doi.org/10.1162/jocn_a_02025

Garrett, D. D., Epp, S., Kleemeyer, M., Lindenberger, U., & Polk, T. A. (2020). Higher performers upregulate brain signal variability in response to more feature-rich visual input. *NeuroImage*, *217*, 116836. https://doi.org/10.1016/j.neuroimage.2020.116836

Garrett, D. D., McIntosh, A. R., & Grady, C. L. (2014). Brain Signal Variability is Parametrically Modifiable. *Cerebral Cortex*, *24*(11), 2931–2940. https://doi.org/10.1093/cercor/bht150

Garrett, D. D., Nagel, I. E., Preuschhof, C., Burzynska, A. Z., Marchner, J., Wiegert, S., Jungehülsing, G. J., Nyberg, L., Villringer, A., Li, S.-C., Heekeren, H. R., Bäckman, L., & Lindenberger, U. (2015). Amphetamine modulates brain signal variability and working memory in younger and older adults. *Proceedings of the National Academy of Sciences*, *112*(24), 7593–7598. https://doi.org/10.1073/pnas.1504090112

Garrett, D. D., Samanez-Larkin, G. R., MacDonald, S. W. S., Lindenberger, U., McIntosh, A. R., & Grady, C. L. (2013). Moment-to-moment brain signal variability: A next frontier in human brain mapping? *Neuroscience & Biobehavioral Reviews*, *37*(4), 610–624. https://doi.org/10.1016/j.neubiorev.2013.02.015

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*(PLoS Computational Biology 6 2010), 34–42. https://doi.org/10.1016/j.cognition.2017.12.014

Hayes, T. R., & Petrov, A. A. (2015). Pupil Diameter Tracks the Exploration–Exploitation Trade-off during Analogical Reasoning and Explains Individual Differences in Fluid Intelligence. *Journal of Cognitive Neuroscience*, *28*(2), 308–318. https://doi.org/10.1162/jocn_a_00895

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., Couzin, I. D., & Group, the C. S. R. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, *19*(1), 46–54. https://doi.org/10.1016/j.tics.2014.10.004

Hogeveen, J., Mullins, T. S., Romero, J. D., Eversole, E., Rogge-Obando, K., Mayer, A. R., & Costa, V. D. (2022). The neurocomputational bases of explore-exploit decision-making. *Neuron*, *110*(11), 1869-1879.e5. https://doi.org/10.1016/j.neuron.2022.03.014

Huddleston, P. T., Behe, B. K., Driesener, C., & Minahan, S. (2018). Inside-outside: Using eye-tracking to investigate search-choice processes in the retail environment. *Journal of Retailing and Consumer Services*, *43*, 85–93. https://doi.org/10.1016/j.jretconser.2018.03.006

Jacob, R. J. K., & Karn, K. S. (2003). The Mind's Eye. *Section 4: Eye Movements in Human—Computer Interaction*, *Vision Research391999*, 573–605. https://doi.org/10.1016/b978-044451020-4/50031-1

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/10.1162/jocn.2010.21548

Joshi, S., Li, Y., Kalwani, R. M., & Gold, J. I. (2016). Relationships between Pupil Diameter and Neuronal Activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex. *Neuron*, *89*(1), 221–234. https://doi.org/10.1016/j.neuron.2015.11.028

Kosciessa, J. Q., Lindenberger, U., & Garrett, D. D. (2021). Thalamocortical excitability modulation guides human perception under uncertainty. *Nature Communications*, *12*(1), 2430. https://doi.org/10.1038/s41467-021-22511-7

Laureiro-Martínez, D., Canessa, N., Brusoni, S., Zollo, M., Hare, T., Alemanno, F., & Cappa, S. F. (2014). Frontopolar cortex and decision-making efficiency: comparing brain activity of experts with different professional background during an exploration-exploitation task. *Frontiers in Human Neuroscience*, *7*, 927. https://doi.org/10.3389/fnhum.2013.00927

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., Hausmann, D., Fiedler, K., & Gonzalez, C. (2015). Unpacking the Exploration–Exploitation Tradeoff: A Synthesis of Human and Animal Literatures. *Decision*, *2*(3), 191–215. https://doi.org/10.1037/dec0000033

Muller, T. H., Mars, R. B., Behrens, T. E., & O'Reilly, J. X. (2019). Control of entropy in neural models of environmental state. *ELife*, *8*. https://doi.org/10.7554/elife.39404

Navarro, D. J., Newell, B. R., & Schulze, C. (2016). Learning and choosing in an uncertain world: An investigation of the explore–exploit dilemma in static and dynamic environments. *Cognitive Psychology*, *85*, 43–77. https://doi.org/10.1016/j.cogpsych.2016.01.001

Nee, D. E. (2019). fMRI replicability depends upon sufficient individual-level data. *Communications Biology*, *2*(1), 130. https://doi.org/10.1038/s42003-019-0378-6

Pajkossy, P., Szőllősi, Á., Demeter, G., & Racsmány, M. (2017). Tonic noradrenergic activity modulates explorative behavior and attentional set shifting: Evidence from pupillometry and gaze pattern analysis. *Psychophysiology*, *54*(12), 1839–1854. https://doi.org/10.1111/psyp.12964

Piray, P., & Daw, N. D. (2021). A model for learning based on the joint estimation of stochasticity and volatility. *Nature Communications*, *12*(1), 6587. https://doi.org/10.1038/s41467-021-26731-9

Polonio, L., Guida, S. D., & Coricelli, G. (2015). Strategic sophistication and attention in games: An eye-tracking study. *Games and Economic Behavior*, *94*, 80–96. https://doi.org/10.1016/j.geb.2015.09.003

Rajkowski, J., Kubiak, P., & Aston-Jones, G. (1994). Locus coeruleus activity in monkey: Phasic and tonic changes are associated with altered vigilance. *Brain Research Bulletin*, *35*(5–6), 607–616. https://doi.org/10.1016/0361-9230(94)90175-9

Russo, J. E., & Rosen, L. D. (1975). An eye fixation analysis of multialternative choice. *Memory & Cognition*, *3*(3), 267–276. https://doi.org/10.3758/bf03212910

Sara, S. J., & Bouret, S. (2012). Orienting and Reorienting: The Locus Coeruleus Mediates Cognition through Arousal. *Neuron*, *76*(1), 130–141. https://doi.org/10.1016/j.neuron.2012.09.011

Schulz, E., Franklin, N. T., & Gershman, S. J. (2020). Finding structure in multi-armed bandits. *Cognitive Psychology*, *119*, 101261. https://doi.org/10.1016/j.cogpsych.2019.101261

Schwartenbeck, P., Passecker, J., Hauser, T. U., FitzGerald, T. H., Kronbichler, M., & Friston, K. J. (2019). Computational mechanisms of curiosity and goal-directed exploration. *ELife*, *8*, e41703. https://doi.org/10.7554/elife.41703

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and Exploration in a Restless Bandit Problem. *Topics in Cognitive Science*, *7*(2), 351–367. https://doi.org/10.1111/tops.12145

Spering, M. (2022). Eye Movements as a Window into Decision-Making. *Annual Review of Vision Science*, *8*(1), 427–448. https://doi.org/10.1146/annurev-vision-100720-125029

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning* (Vol. 135). MIT press.

Tardiff, N., Medaglia, J. D., Bassett, D. S., & Thompson-Schill, S. L. (2021). The modulation of brain network integration and arousal during exploration. *NeuroImage*, *240*, 118369. https://doi.org/10.1016/j.neuroimage.2021.118369

Thomas, A. W., Molter, F., & Krajbich, I. (2021). Uncovering the computational mechanisms underlying many-alternative choice. *ELife*, *10*, e57012. https://doi.org/10.7554/elife.57012

Tomov, M. S., Truong, V. Q., Hundia, R. A., & Gershman, S. J. (2020). Dissociable neural correlates of uncertainty underlie different exploration strategies. *Nature Communications*, *11*(1), 2371. https://doi.org/10.1038/s41467-020-15766-z

Tversky, A., & Edwards, W. (1966). Information versus reward in binary choices. *Journal of Experimental Psychology*, *71*(5), 680–683. https://doi.org/10.1037/h0023123

Waschke, L., Kloosterman, N. A., Obleser, J., & Garrett, D. D. (2021). Behavior needs neural variability. *Neuron*. https://doi.org/10.1016/j.neuron.2021.01.023

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. https://doi.org/10.1037/a0038199

Wu, C. M., Schulz, E., Gerbaulet, K., Pleskac, T. J., & Speekenbrink, M. (2019). Under pressure: The influence of time limits on human exploration. *Psyarxiv*.

2. Disentangling exploration and exploitation: Behavioral and computational mechanisms

Abstract

In the field of exploration-exploitation decision-making, the accurate determination of which type of decision is made by participants is often non-trivial, yet this determination remains a prerequisite to obtaining precise results when contrasting behavior or neurophysiological signals. In the current study, we present a newly developed ExploreExploit task, which combines multiple characteristics of previously used paradigms to allow participants to directly indicate whether they explore or exploit. The task captures naturally-paced exploration-exploitation behavior, encourages high exploration rates throughout the task horizon, and allows for a variety of modifications to the reward structure. Having tested 47 young adults in the lab and 52 young adults online, we show that this task successfully captures behavioral characteristics of exploration-exploitation behavior. We further present a computational model that highlights the role of reward and uncertainty reduction for each behavior type, reflects how learning and forgetting processes contribute to successful task performance, and that illustrates how decision-making processes behind exploration-exploitation decision-making relate to behavior. We demonstrate that the ExploreExploit task shows stable behavioral and computational modeling results in the lab and in online settings. Taken together, these features make our task a novel addition to the collection of exploration-exploitation paradigms and make it particularly suitable to be paired with neuroimaging and physiological methods for further investigation of exploration-exploitation behavior.

Keywords: exploration, exploitation, multi-armed bandit, task design, reinforcement learning

2.1 Introduction

A choice between a familiar rewarding option (exploitation) and an alternative option of unknown reward (exploration) is omnipresent in human lives and determines a wide range of decisions, from everyday food choices to the choice of career or life partner. Though multiple definitions of exploitation and exploration have been used in the literature (e.g., staying with an option or choosing a different one; selecting the option with the highest reward or not; (see Mehlhorn et al., 2015 for a review)), a common perspective is that agents are focused on receiving reward during exploitation, while exploration is driven by information gathering and uncertainty reduction (Cohen et al., 2007; Mehlhorn et al., 2015). Though the question of how humans master the exploration-exploitation trade-off has received substantial attention since it was first espoused (Cohen et al., 2007; Daw et al., 2006; Sutton & Barto, 1998), describing and contrasting exploration and exploitation has remained experimentally challenging for a number of important reasons.

Though a multi-armed bandit with non-stationary reward structure (Daw et al., 2006; Jepma & Nieuwenhuis, 2011; Speekenbrink & Konstantinidis, 2015; Tversky & Edwards, 1966) became a popular paradigm in exploration-exploitation research due to its ability to elicit naturally paced behavior and keep the motivation for exploration high throughout the task horizon, this task requires a computational model to allow experimenters to determine whether a participant has explored or exploited. This makes interpreting experimental results contingent on the choice of the models and their assumptions about exploration strategies, (as discussed in detail in Blanchard & Gershman, 2018). The authors note that without a directly observable behavioral marker, it is hard to reliably categorize exploration and exploitation trials and separate exploratory choices from random error. This lowers the validity of contrast analyses between exploration and exploitation, including in the neurophysiological domains (Blanchard & Gershman, 2018). Conversely, other paradigms use novelty (Averbeck, 2015; Cockburn et al., 2022; Hogeveen et al., 2022) or a combination of reward magnitude and obtained information (Dubois et al., 2021; Wilson et al., 2014) to anchor the definition of exploration trials in behavior. However, these paradigms aim to elicit exploration on specific trials (e.g. by introducing a novel stimulus (Averbeck, 2015) or by manipulating how much information is available about an option (Wilson et al., 2014)) at the cost of not observing naturally paced exploration-exploitation behavior. In addition, only these trials may be analyzed (Wilson et al., 2014), which greatly decreases the trial counts (especially in the exploration condition) and makes such designs difficult to use with neuroimaging methods. Still other paradigms assign different response buttons to exploration and exploitation responses and provide only information as feedback on exploration trials and only reward as feedback on exploitation trials (Blanchard & Gershman, 2018; Navarro et al., 2016; Tversky & Edwards, 1966), thus creating a direct measure of whether participants explore or exploit. The use of a deterministic reward structure (Navarro et al., 2016; Tversky & Edwards, 1966), however, does not encourage exploration throughout the course of the block, and deploying fixed-magnitude probabilistic rewards (Blanchard & Gershman, 2018; Tversky & Edwards, 1966) provides limited possibilities for testing the influence of effects related to the reward magnitude and for scaling the number of bandits.

Here, we present a newly developed ExploreExploit task, a 3-armed bandit task which combines useful features of previously used paradigms to overcome a variety of experimental limitations. With the ExploreExploit task, our design capitalizes on the distinction between exploration and exploitation based on the goal of the respective action: gaining information to reduce uncertainty vs. gaining reward. In our task, participants receive only information as feedback after exploratory responses and only reward after exploitative choices, while also use separate response buttons for indicating their wish to explore or exploit a bandit on a given trial (Blanchard & Gershman, 2018; Navarro et al., 2016; Tversky & Edwards, 1966). A behavioral marker of the trial type coded directly into response buttons allows us to unambiguously assign each trial to either exploration or exploitation in a way that does not require computational modeling (Daw et al., 2006) or evaluating how the chosen option differs from other options in novelty (Hogeveen et al., 2022), reward, or uncertainty (Gershman, 2018; Wilson et al., 2014). Hence, participants themselves indicate whether they have explored or exploited. This allows easy assignment of behavior and neural data to exploration and exploitation, ensuring more precise results when contrasting exploration and exploitation trials. In addition, in designs requiring mode-based trial categorization, the use of model parameters in further analyses of exploration-exploitation data is potentially problematic because these parameters served to create value estimates, which, in turn, were used for trial categorization. A direct measure of trial categorization in our task allows to avoid this issue and to use computational model parameters to support further analyses gaining deeper insights into exploration-exploitation behavior. Another key feature of the ExploreExploit task is its non-stationary reward structure (Daw et al., 2006; Sutton & Barto, 1998), which encourages a constantly high rate of exploration responses throughout the block, thus increasing the number of exploration trials available for analysis. Importantly, a multi-armed bandit task with a non-stationary reward structure allows the observation of naturally paced dynamic switching between exploration and exploitation. The ExploreExploit task uses “magnitude-based” rewards that differ in how much reward each bandit provides on each trial. This makes the reward structure flexibly modifiable, allowing the investigation of various influences of different reward characteristics (e.g. reward range, reward similarity, wins and losses) on exploration-exploitation behavior.

In the current study, we describe behavioral characteristics of exploration and exploitation using the ExploreExploit task, allowing us to observe dynamic exploration-exploitation behavior while unambiguously categorizing the trial type by directly observable participant responses. We aimed to verify the behavioral features of exploration and exploitation seen in previous studies, which often relied on computational modeling for determining trial type. For instance, we expected exploration to take up ca. 30% of trials (Chakroun et al., 2020; Muller et al., 2019) and to remain at that level throughout the block (Blanchard & Gershman, 2018). We also expected continuous exploitation sequences to be longer than continuous exploration sequences (Blanchard & Gershman, 2018; Muller et al., 2019). If choice behavior reflects the focus on gaining reward during exploitation, the bandit with the highest reward should be exploited on the majority of exploitation trials. At the same time, the distribution of exploration trials across bandits should be much more equal, emphasizing the role of information gathering during exploration. Furthermore, we utilized computational modeling to examine the details of behavioral mechanisms underlying exploration-exploitation decision-making.

Having tested 47 young adults in the lab and having replicated both behavioral and computational results with 52 young adults online, we show that the ExploreExploit task is well suited to characterize exploration-exploitation behavior, to investigate the influence of the reward structure and task environment on these behaviors, and to examine computational and neurophysiological mechanisms behind exploration-exploitation choices in future studies.

2.2 Materials and Methods

2.2.1 Lab study

Participants

52 healthy young adults participated in the study. Five participants were excluded due to incorrectly saved data, not understanding the assignment of the response buttons, or because they were identified during the experiment as not understanding the task. The data of excluded participants was used to inform data-based exclusion criteria, which were applied in the online replication study (such as less than 15% exploration or exploitation trials, or less than 50% exploitation trials on which the highest-paying bandit was chosen; see *Data exclusion criteria* in Supplementary Methods for details). The final sample thus consisted of 47 participants (age 18 – 35 years, M = 23.9, SD = 3.9; 27 female, 20 male). Subjects received 10 euros per hour plus a 10-euro bonus for their participation. To increase motivation, the bonus was described as result-dependent at the beginning of the experiment, but all participants who finished the study received the same amount and were debriefed at the end. Participants were recruited thorough an internal participant database at the Max-Planck-Institute for Human Development. The study was approved by the DGPs ethics committee and written informed consent was obtained from each participant.

Procedure

The lab study – including all instructions and experiment presentation – was conducted in German and took ca. 3 hours in total. Participants performed the ExploreExploit task while in the MRI scanner (we focus on behavioral data in this chapter). The experiment consisted of a practice session (ca. 15 min) outside of the scanner, that allowed participants to familiarize themselves with the task, and a main experiment session (ca. 50 min) inside the scanner. During the practice session, participants completed 2 short practice blocks of 25 trials each. The main task session consisted of 5 blocks of 100 trials each. Both during the practice session and during the main task, participants could take self-paced breaks after each block.

The main task was presented with MATLAB R2017b (https://www.mathworks.com) on a Dell Precision Tower 5810 PC running Windows7. Experimental stimuli were projected onto a screen inside the scanner room and viewed with a mirror placed atop of the coil. Additionally, MRI-compatible headphones

were used to play the sound, which marked the beginning of each trial (see task design description below). The responses were given with an MRI-compatible button box (8-Button Bimanual Fiber Optic Response Pad; Current Designs, https://www.curdes.com/), using the index, middle, and ring fingers of each hand.

Task instructions were presented before the practice session. Participants were informed about the reward range and that rewards of each bandit and each block were independent of each other. They were explicitly instructed to use *both* exploration and exploitation. The goal of the task was described as earning as much reward as possible. At the same time, in line with previous studies (Addicott et al., 2014), participants were told that finding the bandit with the highest reward would help them to achieve this goal. The latter instruction was given to ensure that participants aim to find and exploit the best-paying bandit, since the percentage of optimal choice (percentage of exploitation trials on which the highest-paying bandit was chosen) was used as a measure of task performance (see *Optimal choice as a measure of task performance* in Supplementary Methods).

Task design

The ExploreExploit task (programmed in MATLAB 2018a) was created by combining a multi-armed bandit task with a non-stationary reward structure and separate behavioral responses for exploration and exploitation. The task design and trial structure are schematically depicted in **Figure 2-1A**. The bandits were represented by geometric figures: triangle, circle, and square. On each trial, participants had to choose whether to explore or exploit one of the 3 bandits, yielding 6 response possibilities. Crucially, if they chose to explore, they would see how many points the chosen bandit would have yielded at the current trial, but these points would *not* be added to their account. After exploiting a bandit, participants saw non-informative feedback in the form of the first letter of the word for the figure that represented the bandit (German: D – for ***D****reieck* (triangle), V – for ***V****iereck* (square), and K – for ***K****reis* (circle)), but the points that the chosen bandit provided on that trial were added to the participant's account. Thus, exploring a bandit provided information but no reward, while exploiting a bandit provided reward but no information. This allowed to separate the motivation for choosing exploration vs. exploitation. In combination with separate buttons for exploitative and exploratory responses, it allowed for an unambiguous categorization of trials as exploration or exploitation.

Geometric figures were all black presented on a grey background. They were centered vertically on the screen and the central figure was also centered horizontally. The two figures on the sides were placed horizontally such that they divided the distance between the central figure and the end of the screen in half. Experimental script adapted stimuli presentation to the screen size of the computer on which it was run. To minimize changes in luminance, all figures were on screen at all times except for when feedback was presented, at which point a number or a letter appeared in place of the figure that represented the chosen bandit (see **Figure 2-1**).

A trial began with a decision-making period that lasted from 1000 to 2000 ms, drawn from a uniform distribution. At the end of the decision-making period, a sound cue (500 Hz, 200 ms) indicated that a

response could be given. From the beginning of the sound cue participants had max. 1500 ms to give a response. If a response was given earlier, the response phase terminated and the remaining time was not added to any other part of the trial. A variable-length post-choice delay (300-700 ms, uniform distribution) followed the response. Following this, feedback was presented for 1000 ms. The end of feedback marked the end of the trial and the decision-making period for the next trial started directly after that.

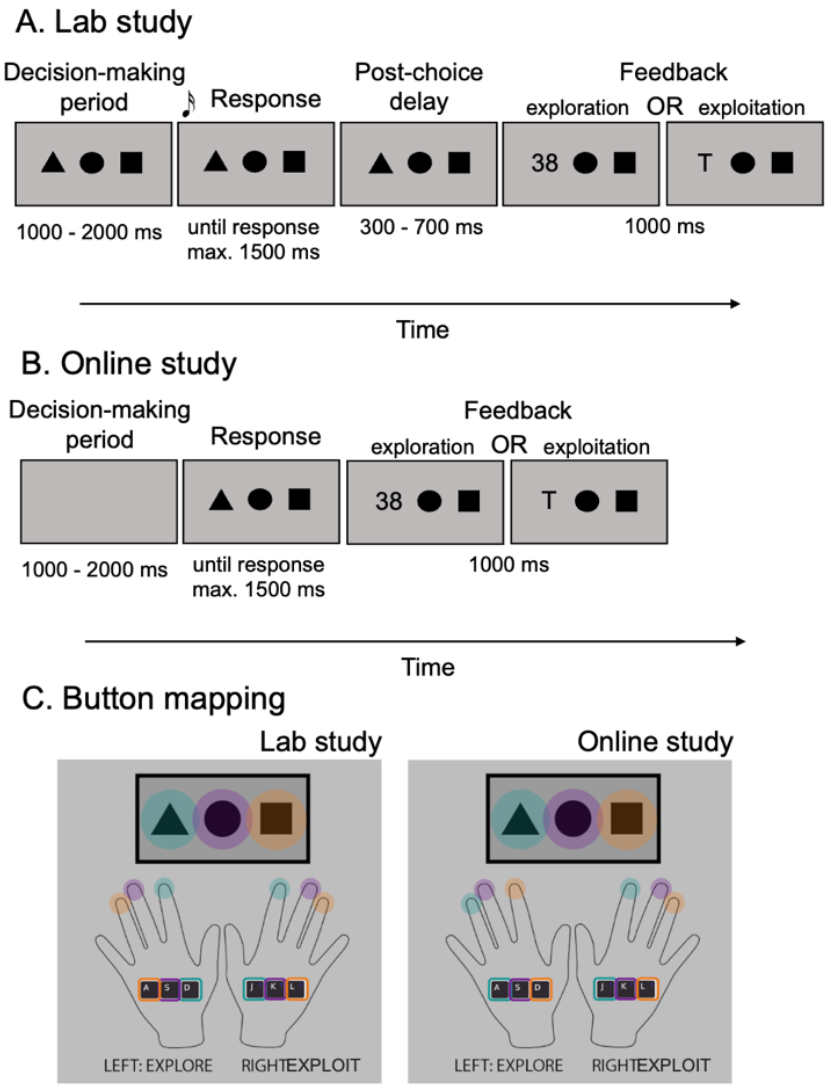


Figure 2-1. Task design. *Trial structure of the ExploreExploit task in the lab study (A) and online study (B). C – Stimulus-response mapping in the lab study (left) and online study (right). For example, in the lab study, triangle corresponds to the index finger and square to the ring finger. Left hand corresponds to exploration and right hand corresponds to exploitation. Hence, to explore the bandit denoted by triangle, one should press a button with the left index finger. To exploit the bandit denoted by square, one would respond with the right ring finger. In the online study, fingers of each hand were mapped to the horizontal position of the figures.*

After each block, participants could take a self-paced break. During this time, they saw information about the total reward they earned in that section of the task (expressed as a percentage of the maximum reward they could earn on all trials in a given block) and how often they made optimal choices

(expressed as a percentage of trials on which the highest-paying bandit was exploited relative to the number of exploitation trials in the block).

A possible response button mapping could be achieved by matching each geometric figure on the screen the same finger on each hand based on their horizontal order (as was done in the online study where a keyboard was used to give responses; discussed below). However, holding the button box in the scanner in the horizontal position that would correspond to the horizontal layout of keys on a keyboard would be difficult and could cause a disruption to the stimulus-response mapping throughout the experiment. For this reason, each of the 3 bandits arranged in a horizontal line on the screen was assigned to the same finger on both hands, starting with the index finger for the left-most bandit (stimulus-response mapping is depicted in **Figure 2-1C**). This allowed participants to hold the button box in any position that they felt comfortable with. Each hand corresponded to either exploration or exploitation. Exploration-exploitation assignment to the hands as well as the position of the bandits on the screen were counterbalanced between participants. A computer keyboard was used to give responses during the practice session with the same figure-to-finger mapping as in the scanner.

Stimuli

The non-stationary reward structure (see Daw et al., 2006 for a detailed description) was created independently for each bandit and each block. Reward points were sampled with SD = 4 around the mean that moved as a random walk with SD = 4 in the range of 10-90 points. For each bandit, rewards were centered at a mean of 50 across all blocks to avoid inducing expectations of any given bandit being more (or less) beneficial than any other. The resulting rewards varied from 1 to 100 points. Three reward structures (each containing 3 bandits and 5 blocks of 100 trials; **Figure 2-S1**) were randomly assigned to participants. Because practice blocks were much shorter (25 trials), separate reward structures were created for practice. As in the main experiment, rewards for each bandit were centered around a mean of 50 throughout all practice trials.

2.2.2 Online replication study

Participants

For the online replication study, we collected data from 54 young adults. Two participants were excluded because their data had less than 15% exploration or exploitation trials, or less than 50% exploitation trials on which the highest-paying bandit was chosen (see *Data exclusion criteria* in Supplementary Methods for details). The resulting sample consisted of 52 participants (age 18 – 35, M = 23.4, SD = 4.1). They were recruited on Prolific (https://www.prolific.co/) using the following criteria: age 18 – 35 years, absence of psychological or neurological disorders, fluent English, and approval rate on Prolific of at least 95%. The experiment could only be done on a PC or laptop. Participants were paid 8 GBP per hour and the experiment took ~1 hour to complete. The study was approved by the ethics committee of the Max-Planck-Institute for Human Development.

Task design and procedure

The ExploreExploit task was adapted for use online and with international participants. In the following, we present information about task and experimental procedure specific to the online study.

The consent form and task instructions (in English) were hosted on GDWG LimeSurvey (https://www.gwdg.de/application-services/online-surveys). Consent was indicated by clicking on a "Yes" button at the end of the consent form. Participants also filled out a short demographic questionnaire and completed a short test to check that they understood the instructions. Only participants who successfully passed the test were redirected to the task that was hosted on Pavlovia (https://pavlovia.org/).

The task design and instructions were similar to the ones used in the lab, but adaptations were made to account for the online nature of the study. The task was programmed in Python 3 and JavaScript, using PsychoPy2 Experiment Builder v2020.1.2 (Peirce et al., 2019) and materials from Wittkuhn et al. (2022). New reward structures were produced for both practice and the main task. Instead of using a sound cue to indicate that participants could start the response, the figures that represented the bandits were not present on the screen during the decision-making period and their appearance was the cue that the response could be given (**Figure 2-1B**). Participants used keyboard keys for a, s, d, and j, k, l with their ring, middle, and index finger of each hand. The stimulus-response mapping assigned the left-most bandit on the screen to the left-most finger of each hand (**Figure 2-1C**).

2.2.3 Statistical analyses

Statistical analyses were performed using R version 4.2.2 (R Core Team, 2022). Main R packages used for data analyses are listed in **Table 2-S1**. Our analyses consist of a series of linear (mixed) models, such as regression and analysis of variance (ANOVA). Whenever necessary, subject ID was used as a random intercept to account for within-subject variance. Paired FDR-corrected (Benjamini & Hochberg, 1995) t-tests were used as follow-ups to ANOVA analyses. We report semi-partial R^2 (Nakagawa & Schielzeth, 2013) for regression and semi-partial η^2 (η^2) (Richardson, 2011) for ANOVA models as measures of effect size. Correlation analysis report Pearson correlation coefficients, unless indicated otherwise. Prior to calculating the median of a response distribution, we excluded extreme outliers: data points that were more than 3 times interquartile range outside of the first or third quartile (Q1 - 3xIQR; Q3 + 3xIQR). Prior to statistical tests, extreme outliers on a subject-level were excluded as well. Due to an unexpected mismatch between software and hardware functionality, participants in the lab study often missed the first trial in the block. We thus omitted the first trial for lab study participants for analyses that included missed trials. Task performance was measured with percentage optimal choice, defined as the number of exploitation trials on which the highest-paying bandit was chosen in relation to the total number of exploitation trials (see *Optimal choice as a measure of task performance* in Supplementary Methods). Exploration percentage was calculated as the number of exploration trials in relation to the number of valid trials, making it a direct opposite of exploitation percentage.

2.2.4 Computational modeling

Computational modeling was done in MATLAB 2020a using custom scripts and materials from Dubois et al. (2021). Reinforcement learning models commonly used to model data from multi-armed bandit tasks in exploration-exploitation studies (e.g. Daw et al., 2006; Gershman, 2018) expect both reward and information to be delivered as feedback on every trial. Hence, they were not directly suitable for modeling the data from the ExploreExploit task. Given that exploration trials yielded only information and not reward, while exploitation trials yielded only reward and no further information, we implemented different expected values for exploration and exploitation, which became a key feature of the computational models we tested (see below).

Best-fitting model for the ExploreExploit task

In the following, we focus on the best-fitting model (the same model showed the best fit for the data of participants in the lab and online). Detailed information on the full model space can be found in the *Computational models* section in Supplementary Methods.

Separate expected values were modeled for exploring and exploiting each bandit. To do so, we capitalized on the Upper Confidence Bound (UCB) algorithm (Auer, 2002) that accounts for the influence of uncertainty about an option. The expected value of exploitation ($V\_exploit_{i,t}$) for bandit i on trial t was defined as the expected reward of this bandit ($Q_{i,t}$):

$$V\_exploit_{i,t} = Q_{i,t}$$

The expected value of exploration ($V\_explore_{i,t}$) was comprised of the sum of the expected reward value ($Q_{i,t}$) of bandit i on trial t and the expected uncertainty ($\sigma_{i,t}$) about this reward, which were weighted by the parameters β_1 and β_2, respectively (estimated as free parameters):

$$V\_explore_{i,t} = \beta_1 * Q_{i,t} + \beta_2 * \sigma_{i,t}$$

For all bandits, the initial expected reward value (Q_0) was fixed to 50 (the middle of the reward range and the mean reward for each bandit over all blocks) and the starting value for expected uncertainty (σ_0) was fixed to 20 (also note that the mean SD across all trials for rewards of all bandits in all reward structures in the lab study was 21).

After a bandit was explored, information (r_t – observed reward on trial t) was received as feedback. A temporal difference (TD) learning model (Sutton, 1988; Sutton & Barto, 1998) with the learning rate α (free parameter) was used to model how one may incorporate newly gained information into the existing belief about the reward of that bandit.

$$Q_{i,t+1} = Q_{i,t} + (r_t - Q_{i,t}) * \alpha$$

At the same time, uncertainty about the reward of the explored bandit decreased:

$$\sigma_{i,t+1} = \sqrt{\sigma_{i,t}^2 - \alpha * \sigma_{i,t}^2}$$

"Forgetting" – a function that makes beliefs about the reward structure less precise, was applied to both expected reward value and expected uncertainty for bandits that were not explored or, in case of exploitation, for all bandits. Forgetting for both expected reward value and expected uncertainty was modeled as a return to starting values (Q_0 and σ_0). Importantly, the winning model included separate forgetting rates λ_1 and λ_2 (free parameters) for expected reward value and expected uncertainty, respectively:

$$Q_{i,t+1} = \lambda_1 * Q_{i,t} + (1 - \lambda_1) * Q_0$$

$$\sigma_{i,t+1} = \lambda_2 * \sigma_{i,t} + (1 - \lambda_2) * \sigma_0$$

Since there were six response possibilities defined by the combinations of three bandits and two actions (exploration and exploitation), six expected values were passed to the softmax choice rule to determine the probability of each response ($P_{i,a,t}$ – probability of applying action *a* to bandit *i* on trial *t*). A trial outcome was then chosen according to the probabilities returned by the softmax algorithm, using inverse temperature τ (free parameter) to determine the stochasticity of the choice; the lower the inverse temperature, the more stochastically the response was chosen (the less it is driven by the largest expected value):

$$P_{i,a,t} = \frac{\exp(\tau * V_{i,a,t})}{\sum_{j,x} \exp(\tau * V_{j,x,t})}$$

where *j* denotes all other bandits and *x* denotes all other actions.

Model selection

Models were fit using the fmincon optimization function in MATLAB and maximum likelihood estimation (MLE) (Wilson & Collins, 2019). The values of free parameters were picked from a uniform distribution with the ranges summarized in **Table 2-S3**. Bayesian information criterion (BIC) was used for model comparison and this was done separately for the MRI and the online group. BIC penalizes model complexity and a lower BIC value indicates a better model fit (Wilson & Collins, 2019).

Parameter recovery

In line with existing recommendations (Wilson & Collins, 2019), we performed parameter recovery for the winning model to check that parameter estimates were reliable. To this end, we simulated 1000 data sets with parameter values derived from a uniform distribution with the same parameter ranges used for

model fitting (**Table 2-S3**) and fit the corresponding model to the simulated datasets to obtain estimated parameter values. We then correlated the original parameter values (with which the data sets were simulated) with the estimated values produced by the model fitting procedure. In addition, we correlated the recovered parameters amongst themselves to verify that the parameter recovery process did not introduce trade-offs between parameters (Wilson & Collins, 2019). Finally, we correlated estimated parameter values and behavioral metrics (e.g. percentage of optimal choice, percentage of exploration trials, and percentage of switch trials (switching between exploration and exploitation)) to check whether model parameters reflected behavior (Danwitz et al., 2022). These correlations were calculated using the Spearman correlation coefficient.

Data simulation

Further, we simulated data using the winning model and each participant's estimated parameter values for this model; we created 10 simulated datasets and picked the one that produced the smallest absolute deviation to the real data in exploration percentage and optimal choice percentage (min (abs (% explore real - % explore simulated) + abs (% optimal choice real - % optimal choice simulated))).

2.3 Results

2.3.1 Behavioral results from the lab study

Task performance

Participants in the lab showed good task performance with a high percentage of optimal choice, ranging from 72% to over 90% (M = 85.3, SD = 4.4; **Figure 2-2**, left), and a low percentage of missed trials (M = 0.8, SD = 1.58; **Figure 2-S4**). High optimal choice percentage demonstrates that participants were motivated and understood the reward structure well.

Participants were instructed to treat rewards in each task block as independent. To verify that they did so, we counted the number of times participants chose the same bandit on the last trial of one block and the first trial of the next block (expressed as a fraction of total number of transitions between blocks; max. 4 transitions for participants with 5 blocks). We then used a linear regression model to test whether the fraction of transitions choosing the same bandit was significantly different from 33% (33% marking the probability of choosing the same bandit by chance): lm ($fraction_{same}$ – 0.33 ~ 1). Number of transitions between blocks on which the same bandit was chosen last on the previous and first on the next block did not differ from chance level (b = 0.02, 95% CI = [-0.02, 0.07], p = 0.34).

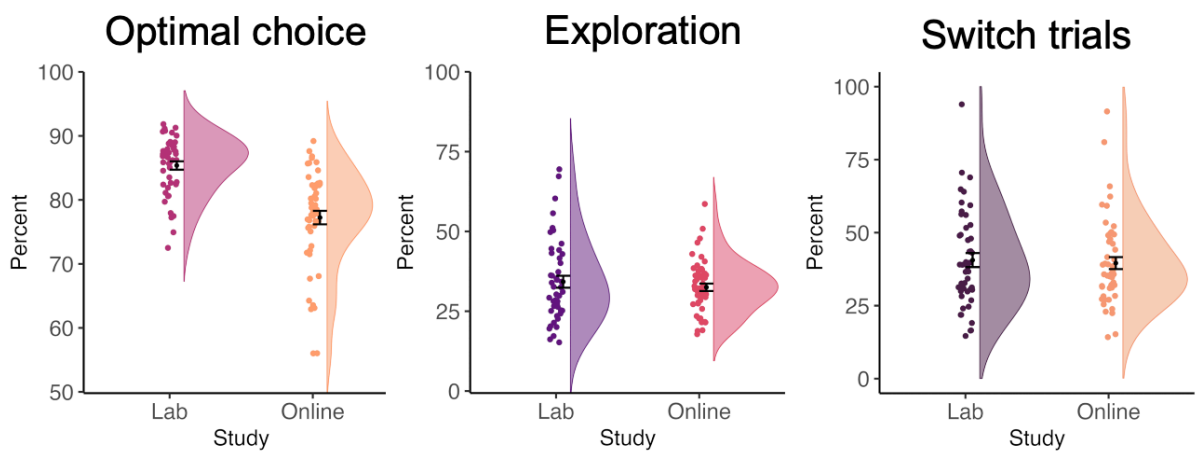


Figure 2-2. Task performance. *Behavioral metrics in the lab and online studies: optimal choice percentage (right), exploration percentage (middle), and percentage of switch trials (right). Optimal choice was calculated as the percentage of exploitation trials on which the highest-paying bandit was exploited. Note that exploitation percentage is the opposite of exploration percentage. Switch trials denote switching between exploration and exploitation.*

ExploreExploit task captures a wide range of exploration-to-exploitation ratios

Participants engaged in exploration on an average of 34% of trials (M = 34.2, SD = 12.8). However, subjects exhibited a wide range of exploration-exploitation ratios (**Figure 2-2**, middle), from highly exploitative behavior (15% exploration) to highly exploratory behavior (70% exploration). Switch trials, on which participants switched between exploration and exploitation, made up 40% of trials (M = 40.5, SD = 16.5, **Figure 2-2**, right).

On average, 1/3 of exploration trials throughout the block

Next, we examined the course of exploration throughout the block (100 trials). For each participant, we calculated the fraction of exploration trials in each position based on the number of total trials available in a given position (max. 5, because there were 5 blocks). The average fraction of exploration throughout the block (excluding trials 1-4, which marked the initial exploration period in the beginning of the block) was 32% (M = 32%, SD = 4%) (**Figure 2-3A**).

We used a linear mixed model to test whether there was a significant change in the course of exploration throughout block trials. The model included fraction of exploration trials as a dependent variable and trial number as an independent variable, in addition to subject ID as a random intercept. The first 4 trials were removed, as they reflect a sharp transition from exploration to exploitation. There was a weak but significant decrease of exploration throughout the block (β = -5.82e-04, 95% CI = [-0.0007, -0.0003], p = 7.52e-08, R^2 = 0.0047). Though significant, this result indicates a decline of 0.06% with an effect size of 0.5%, which we regard as negligible.

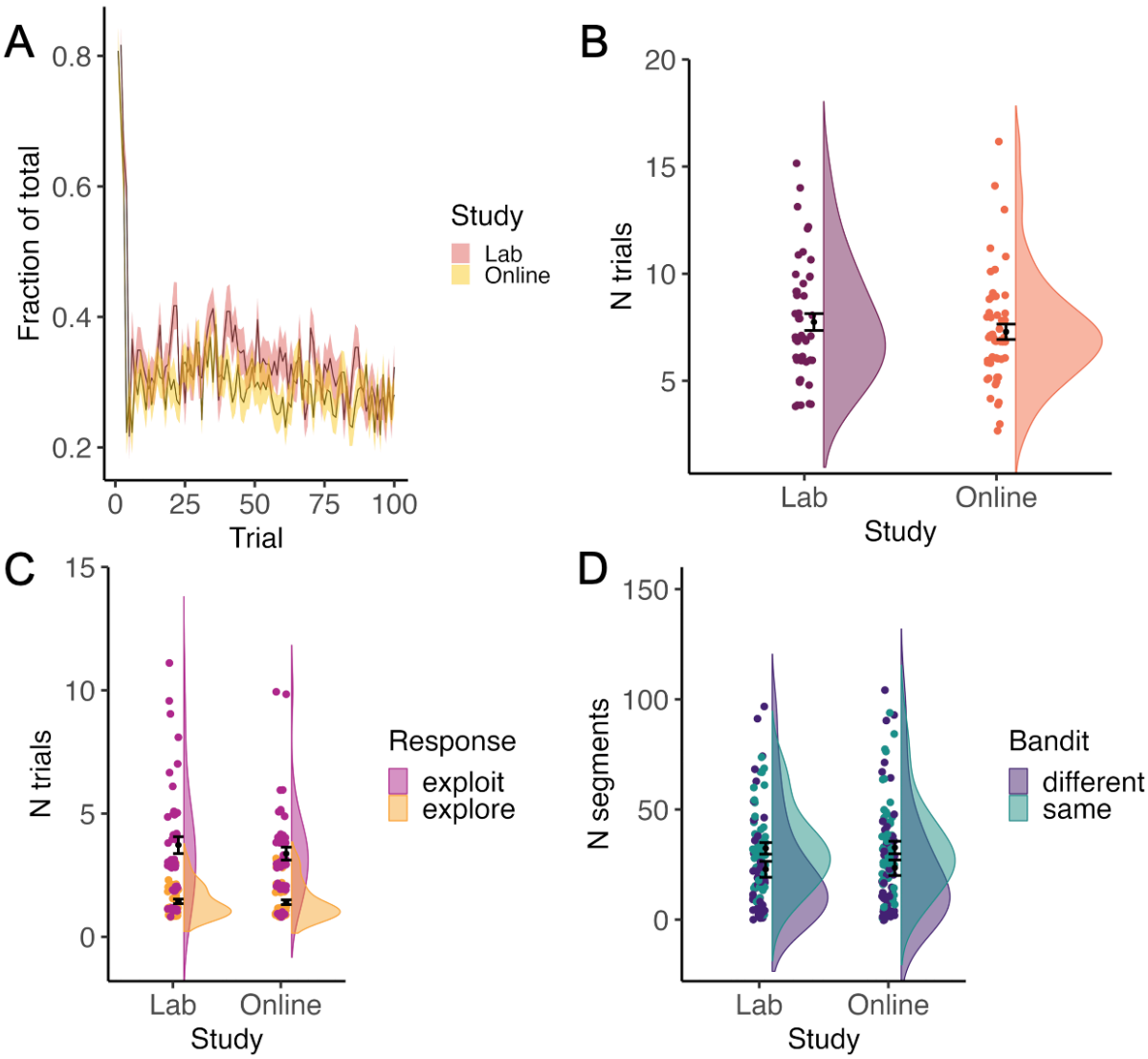


Figure 2-3. Characterizing exploration. *A – Distribution of exploration trials throughout the block, averaged over all blocks. B – How much uncertainty participants tolerate before they explore the bandit again (median number of trials before a bandit is explored again). C – Median number of trials in a continuous exploration and exploitation sequence. D – 1-trial exploration sequences: same – participant explored the same bandit that was exploited on the previous trial, different – participant explored a different bandit from the one they exploited on the previous trial.*

Uncertainty tolerance before exploring again

Next, we checked how many trials passed since the bandit was last explored until it was explored again, that is how high participants' uncertainty about a non-explored option got before they decided to explore it. On average, the median number of trials participants waited until exploring the bandit again was 7 trials (M = 7.7, SD = 2.7) (**Figure 2-3B**). On average, participants did not wait longer than 15 trials (median) until exploring an option again (**Figure 2-3B**).

More time is spent continuously exploiting than exploring

On average, participants exhibited a median continuous sequence length of 3.7 trials (M = 3.7, SD = 2.3) in exploitation and 1.4 (M = 1.4, SD = 0.6) trials exploration (**Figure 2-3C**). A linear mixed model with the median sequence length as a dependent variable and response type as independent variable (plus subject ID as a random intercept) showed that continuous exploitation sequences were significantly longer than exploration sequences (β = -2.28, 95% CI = [-2.94, -1.63], p = 1.34e-08, R^2 = 0.31). Note that one bandit continuously provided the highest reward for, on average, 8 trials (mean and SD between 3 reward structures used in the lab study: M = 8, SD = 14), so participants' behavior, on average, underestimated the time one could spend in exploitation mode.

1-trial sequences dominate exploration

As can be seen in **Figure 2-3C**, exploration was dominated by 1-trial sequences. We used a linear mixed model to test whether 1-trial exploration was used more often on the same bandit that was exploited immediately prior to that exploration trial (checking the value of the preferred bandit). The dependent variable in the model was the number of 1-trial exploration sequences that were applied to either the same bandit that was exploited immediately before that or to a different bandit. The independent variables were previous bandit (same – participants explored the same bandit they were exploiting; different – the explored bandit was different from the one previously exploited), and subject ID as a random intercept. Prior to the analysis, one participant was excluded because of an extreme outlier value on using 1-trial exploration applied to different bandit. There was a weak significant effect of previous bandit (β = 9.52, 95% CI = [1.93, 17.11], p = 0.01, R^2 = 0.04; **Figure 2-3D**): 1-trial exploration sequences were used more often to explore the same bandit, which was exploited immediately before that.

Observing a low reward leads to disengaging from a bandit; observing a high reward is followed by exploiting that bandit

We then examined how the reward seen on an exploration trial relates to the response made on the following trial. For this purpose, we divided exploration and exploitation responses into 2 sub-categories depending on whether the same or different bandit was chosen as the one explored on the trial before. Hence, there were 4 possible responses: explore-same (exploring the same bandit again), explore-different (exploring another bandit), exploit-same (exploiting the bandit that was just explored), and exploit-different (exploiting another bandit).

The number of trials for each response made after seeing each possible value of reward (rewards could take values from 1 to 100) was expressed as a fraction of total number of trials on which that reward value was observed. If a reward value was observed by only one participant in the study, it was omitted to avoid skewing the response distribution. To reduce noise in the response distribution, we binned rewards (ranging 1-100 points) into 10 bins with 10 reward points each.

Results are presented in **Figure 2-4A**. The explore-same category was rarely present in the data, suggesting that one exploration trial was enough to update participants' idea about how a given bandit's reward was changing. Two responses (explore-different and exploit-different) were most prevalent after observing a reward the lower reward range, suggesting that subjects will typically disengage from such low-value bandits. On the other hand, seeing a reward in the upper reward range was predominantly followed by exploit-same responses, indicating that observed reward magnitude was considered sufficient for transitioning to exploitation.

The best-paying bandit dominates exploitation, but more equal distribution of bandits chosen during exploration

To test whether exploration-exploitation behavior of the participants reflected the goal of gathering information vs collecting reward, respectively, we analyzed how the reward rank of the bandit (1 – highest-paying bandit on the current trial, 2 – bandit with middle reward, 3 – lowest-paying bandit) related to exploration and exploitation choices. Most exploitation choices were directed to the bandit with the highest reward, while the distribution of exploration trials was much more equal among bandits (**Figure 2-4B**).

We used a Type III repeated-measures ANOVA model, specifying number of trials in each category as a dependent variable and within-subject factors response type (exploration, exploitation) and bandit rank (1, 2, 3) as independent variables. We also modeled an interaction between them. The F-test showed a significant interaction ($F(2, 92) = 633.25$, $p < 2.2e\text{-}16$, $\eta^2 = 0.93$), which we followed up with a series of paired t-tests (FDR-corrected). Specifically, we compared the same response type between different reward ranks and different response types within the same rank. Results revealed that responses within each rank were significantly different and there were significant differences between the same response for each two ranks (all $p < 0.006$, results are summarized in **Table 2-1**).

2.3.2 Online replication study

Behavioral results seen in the lab study were replicated in the online study.

Participants in the online study generally showed good task performance (optimal choice percentage: M = 77.2, SD = 7.6), though they made significantly fewer optimal choices than participants in the lab study ($\beta = -8.14$, 95% CI = [-10.67, -5.60], $p = 6e\text{-}09$, adj. $R^2 = 0.28$). Percentage of missed, exploration, and switch trials did not differ significantly (all $p < 0.4$) between the two studies. Online participants treated rewards of each block as independent (i.e., the same bandit was chosen last on one block and first on the next at chance level; $b = 0.01$, 95% CI = [-0.06, 0.08], $p = 0.73$).

Like in the lab study, online participants explored 1/3rd of the time (M = 32.5, SD = 8.2), switched 40% of the time (M = 39.5, SD = 14.8), and expressed a wide range (17-60%) of exploration-exploitation ratios (**Figure 2-2**). Exploration declined very slightly throughout the block ($\beta = -3.74e\text{-}04$, 95% CI = [-0.0005, -0.0001], $p = 0.0002$, $R^2 = 0.0022$). The median time most participants waited until exploring the

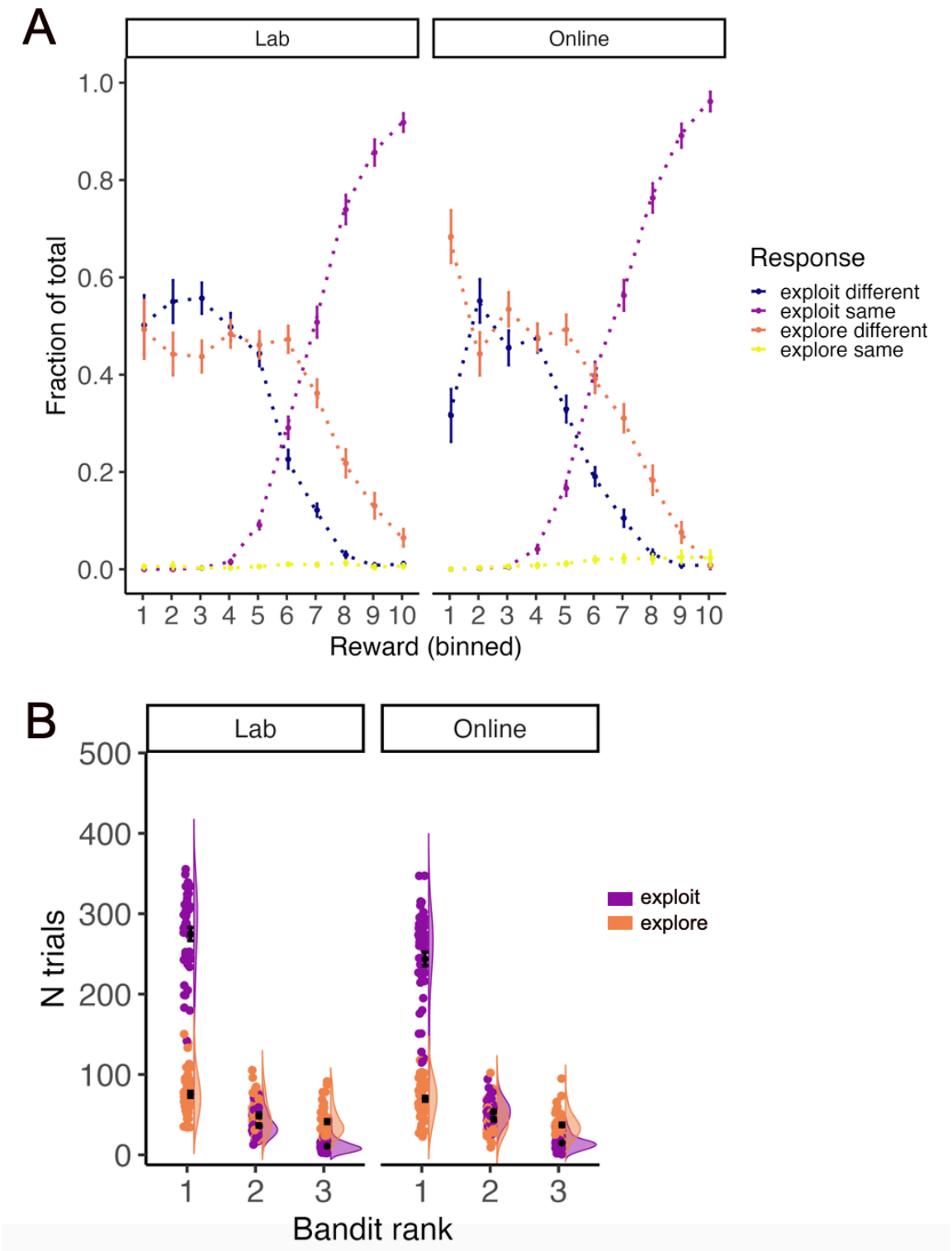


Figure 2-4. Influence of reward on exploration and exploitation. *A – Number of trials a bandit of rank 1 (highest reward on the current trial), 2 (middle reward) or 3 (lowest reward) was explored and exploited. B – Responses made after seeing a reward on a preceding exploration trial. Rewards are binned with 10 reward points per bin. Responses are expressed as a fraction of all trials for a given bin. Same and different refers to the bandit chosen in relation to the one that was explored on the preceding trial. Lab – lab study, online – online study.*

Table 2-1. Exploring and exploiting each bandit rank*. Results of follow-up comparisons (FDR-corrected). df – degrees of freedom, η^2 – semi-partial eta squared.*

Contrast	df	t-ratio	p-value	η^2
Lab study				
exploit rank 1 - exploit rank 2	46	31.25	<.0001	0.95
exploit rank 1 - exploit rank 3	46	33.11	<.0001	0.95
exploit rank 2 - exploit rank 3	46	14.94	<.0001	0.82
explore rank 1 - explore rank 2	46	10.18	<.0001	0.69
explore rank 1 - explore rank 3	46	11.42	<.0001	0.73
explore rank 2 - explore rank 3	46	5.13	<.0001	0.36
exploit rank 1 - explore rank 1	46	17.31	<.0001	0.86
exploit rank 2 - explore rank 2	46	-2.84	0.0067	0.14
exploit rank 3 - explore rank 3	46	-9.58	<.0001	0.66
Online study				
exploit rank 1 - exploit rank 2	51	22.31	<.0001	0.90
exploit rank 1 - exploit rank 3	51	25.73	<.0001	0.92
exploit rank 2 - exploit rank 3	51	21.66	<.0001	0.90
explore rank 1 - explore rank 2	51	9.52	<.0001	0.64
explore rank 1 - explore rank 3	51	10.78	<.0001	0.69
explore rank 2 - explore rank 3	51	5.12	<.0001	0.34
exploit rank 1 - explore rank 1	51	20.24	<.0001	0.88
exploit rank 2 - explore rank 2	51	3.15	0.0027	0.16
exploit rank 3 - explore rank 3	51	-9.44	<.0001	0.63

bandit again was 7 trials (M = 7.2, SD = 2.5), with the upper range going to 16 trials (**Figure 2-3B**). Median length of continuous exploitation sequences was significantly longer than the median length of continuous exploration sequences (β = -1.97, 95% CI = [-2.40, -1.53], p = 5.95e-12, R^2 = 0.33; **Figure 2-3C**). One-trial long sequences were most frequent in exploration (**Figure 2-3C**) and were applied significantly more often to the same bandit (that was exploited prior to exploration trial) than to a different bandit (β = 9.15, 95% CI = [0.64, 17.66], p = 0.03, R^2 = 0.03; **Figure 2-3D**), though the effect was quite weak.

Like in the lab study, observing a reward in the lower reward range on an exploration trial resulted in disengaging from the bandit on the following trial (exploring or exploiting a different bandit), while observing a reward in the higher reward range was followed by exploiting the same bandit on the next trial (**Figure 2-4A**). There was a strong dominance of choosing the bandit with the highest reward (rank 1) on exploitation trials, while exploration trials were much more equally distributed among the bandits of reward ranks (**Figure 2-4B**). An F-test with the number of exploration and exploitation trials applied to bandits of different reward ranks showed a significant interaction (F(2, 102) = 457.86, p < 2.2e-16, η^2 = 0.9) between response type (explore, exploit) and reward rank of the bandit (1, 2, 3). All follow-up comparisons testing the difference between the same response type between different ranks and

different response types within the same rank were significant (all p < 0.002, results are summarized in **Table 2-1**). The effects replicated those seen in the lab study, with the only difference that the bandit with the second-highest reward was exploited more often than it was explored (this effect had the opposite direction in the lab study).

2.3.3 Computational modeling results

Model comparison

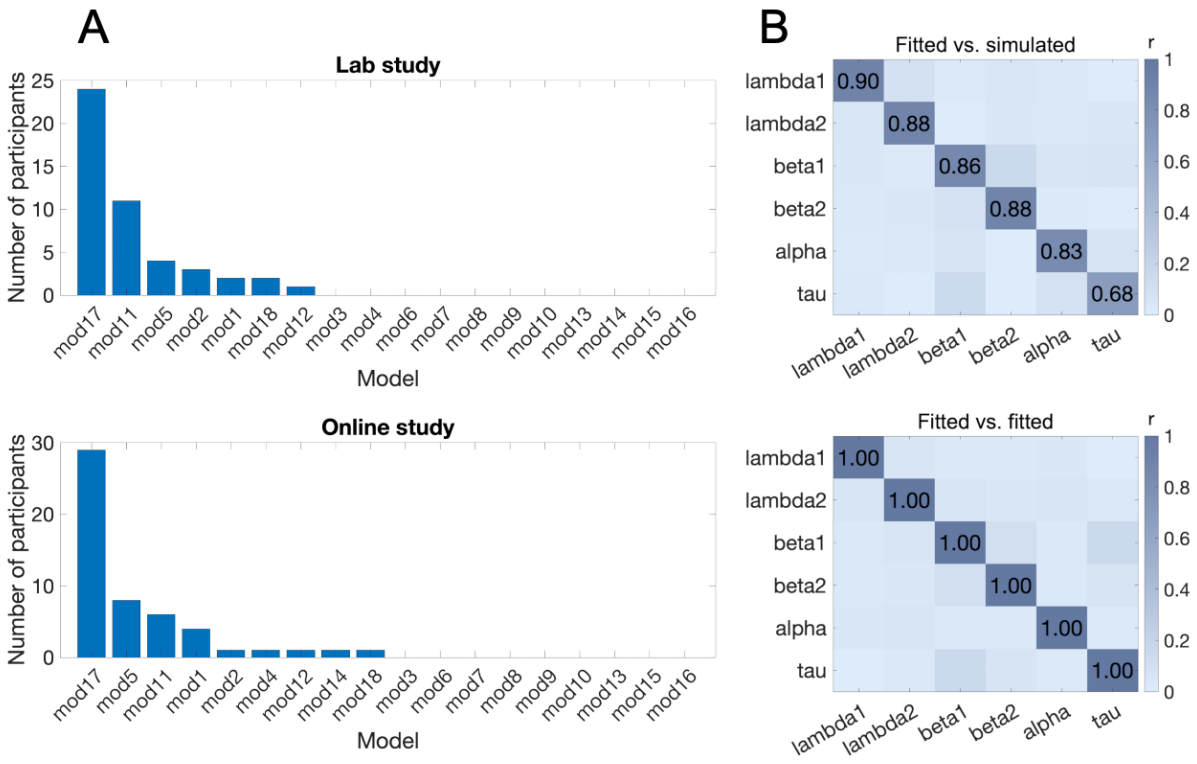


Figure 2-5. Computational modeling results. *A – Model comparison: number of participants with the lowest BIC score for the respective model in the lab study (top) and the online study (bottom). B – Parameter recovery for the winning model: top – correlation between simulated and fitted (recovered) parameter values; bottom – correlation of fitted (recovered) parameter values with each other. Alpha – learning rate (α), beta1 – weight for reward in exploration (β_1), beta2 – weight for uncertainty in exploration (β_2), lambda1 – forgetting rate for reward (λ_1), lambda2 – forgetting rate for uncertainty (λ_2), tau – inverse temperature (τ).*

We assessed how well multiple computational models (within either UCB or discounting model families – cf. *Computational models* in Supplementary Methods) reflected participants' behavior in the ExploreExploit task. Model comparison revealed that the model with the lowest BIC score for most participants in both lab and online studies was model 17 (**Figure 2-5A**). The winning model (described in detail in *Computational modeling* in Methods) was a UCB-type model (with the exploration value for each bandit comprised of a weighted sum of expected reward and uncertainty about it). A special characteristic of this model was the presence of separate forgetting rates (λ_1 and λ_2) for expected reward (Q) and expected uncertainty (σ).

Parameter Recovery

There was a fairly good correspondence between the simulated and recovered parameters of the winning model (**Figure S2-5**), with correlations between simulated and recovered parameters (**Figure 2-5B**) ranging from 0.68 for the inverse temperature parameter (τ) to 0.90 for the forgetting rate for reward (λ_1). The absence of strong correlations between recovered parameters (**Figure 2-5B**) indicated that parameters of the winning model did not appreciably trade-off against each other (Wilson & Collins, 2019).

Correlations between estimated parameter values and behavioral metrics

We computed correlations (Spearman rank correlation) between estimated parameter values and behavioral metrics from each participant's data (optimal choice percentage, exploration percentage, and switch trials percentage). Results demonstrate that estimated model parameters reflected behavior in the ExploreExploit task (**Figure 2-6**, see **Table 2-2** for a summary of all significant correlations). We discuss significant correlation in the following.

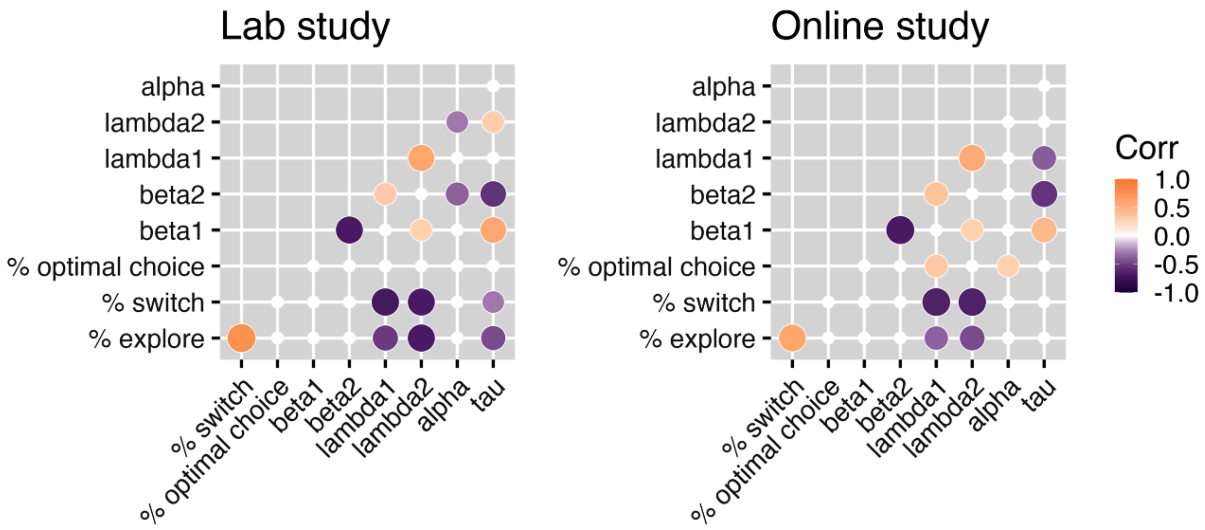


Figure 2-6. Correlations between estimated model parameters and behavioral metrics. *Only significant (all $p < 0.05$, uncorrected) correlations are presented. Alpha – learning rate (α), beta1 – weight for reward in exploration (β_1), beta2 – weight for uncertainty in exploration (β_2), lambda1 – forgetting rate for reward (λ_1), lambda2 – forgetting rate for uncertainty (λ_2), tau – inverse temperature (τ).*

Table 2-2. Correlations between estimated parameter values and behavioral metrics. *Only correlations with a significant ($p < 0.05$, uncorrected) result in at least one of the studies are listed (cf.* ***Figure 2-6****). λ_1 – forgetting rate for expected reward, λ_2 – forgetting rate for expected uncertainty, τ – inverse temperature, α – learning rate, cor - Spearman rank correlation coefficient.*

Parameter	Behavior, %	Study	cor	p-value
λ_1	explore	lab	-0.53	1.23e-04
		online	-0.36	7.03e-03
	switch	lab	-0.71	2.26e-08

λ_1		online	-0.66	7.05e-08
λ_1	optimal choice	lab	-	n.s.
		online	0.32	0.01
λ_2	explore	lab	-0.69	7.74e-08
		online	-0.44	9.40e-04
λ_2	switch	lab	-0.68	1.42e-07
		online	-0.65	1.39e-07
τ	explore	lab	-0.45	1.22e-03
		online	-	n.s.
τ	switch	lab	-0.30	0.03
		online	-	n.s.
α	optimal choice	lab	-	n.s.
		online	0.27	0.04

There was a negative correlation between forgetting rates for reward (λ_1) and uncertainty (λ_2) and exploration percentage, indicating that as forgetting rates decrease (so that expected reward and uncertainty values quickly go back to baseline, where they are the same for all bandits), exploration rate increases.

We found a negative correlation between inverse temperature (τ, higher values = less stochastic choices) and exploration percentage, thus revealing that as choice stochasticity increases, so too does the rate of exploration.

We also noted a positive correlation between learning rate (α) and optimal choice percentage in the online study, indicating that participants who update expected reward and uncertainty stronger (and consequently switch to a different bandit quickly after seeing it providing more reward) choose the highest-paying bandit more often and perform better.

Lastly, optimal choice percentage correlated positively with the forgetting rate for expected reward (λ_1) in the online study. This shows that, as forgetting rate increases and rewards for all bandits are "forgotten" slower (so that learned information about the expected value is regarded as valid for a longer time), participants make more optimal choices. This is likely because they exploit their preferred option longer (as suggested by negative correlations between switch percentage and forgetting rate for reward, and between exploration percentage and forgetting rate for reward – so the higher forgetting rate for reward, the higher exploitation percentage and the lower the switch percentage).

In the online study, correlation patterns largely replicated those seen in the lab study. A correlation between optimal choice percentage and learning rate (α), as well as a correlation between optimal choice percentage and forgetting rate for reward (λ_1) were present in the online study, but not in the lab study, possibly because performance was close to the ceiling in the lab study).

Simulated data captures behavioral features of the empirical data

We simulated data for each participant (combining participants from the lab and online study) using estimated parameters from the winning model. Overall, behavioral characteristics of the simulated data showed fairly good resemblance to the real data. While exploration and optimal choice percentage were well approximated in the simulated data (positioned along the identity line), switch percentage was often overestimated in the lower range and underestimated in the higher range (**Figure 2-7A**).

The length of continuous exploration and exploitation sequences was largely underestimated in the simulated data (**Figure 2-7B**), while it reflected correctly how many trials passed until a bandit was explored again (**Figure 2-7C**). The distribution of exploration trials across the reward ranks matched that found in the real data, but the number of exploitation trials allocated to bandit ranks 2 and 3 was overestimated (**Figure 2-7D**).

2.4 Discussion

In the current study, we presented the ExploreExploit task, which by combining key characteristics of previously used paradigms allows for a reliable and detailed investigation of human exploration-exploitation behavior. Successful replication of both behavioral and computational modeling results in the online experiment demonstrates that ExploreExploit task can be employed in multiple experimental contexts and produces stable results regardless of experimental setting.

2.4.1 The value of the ExploreExploit task lies in successfully combining important features of previously used paradigms

In contrast to multi-armed bandit paradigms which use a computational model to determine the trial type (e.g. Daw et al., 2006), the ExploreExploit task provides a behavioral marker of the trial type coded directly into response buttons (Tversky & Edwards, 1966), thus allowing participants to directly indicate whether they are exploring or exploiting. This feature analyses contrasting exploration and exploitation trials more robust due to precise alignment of neural and physiological data with behavior (trial category). The task is thus well suited to examine neurocognitive and physiological (such as eye tracking) correlates of exploration-exploitation behavior.

In line with previous literature, exploration trials comprised, on average, ~30% of trials (Chakroun et al., 2020; Muller et al., 2019) and exploration rate was close to 1/3 throughout the block (Blanchard & Gershman, 2018). In contrast to tasks with deterministic rewards (e.g. Navarro et al., 2016), the ExploreExploit task ensures a higher proportion of exploration trials, continuously encouraging exploration throughout the block. This feature further supports the use of the ExploreExploit task with neuroimaging methods, allowing one to compute more reliable contrasts based on a higher number of exploration trials (Nee, 2019).

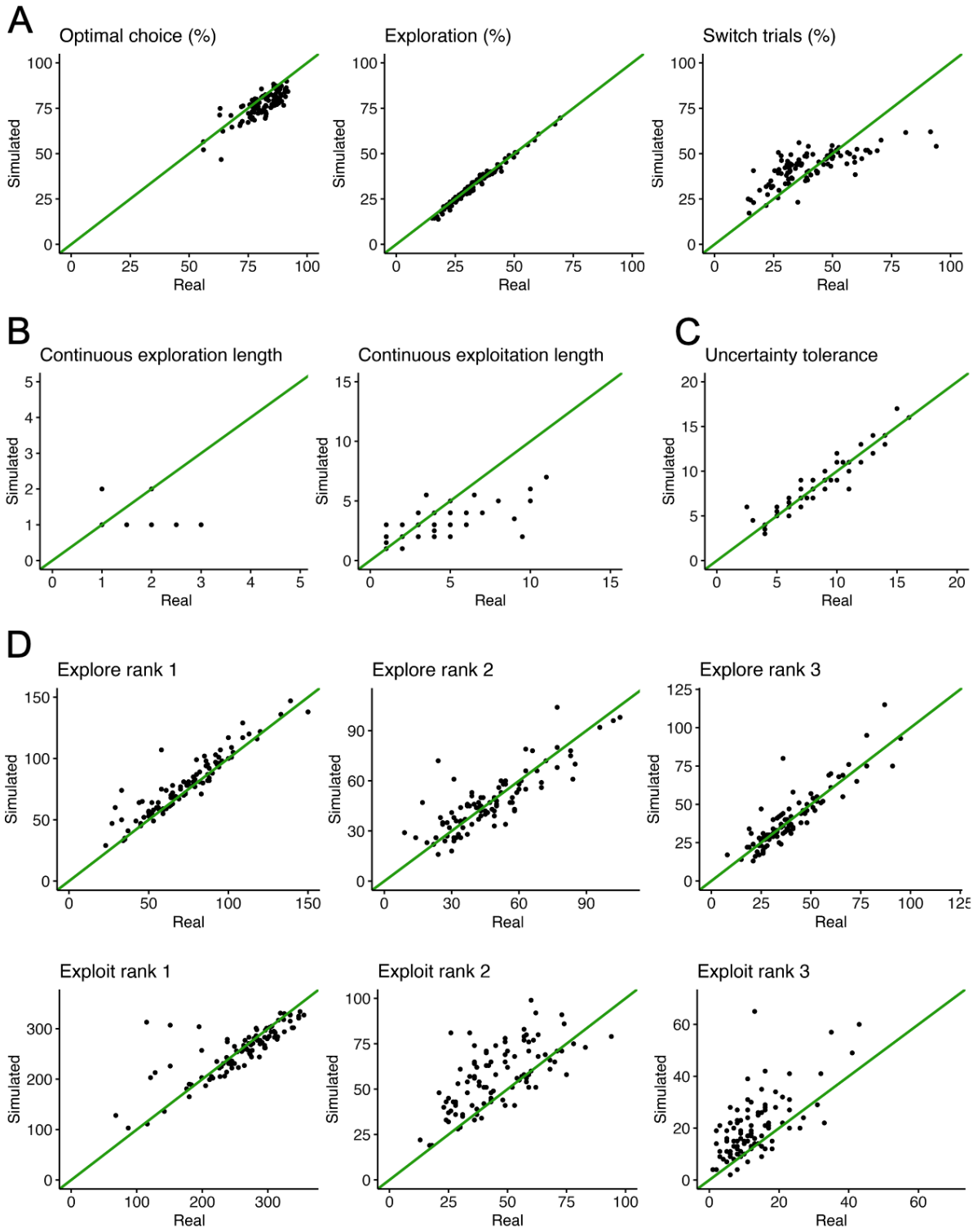


Figure 2-7. Real data for each participant plotted vs their simulated data. *Green – identity line. A – Behavioral metrics: percentage of optimal choice (left), exploration (middle), and switch trials (right). B – Median number of trials in continuous exploration (left) and exploitation (right) sequences. C – Median number of trials until a bandit is explored again (how much uncertainty is tolerated). D – Upper panel: number of trials exploring bandit of rank 1 (left), 2 (middle) and 3 (right); lower panel: number of trials exploiting bandit of rank 1 (left), 2 (middle) and 3 (right).*

In contrast to exploration-exploitation paradigms which focus on specific trials at the cost of observing a natural flow of behavior (e.g. Wilson et al., 2014), the ExploreExploit task probed naturally-paced exploration-exploitation behavior by capturing highly explorative, exploitative, and switching behavior when it occurred. Most participants exploited more than they explored, using longer continuous exploitation sequences and shorter continuous exploration sequences, indicating that they understood and could follow the underlying non-stationary reward structure well.

In line with the conceptual distinction between exploration and exploitation as focusing on gaining reward vs. information (Blanchard & Gershman, 2018; Mehlhorn et al., 2015), participants mostly exploited the highest-paying bandit, while exploration trials were much more evenly distributed between the bandits of all reward ranks. Notably, the reward rank of the bandit played a distinct role in both exploitation and exploration, with more trials of both types spent engaging with higher ranked bandits. In addition, we extend previous findings showing that participants explore more after receiving a lower reward and less after receiving a high reward (Song et al., 2019) by presenting a more detailed picture. In our study, exploitation clearly dominated after a reward in the upper reward range was observed. However, seeing a reward in the lower range was not necessarily followed by exploration responses, but instead by responses indicating disengagement from the low-reward option; participants either explored or exploited another bandit. Such response functions for each response category provide a useful tool to directly observe the influence of reward structure manipulations on exploration-exploitation behavior.

A reward structure based on reward magnitude (as opposed to reward probability) provides a wider range of opportunities for examining the influence of the reward structure features on exploration-exploitation behavior. Such features include magnitude-related reward characteristics (e.g. changes in mean reward and volatility of the reward structure (Speekenbrink & Konstantinidis, 2015)) or a combination of magnitude- and probability-based features (e.g. sensitivity to gains and losses (Tom et al., 2007)). Importantly, magnitude-based rewards also allow for easier scaling of the number of bandits, since multiple magnitude-based options are more easily discriminated than multiple probability-based alternatives. Increasing the number of bandits could be useful to address further questions, such as generalization in exploration-exploitation behavior in environments with spatially correlated rewards (Schulz et al., 2018; Wu et al., 2018).

2.4.2 Our computational model reflects adaptations of exploration-exploitation decision-making to the task, as well as relationships between decision-making processes and behavior

The ExploreExploit task provides a direct behavioral marker of the trial type, so that computational modeling is not required to categorize trials. Model parameters can thus be used to gain deeper insight into the mechanics of the exploration-exploitation decision-making without potential issues caused by model parameters being used to produce reward estimates, which are typically then used to categorize the trials. For example, learning rates in our study are high, reflecting an adaptation to a changing reward

structure, in which most recent information possesses highest relevance (Behrens et al., 2007; Courville et al., 2006). In contrast, the learning rate in a rarely-changing environment would be lower, since occasional changes are treated as outliers rather than indicators of change (Piray & Daw, 2021). Further, a negative correlation between the weights for reward and uncertainty in the expected value of exploration suggests that the more uncertainty was taken into account, the weaker was the influence of reward, and vice versa. While exploration is modeled as driven by a combination of reward and uncertainty, this result shows how motivation to know more about the bandit with most reward and motivation to know more about the most uncertain bandit might be trading off. In addition, our computational model contains separate forgetting rates for reward and uncertainty, thus emphasizing the distinctiveness of reward and uncertainty information.

In addition, estimated parameter values reflected behavioral metrics in the ExploreExploit data, offering further insights into how the elements of the decision-making process might influence behavioral performance. A simulation study (Danwitz et al., 2022) on computational modeling for a multi-armed bandit task (Daw et al., 2006) reported a positive relationship between the optimal choice percentage and the tendency to choose items with the higher expected value (softmax stochasticity parameter), as well as a positive relationship between the optimal choice percentage and exploration bonus parameter (parameter denoting the weight for uncertainty in the expected value). This study also reports positive relationship between the switch percentage and exploration bonus parameter, as well as a negative relationship between switch percentage and optimal choice percentage. While our results also show a positive association of exploration and switch percentage, suggesting that longer exploitation sequences might represent a more typical type of exploration-exploitation balance, we don't find a correlation between switch percentage and optimal choice. The latter effect is likely explained by the fact that switching between bandits is considered in (Danwitz et al., 2022), so that higher switch percentage by definition involves choosing sub-optimal options more often, while switching in our study reflects transitions between exploring and exploiting. Extending the findings of (Danwitz et al., 2022), our study highlights the importance of the learning rate (quickly reacting to change) and forgetting rate for reward (preserving a certain degree of stability of the estimates) for high optimal choice percentage in a changing environment.

Though some behavioral features (like optimal choice percentage, exploration percentage, and exploring different reward ranks) were well approximated in simulated data, others (like switching between exploration and exploitation and the length of continuous exploration and exploitation sequences) were less well approximated. More research is needed to fine-tune our computational model to capture these behavioral characteristics.

2.4.3 Using the ExploreExploit task in future research

Thanks to the combination of a flexible reward structure and unambiguous separation of exploration and exploitation responses, the ExploreExploit task can be used in the future to verify the effects of the reward structure manipulations on exploration-exploitation behavior previously presented in the

literature (e.g. higher volatility leading to higher exploration rates (Speekenbrink & Konstantinidis, 2015)). The frequency of each response type based on characteristics of the reward structure (sigmoid functions in **Figure 2-4**) can easily be read out in the ExploreExploit data, providing a useful tool for observing how manipulations of the reward structure affect exploration-exploitation behavior. The features of the reward structure that can be manipulated in the ExploreExploit paradigm include a wide array of characteristics, e.g. reward range (Clarenau et al., 2024), reward stochasticity and volatility (Piray & Daw, 2021; Speekenbrink & Konstantinidis, 2015), sensitivity to wins and losses (Tom et al., 2007), and generalization (Schulz et al., 2018; Wu et al., 2018). Moreover, the winning computational model in our study provides a tool for simulating the influence of changes in the reward structure on behavior and, if needed, fine-tune rewards to elicit desired behavior.

Although the ExploreExploit task has shown stable results in lab and online, it also has the potential to be visually implemented as a game (cf. Dubois & Hauser, 2022) and be presented in a smartphone app, which could greatly increase the outreach, potential sample sizes and diversity of the samples. Smartphone-based experiments were reported to produce comparable results to in-lab studies (Pin & Rotesi, 2023). Furthermore, the ExploreExploit task is particularly well suited to be paired with neurocognitive and physiological methods to examine the correlates of exploration and exploitation behavior. Unambiguous categorization of the trial type based on a direct behavioral measure increases validity of contrasting neural data between exploration and exploitation due to a more precise pairing between behavior (trial type) and neural signals. For a computational model to achieve the same results at trial categorization, it would have to be extremely precise in estimating behavior and be based on correct assumptions about how this behavior is produced (e.g. exploration strategies (Blanchard & Gershman, 2018)). In addition, our task produces a relatively high number of exploration trials, which is beneficial for analyses that require a large number of trials to be viable (e.g. fMRI analyses (Nee, 2019)).

2.4.4 Summary

The ExploreExploit task combines a multitude of useful features to allow for a detailed and robust investigation of exploration-exploitation behavior. They include providing a direct readout of the trial type, capturing naturally-paced exploration-exploitation behavior, allowing for a variety of modifications of the reward structure, and encouraging relatively high exploration rates throughout the task horizon. In addition, our task shows stable behavioral and computational modeling results in multiple experimental settings. These features make ExploreExploit a novel and easily implemented addition to the collection of exploration-exploitation paradigms in the literature today.

Pairing the ExploreExploit task with neurocognitive and physiological methods could be ideal for understanding the biological bases of exploration-exploitation. We examine such pairings in the next two chapters of this dissertation; Chapter 3 investigates the neural correlates of exploration-exploitation behavior as measured with fMRI, while Chapter 4 examines the eye tracking signatures of exploration-exploitation behavior.

2.5 References

Addicott, M. A., Pearson, J. M., Froeliger, B., Platt, M. L., & McClernon, F. J. (2014). Smoking automaticity and tolerance moderate brain activation during explore–exploit behavior. *Psychiatry Research: Neuroimaging*, *224*(3), 254–261. https://doi.org/10.1016/j.pscychresns.2014.10.014

Auer, P. (2002). Using Confidence Bounds for Exploitation-Exploration Trade-offs. *Journal of Machine Learning Research*, *3*, 397–422.

Averbeck, B. B. (2015). Theory of Choice in Bandit, Information Sampling and Foraging Tasks. *PLOS Computational Biology*, *11*(3), e1004164. https://doi.org/10.1371/journal.pcbi.1004164

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/10.1038/nn1954

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

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/10.3758/s13415-017-0556-2

Chakroun, K., Mathar, D., Wiehler, A., Ganzer, F., & Peters, J. (2020). Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making. *ELife*, *9*, e51260. https://doi.org/10.7554/elife.51260

Clarenau, V. C. von, Appelhoff, S., Pachur, T., & Spitzer, B. (2024). Over- and Underweighting of Extreme Values in Decisions From Sequential Samples. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001530

Cockburn, J., Man, V., Cunningham, W. A., & O'Doherty, J. P. (2022). Novelty and uncertainty regulate the balance between exploration and exploitation through distinct mechanisms in the human brain. *Neuron*, *110*(16), 2691-2702.e8. https://doi.org/10.1016/j.neuron.2022.05.025

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/10.1098/rstb.2007.2098

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*(7), 294–300. https://doi.org/10.1016/j.tics.2006.05.004

Danwitz, L., Mathar, D., Smith, E., Tuzsus, D., & Peters, J. (2022). Parameter and Model Recovery of Reinforcement Learning Models for Restless Bandit Problems. *Computational Brain & Behavior*, *5*(4), 547–563. https://doi.org/10.1007/s42113-022-00139-0

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. https://doi.org/10.1038/nature04766

Dubois, M., Habicht, J., Michely, J., Moran, R., Dolan, R. J., & Hauser, T. U. (2021). Human complex exploration strategies are enriched by noradrenaline-modulated heuristics. *ELife*, *10*, e59907. https://doi.org/10.7554/elife.59907

Dubois, M., & Hauser, T. U. (2022). Value-free random exploration is linked to impulsivity. *Nature Communications*, *13*(1), 4542. https://doi.org/10.1038/s41467-022-31918-9

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*(PLoS Computational Biology 6 2010), 34–42. https://doi.org/10.1016/j.cognition.2017.12.014

Hogeveen, J., Mullins, T. S., Romero, J. D., Eversole, E., Rogge-Obando, K., Mayer, A. R., & Costa, V. D. (2022). The neurocomputational bases of explore-exploit decision-making. *Neuron*, *110*(11), 1869-1879.e5. https://doi.org/10.1016/j.neuron.2022.03.014

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/10.1162/jocn.2010.21548

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., Hausmann, D., Fiedler, K., & Gonzalez, C. (2015). Unpacking the Exploration–Exploitation Tradeoff: A Synthesis of Human and Animal Literatures. *Decision*, *2*(3), 191–215. https://doi.org/10.1037/dec0000033

Muller, T. H., Mars, R. B., Behrens, T. E., & O'Reilly, J. X. (2019). Control of entropy in neural models of environmental state. *ELife*, *8*. https://doi.org/10.7554/elife.39404

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Navarro, D. J., Newell, B. R., & Schulze, C. (2016). Learning and choosing in an uncertain world: An investigation of the explore–exploit dilemma in static and dynamic environments. *Cognitive Psychology*, *85*, 43–77. https://doi.org/10.1016/j.cogpsych.2016.01.001

Nee, D. E. (2019). fMRI replicability depends upon sufficient individual-level data. *Communications Biology*, *2*(1), 130. https://doi.org/10.1038/s42003-019-0378-6

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pin, P., & Rotesi, T. (2023). App-based experiments. *Journal of Economic Psychology*, *99*, 102666. https://doi.org/10.1016/j.joep.2023.102666

Piray, P., & Daw, N. D. (2021). A model for learning based on the joint estimation of stochasticity and volatility. *Nature Communications*, *12*(1), 6587. https://doi.org/10.1038/s41467-021-26731-9

Richardson, J. T. E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, *6*(2), 135–147. https://doi.org/10.1016/j.edurev.2010.12.001

Schulz, E., Wu, C. M., Huys, Q. J. M., Krause, A., & Speekenbrink, M. (2018). Generalization and Search in Risky Environments. *Cognitive Science*, *42*(8), 2592–2620. https://doi.org/10.1111/cogs.12695

Song, M., Bnaya, Z., & Ma, W. J. (2019). Sources of suboptimality in a minimalistic explore–exploit task. *Nature Human Behaviour*, *3*(4), 361–368. https://doi.org/10.1038/s41562-018-0526-x

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and Exploration in a Restless Bandit Problem. *Topics in Cognitive Science*, *7*(2), 351–367. https://doi.org/10.1111/tops.12145

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*(1), 9–44. https://doi.org/10.1007/bf00115009

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning* (Vol. 135). MIT press.

Team, R. C. (2022). *R: A language and environment for statistical computing*. URL https://www.R-project.org/

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The Neural Basis of Loss Aversion in Decision-Making Under Risk. *Science*, *315*(5811), 515–518. https://doi.org/10.1126/science.1134239

Tversky, A., & Edwards, W. (1966). Information versus reward in binary choices. *Journal of Experimental Psychology*, *71*(5), 680–683. https://doi.org/10.1037/h0023123

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *ELife*, *8*, e49547. https://doi.org/10.7554/elife.49547

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. https://doi.org/10.1037/a0038199

Wittkuhn, L., Krippner, L. M., & Schuck, N. W. (2022). Statistical learning of successor representations is related to on-task replay. *BioRxiv*, 2022.02.02.478787. https://doi.org/10.1101/2022.02.02.478787

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/10.1038/s41562-018-0467-4

3. Uncertainty-driven brain signal variability adapts exploration-exploitation behavior to a changing environment

Abstract

Past research has shown that uncertainty-driven brain signal variability could be a useful marker to understand how behavior adapts to changing environmental demands. However, it remains unknown whether brain signal variability could be a neural mechanism supporting successful adaptation of exploration-exploitation behavior to a changing environment. In the present study, we administered a newly designed ExploreExploit task that allows unambiguous separation of exploration and exploitation trials during fMRI. In a sample of 40 younger adults, we demonstrate that BOLD signal variability driven by uncertainty could provide a neural mechanism for flexibly switching between exploration and exploitation. We show that better and more flexible performance was related to greater BOLD variability decreases during exploration. Moreover, participants who showed more flexible behavior (not staying too long continuously in exploitation mode), increased BOLD variability less strongly during exploitation, while at the same time exhibiting higher levels of BOLD variability. Higher levels of BOLD signal variability in the beginning of exploitation might thus enable participants to move out of the exploitation mode faster and allow them to more flexibly switch between exploration and exploitation. Our results present a broad network of brain regions underlying neural variability effects, including those associated with uncertainty processing and behavioral flexibility, thus emphasizing the importance of these functions for balancing exploration and exploitation in a changing environment.

Keywords: exploration, exploitation, fMRI, brain signal variability, reinforcement learning, uncertainty

3.1 Introduction

Past neuroimaging work on exploration-exploitation has mostly concentrated on mapping brain regions to exploration and/or exploitation independently, highlighting a large number of vastly distributed brain regions that might be involved in one or both modes, including frontal, temporal, sensorimotor, parietal and occipital regions, as well as the cingulate cortex, cerebellum, and multiple subcortical structures, such as nucleus accumbens (ventral striatum), caudate and putamen (dorsal striatum), amygdala and thalamus (Addicott et al., 2014; Badre et al., 2012; Blanchard & Gershman, 2018; Boorman et al., 2009; Chakroun et al., 2020; Cockburn et al., 2022; Daw et al., 2006; Dombrovski et al., 2020; Hogeveen et al., 2022; Muller et al., 2019; Tardiff et al., 2021; Tomov et al., 2020). However, the neural mechanisms allowing one to flexibly switch between exploration and exploitation remain poorly understood. Previous work has highlighted the role of neural variability in providing a basis for flexibility needed to successfully adapt to environmental demands – from "computational noise" being an integral part of learning and decision-making (Findling & Wyart, 2021) to the essential role of variability for brain and behavioral development across the lifespan (Lindenberger & Lövdén, 2019). Moment-to-moment variability of brain signals provides a promising and underexplored angle to better understand the relationship between brain activity and behavior (Garrett, Samanez-Larkin, et al., 2013; see Waschke et al., 2021 for a recent review).

Uncertainty of the environment plays a crucial role in exploration-exploitation behavior; while higher levels of uncertainty invite exploration, allowing to gain information about the options and thus decrease uncertainty (at the cost of potentially forgoing the best reward), known and stable environments warrant exploitation, allowing to focus on collecting reward (at the cost of growing uncertainty about non-chosen options, if their values change over time) (Bond et al., 2021; Cohen et al., 2007; Doya, 2008; Mehlhorn et al., 2015). Uncertainty has long been hypothesized as a driving force behind brain signal variability (see Garrett, Samanez-Larkin, et al., 2013; and Waschke et al., 2021 for a detailed discussion). The variability of the BOLD signal lends itself as a potential candidate for a mechanism that underpins flexible switching between exploration and exploitation, which is needed to successfully adapt to changes in the environment. Previous studies have highlighted the link between uncertainty and brain signal variability, including BOLD signal in particular (Waschke et al., 2021). Increases in brain signal variability have been shown in tasks requiring the brain to be more attuned to the environment due to higher perceptual (Garrett et al., 2014; Orbán et al., 2016) or rule (Kosciessa et al., 2021) uncertainty, working memory load (Garrett et al., 2015; Guitart-Masip et al., 2016) and feature-richness of the perceptual input (Garrett et al., 2020). Increasing uncertainty about how many perceptual features were relevant for making a response (which could be thought of as rule uncertainty (Bach & Dolan, 2012)) led to a parametric increase in neural variability measured with EEG (Kosciessa et al., 2021). Though the relationship between brain signal variability may be more complex. As uncertainty or task demands parametrically increase, brain signal variability often plateaus (Kosciessa et al., 2021) or adopts an inverted-U shape (Garrett et al., 2014, 2015), possibly explained by exceeding available processing resources (Waschke et al., 2021).

Notably, both higher levels of BOLD signal variability and its stronger modulation with uncertainty have been associated with better task performance (Waschke et al., 2021). For example, participants who showed higher levels of BOLD variability, made fewer errors both in a task requiring cognitive flexibility (switching task) and in a task requiring cognitive stability (distractor inhibition task) (Armbruster-Genç et al., 2016). Administering amphetamine resulted in higher levels BOLD signal variability and more consistent behavioral performance in older adults performing a task with three different levels of cognitive load (though the effects were complex) (Garrett et al., 2015). Better behavioral performance in the form of higher accuracy or more stable reaction times was observed in participants who increased BOLD signal variability more in response to more uncertain stimuli (Garrett et al., 2020) or more uncertain tasks (Grady & Garrett, 2018). In a recent study, participants who showed stronger decrease of BOLD signal variability during learning (as uncertainty decreased with each learning step), also showed higher accuracy in estimating the underlying stimulus distribution, possibly indicating a more efficient belief updating process (Skowron et al., 2024).

Though brain signal variability has been brought in connection with behavioral adaptability and cognitive flexibility, it's role in the context of exploration-exploitation behavior – or even in the context of reinforcement learning in general (Waschke et al., 2021) – has not yet been examined. In the current study, we employ the ExploreExploit task during fMRI in a sample of 40 younger adults to examine (a) how changes in uncertainty during exploration and exploitation might drive BOLD signal variability, and (b) how uncertainty-driven BOLD variability could provide a neural mechanism underlying the ability to balance exploration and exploitation in a changing environment. We expected changes in BOLD signal variability to follow the direction of changes in uncertainty: variability increasing when uncertainty increased and decreasing when uncertainty decreased. In contrast to previous (Garrett et al., 2014, 2015, 2020; Kosciessa et al., 2021), which assumed different levels of uncertainty based only on task design, we use computational modeling to estimate uncertainty directly. The ability to estimate three different types of uncertainty from our computational model allowed us to examine which uncertainty type had strongest influence on brain signal variability in the context of exploration-exploitation behavior by contrasting trials in which uncertainty of each type was changing. We also expected that participants, who modulate BOLD signal variability more strongly in the direction of uncertainty change, would show better behavioral performance (exploit highest-paying bandit more often) and more flexible behavior (switch between exploration and exploitation more often). Furthermore, our study makes an important contribution to neuroimaging research on exploration-exploitation decision-making by investigating which neural systems underpin exploration and exploitation in the context of BOLD signal variability, as they may differ greatly from the brain regions found in studies based on mean BOLD signal (Garrett, Samanez-Larkin, et al., 2013).

3.2 Materials and Methods

Sample characteristics, task design, experimental procedure, and computational modeling were described in detail in Chapter 2. Here, we briefly reiterate key information and add fMRI-specific details.

3.2.1 Participants

Fifty-two young adults took part in the experiment. All participants fulfilled standard MRI-compatibility criteria (e.g., no metal in the body, no claustrophobia, no pregnancy), were right-handed, and had normal or corrected-to-normal vision. No *a priori* sample size calculations were performed, but we aimed to collect 50 complete data sets, which is considerably more than the sample sizes of comparable neuroimaging studies in the field (Blanchard & Gershman, 2018; Boer et al., 2017; Cockburn et al., 2022; Daw et al., 2006; Muller et al., 2019; Tomov et al., 2020). In addition to 5 participants excluded for the analyses of behavioral data (see Chapter 2 for details), we excluded another 7 participants because their fMRI data did not meet the criteria necessary for the intended analyses (see *Calculating IQR BOLD* section for details). In addition, 6 additional participants were excluded from analyses based on not having adequate data on sequences of 5 continuous exploitation trials (see below). The final sample thus consisted of 40 participants (and 34 participants for analyses based on a sequence of 5 continuous exploitation trials).

3.2.2 Task design and computational model

Participants performed the ExploreExploit task in the MRI scanner. On each trial, they made a decision to either explore or exploit one of the three bandits. The key feature of this task is that feedback consisted only of information (but not reward) on exploration trials and only of non-stationary reward (but not information) on exploitation trials. This, in combination with using separate response buttons for each response, allowed for the unambiguous categorization of exploration and exploitation trials during behavior (eg. Blanchard & Gershman, 2018). Categorizing trials based on a direct measure provided by participants themselves, rather than computational modeling, eliminates (or at least minimizes) issues that might obscure results of contrasting neural data between exploration and exploitation trials, such as trial categorization being inherently dependent on the type of the model or exploration being indistinguishable from random error (eg. Blanchard & Gershman, 2018).

The computational model we used (see Chapter 2) provides separate expected values for exploring and exploiting each bandit (3 exploration and 3 exploitation values). The updating process during exploration includes (1) learning for the explored bandit: new information is incorporated into the estimated reward value and uncertainty (sigma) about this value becomes smaller, and (2) forgetting for the unexplored bandits: an information leak brings the expected reward values and uncertainty (sigma) about reward back to their starting values. During exploitation, forgetting is applied to all bandits, since feedback contains no information on exploitation trials.

3.2.3 Capturing multiple types of uncertainty with computational modeling

Capitalizing on our winning computational model (see Chapter 2), we identified three uncertainty-related variables which could be used to disentangle the influence of different types of uncertainty on IQR BOLD in the context of exploration and exploitation. These variables represent prior and posterior estimation

uncertainty (sigma prior and sigma posterior), as well as choice uncertainty (**Figure 3-1**), which we specify below (cf. Bruckner et al., 2022).

Estimation uncertainty

Prior sigma – uncertainty about the estimated reward value – can be considered a proxy for the SD of a belief distribution in a Bayesian framework (Pearson et al., 2011). It represents the uncertainty that a person has about the value of a bandit *prior to* making a choice. Posterior sigma, on the other hand, can be thought of as a proxy for the SD of a posterior belief distribution, representing the uncertainty *after* a choice has been made. Notably, though prior and posterior sigma values existed for each bandit on each trial, the measure we used in our analyses represents total prior or posterior uncertainty on each trial, calculated by summing prior sigma values and posterior sigma values across all 3 bandits (cf. Chakroun et al., 2020; Tomov et al., 2020). This approach allows to examine the total level of uncertainty on a given trial and simplifies the analyses. For the sake of simplicity, we refer to total prior uncertainty and total posterior uncertainty simply as prior sigma and posterior sigma, respectively.

During exploration, information is learned for one bandit and forgotten for the other two, so the uncertainty that makes up the sum of the posterior sigma values on the current trial – and the sum of the prior sigma values on the next trial – decreases. In contrast, the reward estimates of all bandits are forgotten during exploitation, so the sum of the posterior sigma values on the current trial – and the sum of the prior sigma values on the next trial – increases.

Choice uncertainty

In addition, we utilize information entropy of the response distribution (Muller et al., 2019; Shannon, 1948) to calculate an *entropy* measure, which reflects how uncertain a choice may be (cf. Bruckner et al., 2022). Entropy was calculated as a negative sum across all bandits of the product of the response probability and its log value:

$$Entropy = -\sum_{i=1}^{n} p(x_i)\log\big(p(x_i)\big),$$

where $p(x_i)$ denotes the response probability $p(x)$ of bandit i. Of note, we calculated entropy separately for exploration and exploitation responses. On exploration trials, entropy was calculated only across exploration response probabilities (3 out of 6). Likewise, only exploitation response probabilities (another 3 out of 6) were used to calculate entropy on exploitation trials. Low entropy values indicate more deterministic choices (the value of 0 corresponding to a 100% deterministic choice), while higher entropy values indicate more uncertain choices (e.g., a value of 1.09 in the case of 3 options indicates a flat response probability distribution in which each response has a probability of 1/3).

3.2.4 Utilizing uncertainty to inform hypotheses

A key aim of the current study was to investigate the link between uncertainty and BOLD signal variability in the context of exploration-exploitation decision-making. For this purpose, we examined computational modeling results to identify the relationships between the variables (prior sigma, posterior sigma, and entropy) that represented different types of uncertainty. First, we broadly hypothesized that BOLD signal variability should increase when uncertainty increases (Grady & Garrett, 2018). To create contrasts that would allow to delineate the relationship between BOLD signal variability and different types of uncertainty, we examined the relationships between uncertainty-related variables present in our computational model. We aimed to identify such types of exploration and/or exploitation trials on which these variables could be observed to change in the same or different directions. As a result, we produced trial contrasts that could reveal (1) whether BOLD signal variability tracked level of uncertainty on exploration and exploitation trials and (2) to which uncertainty type BOLD signal variability was most robustly related. We describe the hypotheses in detail in *Results* section.

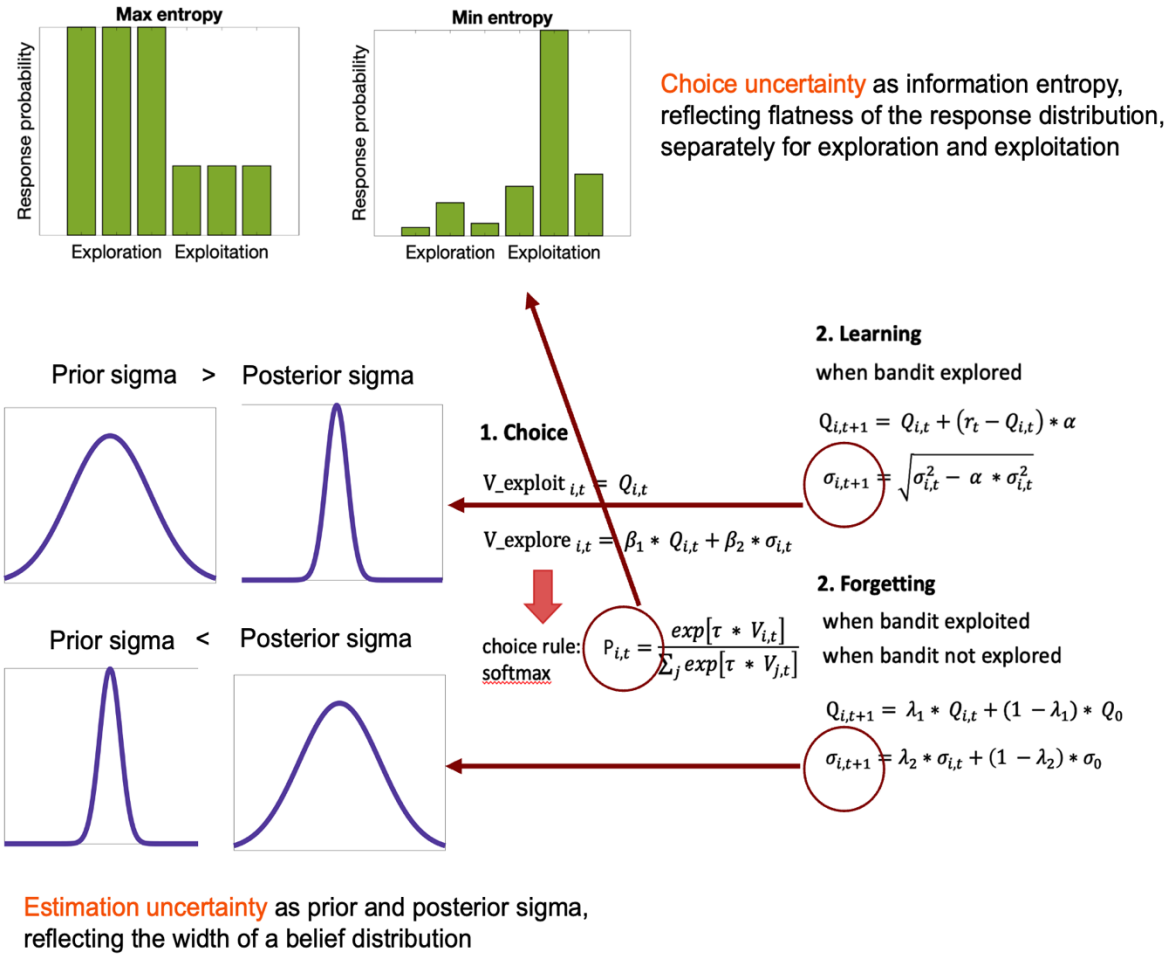


Figure 3-1. Schematic representation of choice and estimation uncertainty. *Equations refer to the computational model used in our study. If a bandit is explored, its estimation uncertainty decreases (prior sigma is larger than posterior sigma). If a bandit is not explored (or in case of exploitation), its value is forgotten, so the estimation uncertainty increases (prior sigma is smaller than posterior sigma). Information entropy was calculated separately for exploration and exploitation responses based only on exploration or exploitation response probabilities, respectively. Higher entropy values reflect a flatter response distribution and more uncertain choice; lower entropy values indicate a more deterministic choice.*

3.2.5 MRI data acquisition

The experiment was conducted with a 3T Siemens TimTrio MRI scanner (Erlangen, Germany) and a 32-channel head coil. We used a T1-weighted MPRAGE sequence (192 saggital slices, voxel size 1 x 1 x 1 mm, TR = 2500 ms, TE = 4.77 ms, FoV = 256 mm, flip angle = 7º) to collect structural MRI data, and a multiband EPI sequence (MB factor = 4, 40 transverse slices, voxel size 3 x 3 x 3 mm, TR = 645 ms, TE = 30 ms, FoV = 222 mm, flip angle = 60º) to collect functional MRI data during each task block. Subsequently, two short (5 volumes) sequences with the same parameters as the EPI sequence but opposite phase encoding directions (A>P, P>A) were collected to be used for distortion correction during data preprocessing. Recorded data were processed according to brain imaging data structure (BIDS) format guidelines (Gorgolewski et al., 2016).

3.2.6 MRI data preprocessing

Collected images were converted from DICOM to NIfTI format using HeuDiConv (version 0.9.0, https://heudiconv.readthedocs.io/) and ReproIn heuristic (version 0.6.0, Castello et al., 2020). Brain extraction of T1-weighted images was done with ANTs (Avants et al., 2014). Functional MRI data was preprocessed using FSL (version 5.0.11, Smith et al., 2004). The first 12 volumes of each functional image were discarded to ensure steady-state tissue magnetization. Field maps needed to correct distortions caused by B0 field inhomogeneities were created using FSL topup (Andersson et al., 2003) and short EPI sequences recorded with opposite phase-encoding directions. Preprocessing of functional data was done for each run separately using FSL FEAT (Woolrich et al., 2001) and included motion correction, smoothing (7 mm kernel), and unwarping with previously prepared field maps. In addition, the data was detrended (at a 3rd order polynomial) using spm_detrend function from SPM12 (www.fil.ion.ucl.ac.uk/spm/software/spm12/) and filtered with an 8th-order Butterworth filter (high-pass cut-off: 0.01 Hz) implemented in MATLAB. Functional images were registered with ANTs first to the individual T1-weighted structural image (using 6 DOF) and then to a standard 3 mm MNI152 template (using linear rigid-body transformation). To further reduce the influence of noise in the data (Garrett et al., 2010, 2015; Kosciessa et al., 2021), we performed spatial independent component analysis (ICA) using FSL MELODIC (Beckmann & Smith, 2004). Components were manually classified as signal or noise (see Garrett et al., 2014; Kosciessa et al., 2021 for details on classification criteria) and those flagged as noise were regressed out of the data via the regfilt function in FSL. The data for each run was demeaned and masked with a grey matter mask based on MNI152 template grey matter tissue prior (at probability of 0.25).

3.2.7 MRI data analyses

Calculating IQR BOLD

MRI data analyses were performed using MATLAB 2020a (https://www.mathworks.com) and custom scripts based on materials from (Kloosterman & Garrett, n.d.; Kosciessa et al., 2021). In the current

study, we opt for the interquartile range (IQR) of the BOLD signal distribution (IQR BOLD) as a measure to express brain signal variability. Like standard deviation (SD), IQR is a measure of dispersion of a distribution, but it is based on data between the 25th and the 75th percentiles (Dekking et al., 2005), and is thus less susceptible to outliers than SD. In light of a limited number of trials, especially in case of exploration, we opted for IQR as a measure of BOLD signal variability. Exploration and exploitation trials were grouped into conditions (details provided in *Results*), for which we then calculated IQR BOLD to be used in subsequent analyses.

To reliably calculate the IQR, we aimed to have at least 20 data points (i.e. TRs) per condition. This corresponded to a minimum of 4 trials per condition, since each trial provided a minimum of ca. 6 TRs. Participants whose data did not have at least 4 trials in a certain condition were excluded from analyses involving that condition. In particular, conditions based on a sequence of consecutive exploration/exploitation trials were affected. Seven participants did not have enough trials in each position in a sequence of 3 continuous exploration or exploitation trials to calculate BOLD variability, which would prevent them from being included in some task PLS and all behavioral PLS analyses.

To calculate IQR BOLD, we first used an in-house version of the Variability Toolbox (https://github.com/LNDG/vartbx) implemented in SPM12 to compute voxel-wise GLM beta estimates for each TR (Haynes, 2015). For this purpose, the data for all runs of each participant were concatenated. Each TR was specified as a regressor and convolved with the canonical HRF function, resulting in a 4D NIfTI file containing whole-brain beta maps *for each time point*. Next, we identified trials that composed each condition of interest and the TRs that fell into these trials. For each condition, we then calculated IQR over the condition-specific beta estimates, within voxel. This approach provides a number of advantages over calculating one beta estimate per trial for specific contrasts, as was done in past work (Garrett et al., 2010, 2014). First, trials can be post-hoc grouped into conditions without the need to rerun the first-level GLM analysis to obtain beta estimates for each condition. In addition, a variability-based analysis benefits from an increased number of data points produced by obtaining a beta estimate for each TR instead of each trial (cf. LSS method (Arco et al., 2018; Mumford et al., 2014)).

Multivariate PLS analysis

To investigate the role of IQR BOLD as a potential mechanism providing flexibility for switching between exploration and exploitation, we then used multivariate PLS analyses (McIntosh et al., 1996). In the following, we present a summary of the key procedures of the PLS analysis (see Krishnan et al. (2011) for a detailed description of the method). Task PLS is used to examine the average effects of the task design on brain activity, while Behavioral PLS reveals how individual differences in neural activity are linked to individual differences in behavior or group characteristics (Krishnan et al., 2011). Singular value decomposition (SVD) is used to decompose a design (task PLS) or behavior (behavior PLS) x brain data (voxels) matrix into 3 matrices: a matrix of singular vectors representing task design (task PLS) or behavior (behavior PLS) weights, another matrix of singular vectors representing brain data weights, and a diagonal matrix of singular values (or so-called latent variables, LVs). The LV matrix that

represents design or behavior variables is called “design” or “behavior scores”, while the matrix that represents the brain data is called “brain scores” (Krishnan et al., 2011).

For both Task and Behavioral PLS, statistical significance of singular values associated with LVs was assessed using 1000 permutations. To this end, the rows of the original matrix containing brain data (voxels) were shuffled, while behavior or design variables remained unchanged (Krishnan et al., 2011; McIntosh et al., 1996). SVD was then repeated to produce a new set of singular values. A sampling distribution of singular values under the null hypothesis was based on results of 1000 permutation tests. If the p-value associated with the LV was < 0.05, it was considered statistically significant.

Further, a bootstrapping procedure with 1000 resampling steps was used to identify a set of voxels (spatial pattern) that reliably related to task conditions (task PLS) or behavior variables (behavior PLS) within an LV. To this end, participants were resampled 1000x with replacement (Efron & Tibshirani, 1986; Krishnan et al., 2011). Each voxel weight (from the original data) was then subsequently divided by its bootstrapped standard error, resulting in bootstrap ratios (BSRs) that are considered a type of non-parametric z-score (Efron & Tibshirani, 1986; McIntosh et al., 1996).

We first employed Task PLS to investigate the relationship between uncertainty and IQR BOLD during exploration and exploitation. Moreover, we aimed to disentangle the influence of different types of uncertainty (estimation uncertainty: prior sigma, posterior sigma; choice uncertainty: entropy) on IQR BOLD in exploration and exploitation based on relationships between uncertainty-related variables in our computational modeling results (see Chapter 2 and section X above, *Utilizing uncertainty to inform hypotheses*).

We employed Behavioral PLS to examine whether modulation of IQR BOLD with uncertainty in exploration and exploitation could be a potential mechanism underlying task flexibility and optimal performance. Specifically, we investigated the link between voxel-wise modulation of IQR BOLD (from trial 1 to 3) in exploration and exploitation and (a) optimal choice percentage (level of performance) and (b) switch percentage (task flexibility). Optimal choice percentage was defined as how often participants exploited the best-paying bandit in relation to all exploitation trials, while switch percentage reflected the degree of changing from exploration to exploitation mode (and vice versa) relative to all trials.

We used (mixed) linear models (as implemented in lme4 package (Bates et al., 2014) in R (version 4.2.2, R Core Team, 2022) to provide further statistical support to some of the PLS analyses. Effect sizes were quantified using the partR2 package (Stoffel et al., 2021) to calculate semi-partial R^2 (R^2) for all predictors in the model (Nakagawa & Schielzeth, 2013).

Reporting results

We thresholded BSRs at +/-3, unless otherwise indicated. We report brain region results with a minimum cluster size of 25 voxels and a minimum distance of 10 mm between the clusters. Harvard-Oxford cortical and subcortical atlases (https://cma.mgh.harvard.edu/) in FSLeyes (McCarthy, 2023) were used

to automatically label the brain regions at the peak coordinates of the cluster. In addition, we used the Oxford thalamic connectivity probability atlas (Behrens et al., 2003) to visualize connections of different thalamic regions. For one cluster of above-threshold BSRs, peak coordinates did not provide a label in either a cortical or subcortical atlas, so we visually inspected the cluster and identified the region label based on the near-peak voxels that could be assigned to a specific brain region.

3.3 Results

3.3.1 Elucidating the relationships between prior estimation uncertainty, posterior estimation uncertainty, and choice entropy

To identify which contrasts of exploration and/or exploitation trials could be used to examine the relationship between IQR BOLD and uncertainty in exploration-exploitation decision-making, we first examined how different types of uncertainty related to each other. We then used these relationships as a basis for creating hypotheses for IQR BOLD analyses (presented in the next section).

All uncertainty types show the same direction of change in a sequence of trials

Our computational modeling results (Chapter 2) revealed that values of all uncertainty-related variables changed in the same manner within a sequence of exploration or exploitation trials; uncertainty decreased with each consecutive exploration trial due to learning new information, and increased with each consecutive exploitation trial due to forgetting (**Figure 3-2A**). In addition, the sequence length of consecutive trials differed in exploration and exploitation, with exploration sequences being mostly composed of a maximum of three trials, while exploitation sequences were often longer. Therefore, examining the IQR BOLD across the totality of trials in each position in a sequence could show (1) whether IQR BOLD generally tracks the uncertainty levels (increases during exploitation and decreases during exploration, like uncertainty does), and (2) whether the increase of IQR BOLD plateaus or takes on an inverted U-shape as uncertainty keeps growing during longer sequences of exploitation trials, possibly reflecting that the task has become too demanding (Garrett et al., 2014, 2015).

Negative relationship between choice and estimation uncertainty on exploitation trials

First, we searched for a way to separate the association between IQR BOLD and sigma (estimation uncertainty) vs. entropy (choice uncertainty) in exploration and exploitation. For most participants, within-subject correlations of both prior and posterior sigma with entropy revealed a positive relationship between both sigma variables and entropy in exploration trials, and a negative relationship in exploitation trials (**Figure 3-2B**). A deeper examination of the relationship between sigma and entropy in exploitation revealed that the negative correlation was driven by switch trials (the first trial of an exploitation sequence) and arose because when the difference between bandits was large enough to make the choice rather deterministic (lower entropy), participants tended to explore in rather short sequences thus not allowing sigma to decrease much (higher sigma). Consequently, both sigma and

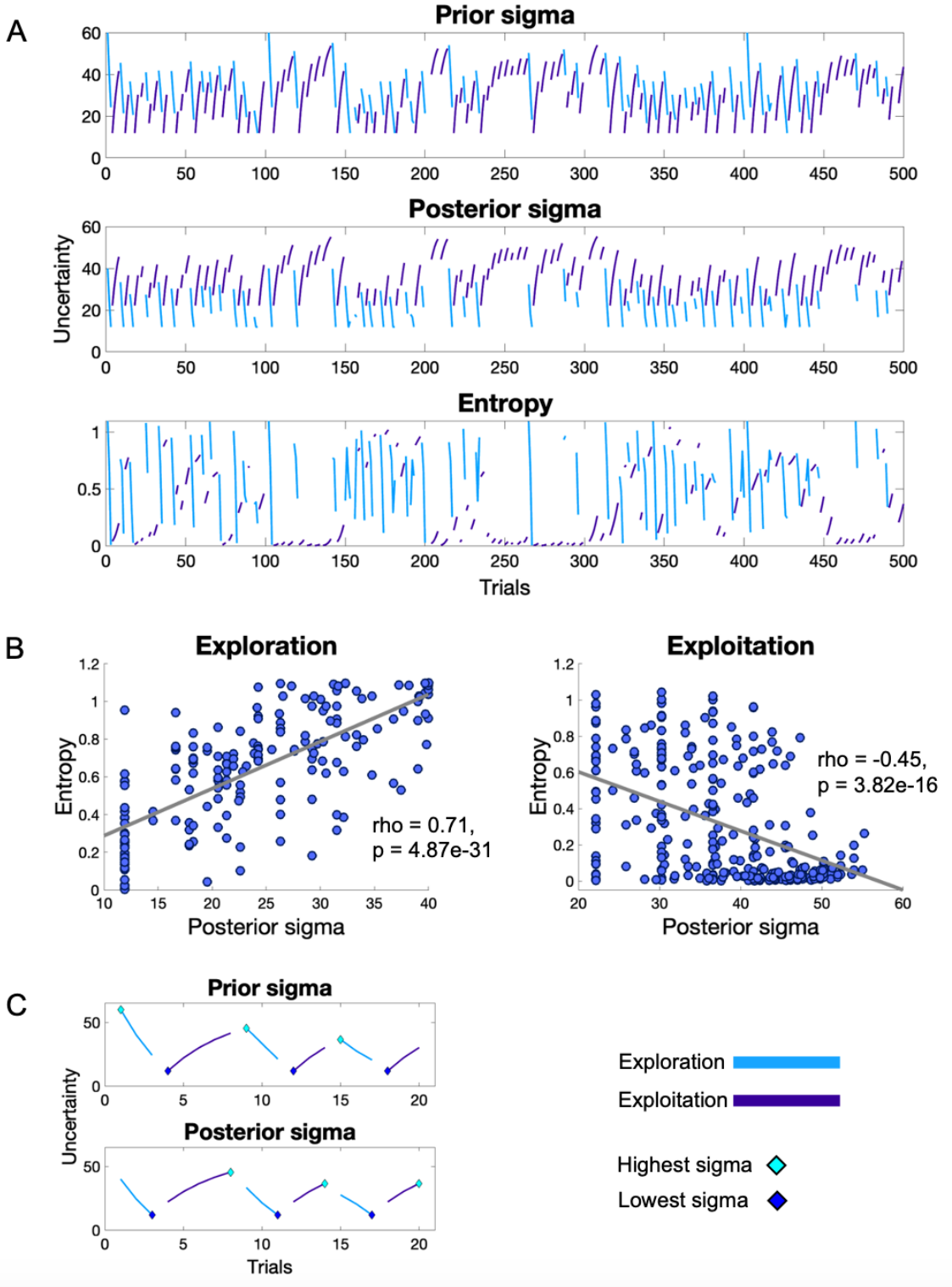


Figure 3-2. Uncertainty measures in the data of an exemplary subject. *A – prior sigma (top), posterior sigma (middle) and entropy (bottom) decrease in exploration and increase in exploitation trial sequences. Only sequences with 2 and more trials are plotted here. B – positive correlation between posterior sigma and entropy in exploration (left) and negative correlation between posterior sigma and entropy in exploitation (right). Correlation expressed as Spearman's correlation coefficient. C – for prior sigma (top panel), the highest level of uncertainty is observed on explore switch trials and lowest level in exploit switch trials; for posterior sigma (bottom panel), the highest uncertainty level is observed on the last exploit trial (noswitch) and lowest uncertainty level was observed on the last explore trial (noswitch). First 20 trials (same as in panel A) were selected for visualization.*

entropy would drive IQR BOLD in the same direction in exploration, but in different directions in exploitation trials.

Switch trials differentiate between prior and posterior estimation uncertainty

Next, we looked for an analysis that would allow us to separate the link between IQR BOLD and prior vs. posterior sigma variables. Contrasting switch (first trial in a sequence) and noswitch (other trials) trials in exploration vs exploitation could provide such separation. This relationship arises because (1) posterior sigma on one trial becomes prior sigma on the next, and (2) uncertainty in a sequence of trials decreases to the minimum on the last exploration trial before a switch to exploitation (reflecting that information gained on each exploration trial made the beliefs about the reward structure more precise) and increase to the maximum on the last exploitation trial before a switch to exploitation (reflecting that receiving no information as feedback on exploitation trials makes beliefs about the reward structure more uncertain with each exploitation trial). For example, in case of prior sigma (**Figure 3-2C**), it is high on the first trial in an exploration sequence (switch trial). As exploration goes on and more is learned about the reward structure, prior sigma on each successive exploration trial in a sequence will be smaller, resulting in the smallest value of prior sigma on the last trial of the exploration sequence (noswitch trial). At this point the participant has a precise enough idea of the reward structure and switches to exploitation. The posterior sigma from the last exploration trial becomes prior sigma of the first exploitation trial in a sequence (a switch trial) and is, therefore, lower than the values of prior sigma on subsequent exploitation trials in the sequence (which increase reflecting the forgetting process). The last trial of the exploitation sequence (noswitch trial) will thus have the highest value of prior sigma, at which point uncertainty becomes too high so that participants decide to explore, so that the posterior sigma from the last exploitation trial becomes prior sigma on the first exploration trial. The opposite is the case for posterior sigma (**Figure 3-2C**). Consequently, posterior sigma was lowest in explore noswitch and highest in exploit noswitch category, while both switch categories had more similar values in between. Conversely, prior sigma was lowest on exploit switch trials and highest on explore switch trials, while noswitch trials had more similar values in between.

3.3.2 IQR BOLD tracks uncertainty in exploration and exploitation

First, we utilized a multivariate task PLS analysis to examine whether the level of IQR BOLD in a exploration or exploitation trials reflects the direction of uncertainty change. Exploration sequences of more than 3 trials were rare, we limited the number of trials in a sequence to 3 for both exploration and exploitation. This analysis thus included 6 conditions based on combinations of exploration/exploitation and trial position (1, 2, 3) in a sequence of 3 consecutive trials of the same type. Importantly, only trials that belonged to a sequence of at least 3 trials were included in analyses involving these conditions (trials that belonged to sequences of 1 or 2 trials were thus excluded).

If IQR BOLD is related to uncertainty in exploration and exploitation, we should see parametric change of IQR BOLD that reflects the direction of uncertainty change in a sequence of exploration and exploitation trials, with IQR BOLD increasing – in line with uncertainty – from trial 1 to trial 3 in

exploitation, and decreasing from trial 1 to trial 3 in exploration (**Figure 3-3A**). Next, we used brain scores (reflecting the brain-dependent latent score from the PLS results) for each participant as a dependent variable in a mixed model, with behavior category (exploration, exploitation) and trial position (1, 2, 3) in a sequence as independent variables, and participant ID as a random intercept. The presence of an interaction would provide further statistical support for a parametric increase of IQR BOLD in exploitation and a decrease in exploration.

In line with our expectation, we found a parametric effect in our PLS results expressing a decrease of IQR BOLD from trial 1 to 3 in exploration and an increase from trial 1 to 3 in exploitation (LV1: permuted $p < 0.001$, **Figure 3-4A**). IQR BOLD levels in trial 1 were most similar between exploration and exploitation, with increasing separation in later trials. Mixed modeling confirmed a significant interaction between behavior category (exploration, exploitation) and trial number (1, 2, 3) ($t(197) = -5.87$, $p = 1.82e-08$, $R^2 = 0.05$).

Strong effects (BSR threshold = +/-3) were observed across a large proportion of the brain (**Figure 3-4B**, see **Table 3-S1** for cluster peak coordinates). Notably, while there was a small separate cluster in the bilateral thalamus, another group of voxels from a larger cluster was clearly localized to the left thalamus, spanning mostly thalamic regions with structural connections to prefrontal and temporal brain areas (**Figure 3-4C**). To better assess the spatial distribution of the effects, we increased the BSR threshold to 5 (**Figure 3-S1**) to better allow localization of brain regions showing the strongest effect. Doing so revealed multiple clusters in frontal (frontal pole, inferior and superior frontal gyri, frontal operculum), lateral occipital, and temporal (3 temporal gyri, temporal fusiform cortex) regions, as well as in the paracingulate gyrus, precuneus, and the hippocampus (**Table 3-S1**).

To dig deeper into how IQR BOLD might track uncertainty in a sequence of consecutive trials in our task, we next analyzed consecutive exploitation trials in positions 1 to 5, thus allowing both linear and nonlinear effects to be estimated. As with our 3-trial sequence analyses above, only trials that belonged to a sequence of at least 5 consecutive exploitation trials were included. This analysis was only possible for exploitation, since sequences with more than three trials in exploration were rare. If IQR BOLD showed a quadratic effect (**Figure 3-3B**) in a sequence of 5 consecutive exploitation trials, this could indicate that, as uncertainty and the corresponding mental load increase further during exploitation, IQR BOLD tracks them only up to a certain point after which further linear increases become too resource intensive. To fully assess whether IQR BOLD showed a quadratic effect, we additionally ran a mixed model with brain scores as a dependent variable, a linear and a quadratic effect of the trial position (1, 2, 3, 4, 5) in a sequence as independent variables, and participant ID as a random intercept.

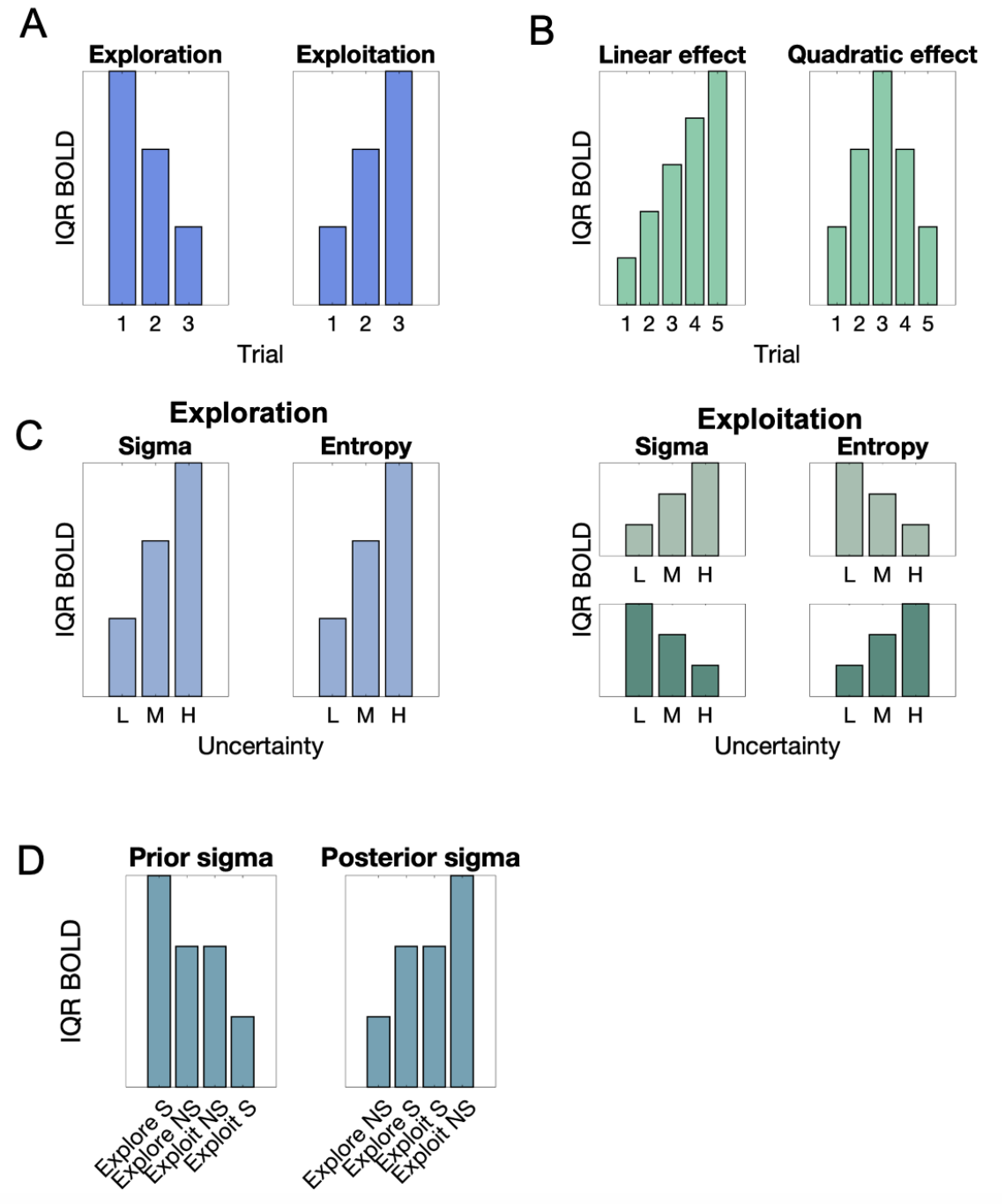


Figure 3-3. Schematic representation of hypotheses for Task PLS analyses. *A – parametric decrease of IQR BOLD in a sequence of 3 exploration trials and increase in a sequence of 3 exploitation trials. B – parametric increase of IQR BOLD in a sequence of 5 exploitation trials (linear effect); alternative: inverted-U shape (quadratic effect). C – Left: in exploration, IQR BOLD should increase as the level of both sigma and entropy increases. Right: in exploitation, the level of IQR BOLD increases with either sigma or entropy levels (L, M, H – low, medium, high). D – IQR BOLD is highest on explore switch trials, lowest on exploit switch trials and has middle values on noswitch trials of both types; alternative: IQR BOLD is highest on exploit noswitch trials, lowest on explore noswitch trials and has middle values on switch trials of both types.*

Task PLS results revealed that IQR BOLD level increased from trial 1 to 3 and then stagnated from trial 3 to 5 (LV1: permuted p < 0.001, **Figure 3-S2**). Brain regions observed in this analysis included the posterior cingulate and lateral occipital cortex, as well as frontal (frontal pole, orbitofrontal cortex, inferior frontal gyrus) and temporal (middle and superior temporal gyri) regions, and in the hippocampus (**Table 3-S2**). Mixed modeling confirmed unique linear (t(134) = 4.88, p = 2.97e-06, R^2 = 0.03) and quadratic (t(134) = -3.68, p = 0.0003, R^2 = 0.02) effects of trial.

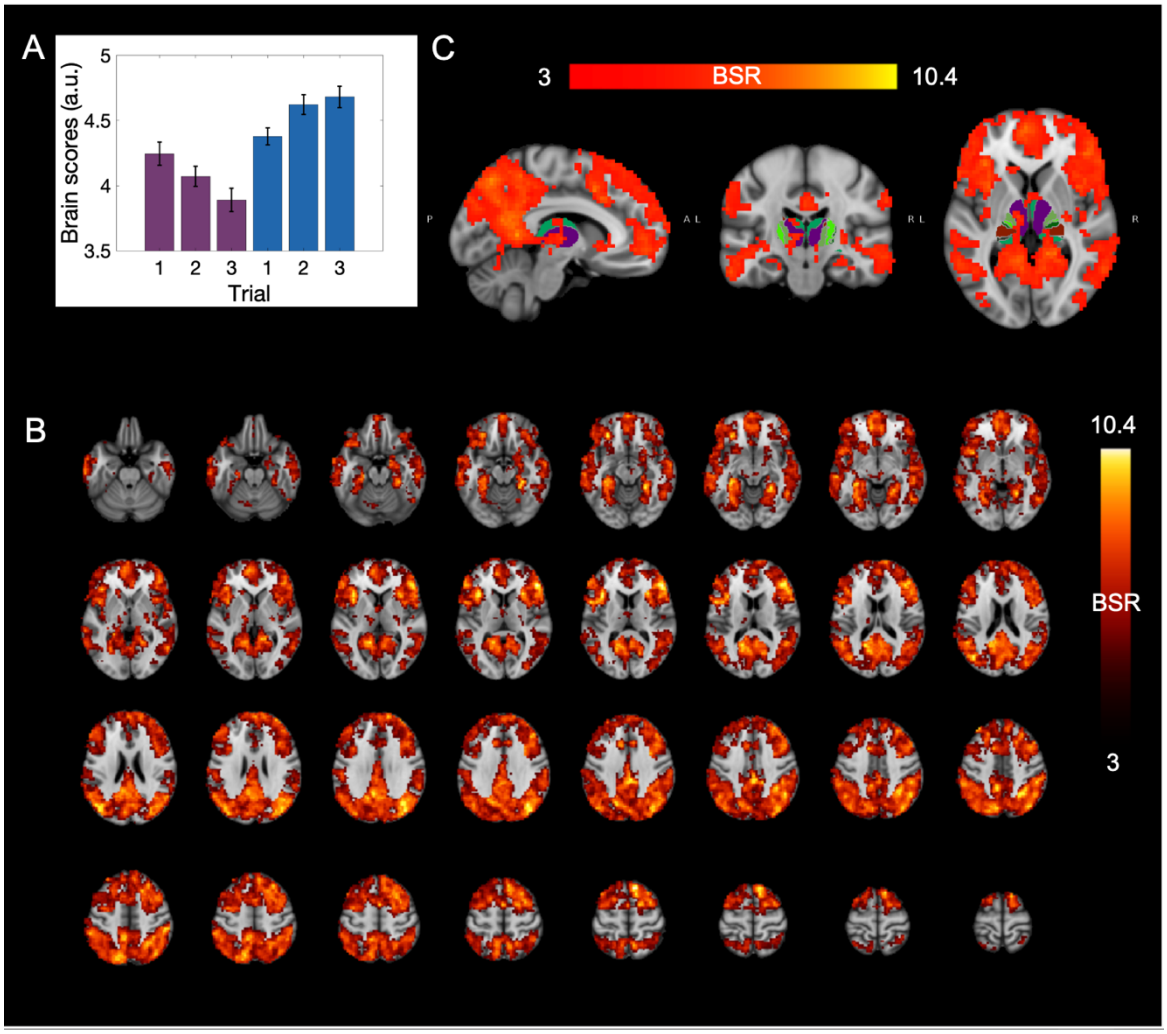


Figure 3-4. Results of task PLS analysis with IQR BOLD sequences of 3 exploration and exploitation trials thresholded at BSR +/-3. *A – IQR BOLD (expressed as brain scores) levels in trials 1, 2, 3 in exploration (purple) and exploitation (blue). Error bars – SEM. B – axial brain view. MNI coordinates of the first slice: z = -17. Each next slice increases z coordinate in increments of 3. BSR – bootstrap ratio. C – sagittal (left), coronal (middle) and axial (right) slices (MNI coordinates x, y, z: -7, -19, 2) showing thalamus activation, localized mostly to parts of the thalamus connected to prefrontal (purple) and temporal (turquoise) regions, according to the Oxford thalamic connectivity probability atlas. L, R – left, right. A, P – anterior, posterior.*

3.3.3 IQR BOLD changes in the direction of estimation uncertainty rather than choice uncertainty during exploitation

We then examined whether IQR BOLD was more strongly related to estimation or choice uncertainty by testing the relationship of IQR BOLD to both sigma (estimation uncertainty) and entropy (choice uncertainty). Since uncertainty estimates were continuous in nature, we grouped them into low, medium, and high levels. Conditions included in this analysis were based on combinations of exploration/exploitation and three levels of posterior sigma/entropy. Since prior and posterior sigma showed similar associations to entropy and posterior sigma plays a prominent role in our results, we present posterior sigma for this condition to make our results more succinct. The levels of sigma and entropy were calculated individually for each participant to yield approximately the same number of trials in each low/med/high category.

Since sigma and entropy showed a positive relationship in exploration trials, we expected IQR BOLD to show an increase from low to high values of both sigma and entropy in exploration trials. On the other hand, since sigma and entropy showed a negative relationship in exploitation trials, we expected IQR BOLD to be highest when either sigma is high and entropy is low (suggesting that IQR BOLD follows the direction of estimation uncertainty) or when entropy is high and sigma low (suggesting that IQR BOLD follows choice uncertainty) (**Figure 3-3C**).

As expected, IQR BOLD in exploration was highest when both sigma and entropy were highest, and decreased as both sigma and entropy decreased (LV1: permuted $p < 0.001$, **Figure 3-S3**). The effect was present in the posterior cingulate cortex, frontal regions (inferior and superior frontal gyri, orbitofrontal cortex, and multiple clusters in the frontal pole), temporal (middle and superior temporal gyri, temporal pole), parietal (supramarginal gyrus) and occipital (lateral occipital cortex, lingual gyrus) areas, as well as subcortical regions (caudate and putamen; **Table 3-S3**). For exploitation trials, IQR BOLD was highest when sigma was highest and entropy lowest (LV1: permuted $p < 0.001$, **Figure 3-5**). Further, IQR BOLD decreased with decreasing levels of sigma/increasing levels of entropy. The effect was observed in the anterior and posterior cingulate cortices, as well as in the frontal regions (inferior gyrus and multiple clusters in the frontal pole), in multiple temporal regions (middle and superior temporal gyri, temporal pole) as well as in the hippocampus, in the lateral occipital cortex, and the supramarginal gyrus (**Table 3-S4**).

3.3.4 IQR BOLD changes in the direction of posterior estimation uncertainty rather than prior estimation uncertainty

Examining switch and noswitch trials in exploration and exploitation allows us to disentangle whether IQR BOLD rather moves in the direction of prior sigma (reflecting estimation uncertainty that exists prior to making a choice) or in the direction of posterior sigma (reflecting estimation uncertainty that results from making a choice). Conditions included in this analysis were based on combinations of

exploration/exploitation and switch/noswitch categories (switch trials being the first trial in a sequence of continuous exploration or exploitation trials, noswitch trials – all other trials).

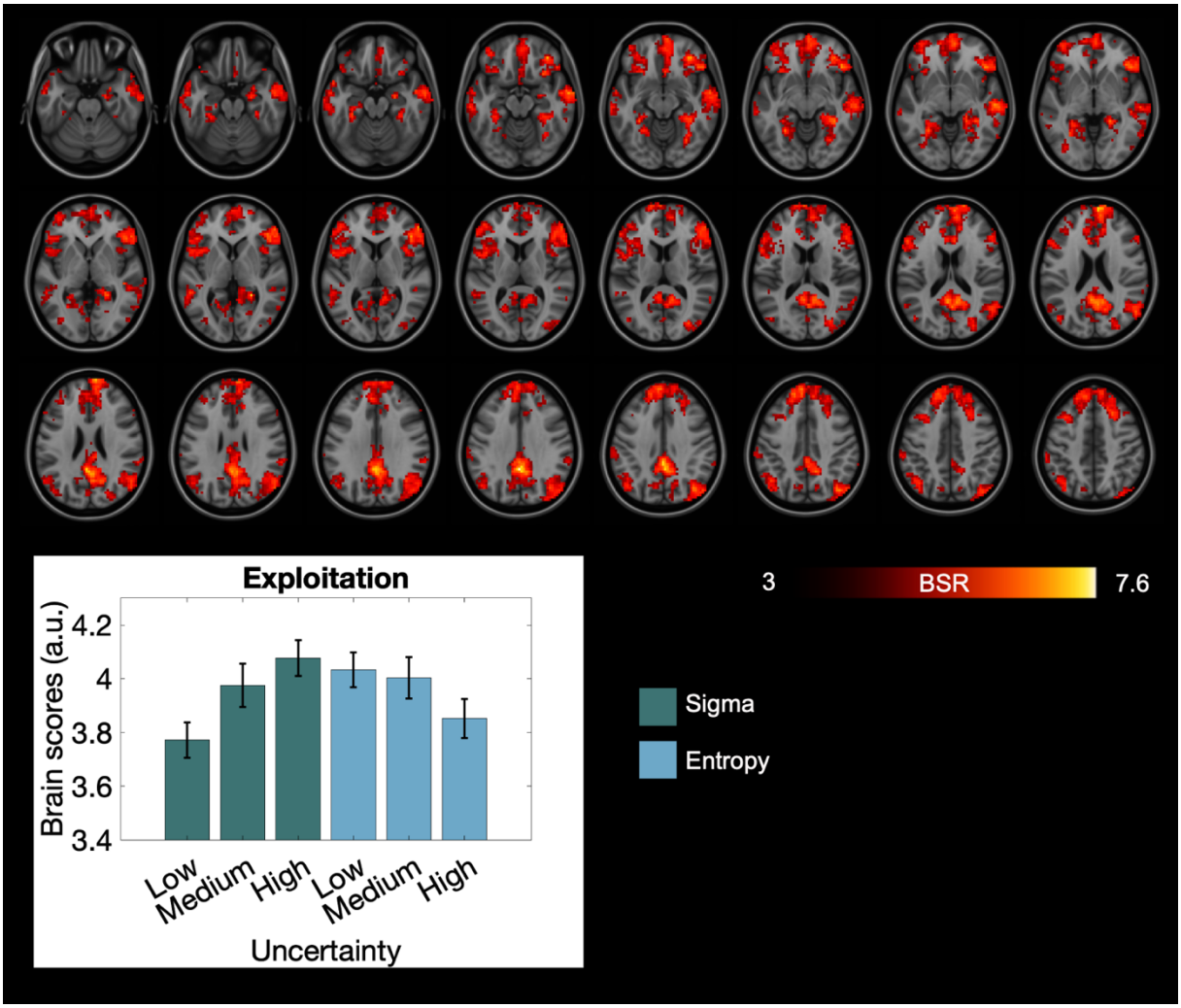


Figure 3-5. Results of task PLS analysis with IQR BOLD at low, medium and high levels of sigma and entropy in exploitation trials. *Top – axial brain view. MNI coordinates of the first slice: z = -24. Each next slice increases z coordinate in increments of 3. BSR – bootstrap ratio. Bottom – IQR BOLD levels (expressed as brain scores) at low, medium and high levels of sigma (green) and entropy (blue). Error bars – SEM.*

If IQR BOLD is more strongly related to prior sigma, we would expect IQR BOLD to be highest in the explore switch condition (when prior sigma is highest) and lowest in the exploit switch condition (when prior sigma is lowest), with noswitch conditions being in the middle (**Figure 3-3D**). On the other hand, if IQR BOLD rather reflects posterior sigma, we would expect IQR BOLD to be highest in exploit noswitch condition (highest posterior sigma) and lowest in explore noswitch condition (lowest posterior sigma), with switch trials being in the middle.

We found that explore and exploit trials could be reliably separated from each other, with IQR BOLD being highest in the exploit noswitch and lowest in the explore noswitch condition (LV1: permuted p <

0.001, **Figure 3-6**), corresponding to the level of posterior estimation uncertainty. The spatial distribution of the effect covered large areas of the brain when BSR threshold was set to +/-3. Of note, only a small, single cluster (the precentral gyrus) showed the opposite effect (higher IQR BOLD in exploration than in exploitation; highest IQR BOLD in explore noswitch and lowest in exploit noswitch). At an even more conservative threshold of BSR +-6 (**Figure 3-S4**), cluster peaks most related to the effect could be localized to the middle frontal gyrus (dlPFC), posterior cingulate cortex, angular gyrus, lingual gyrus and lateral occipital cortex, as well as middle temporal gyrus and the hippocampus (**Table 3-S5**).

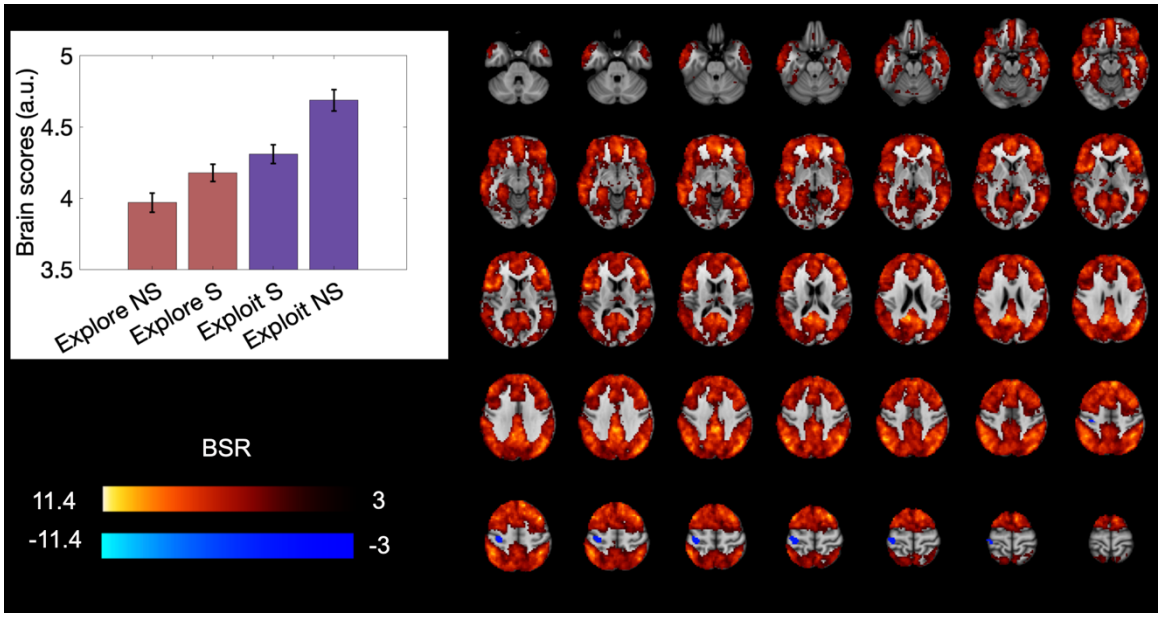


Figure 3-6. Results of Task PLS analysis with IQR BOLD in switch and noswitch exploration and exploitation trials thresholded at BSR +/-3. *Left – IQR BOLD levels (expressed as brain scores). Error bars – SEM. NS, S – noswitch, switch. Right – axial brain view. MNI coordinates of the first slice: z = -32. Each next slice increases z coordinate in increments of 3. BSR – bootstrap ratio.*

3.3.5 Higher optimal choice percentage is associated with higher level of IQR BOLD

Next, we examined how IQR BOLD might be differentially related to behavior in exploration and exploitation. To do so, we first estimated IQR BOLD level for each condition of interest, and then also voxel-wise beta weights (slopes) for IQR BOLD *change* across sequences of 3 consecutive exploration or exploitation trials (level and change analyses were performed separately). We expected higher level of IQR BOLD and stronger modulation of IQR BOLD in the direction of uncertainty to be positively related to behavioral performance in the form of optimal choice.

There was no significant association between IQR BOLD modulation in sequences of 3 consecutive trials and optimal choice percentage on the latent level (LV1: permuted p = 0.42). Any further interpretation should thus be treated with caution. The direction of the relationship suggested higher optimal choice percentage (better performance) was related to a stronger IQR BOLD decrease in

exploration, but, in contrast to our expectations, it was related to a more modest increase of IQR BOLD in exploitation (**Figure 3-S5**, **Table 3-S6**).

To better understand the role of IQR BOLD for performance in our task, we ran a behavior PLS analysis with optimal choice percentage and IQR BOLD level (as opposed to modulation) on trials 1, 2, and 3 separately for exploration and exploitation. For both exploration (LV1: permuted $p < 0.001$, **Figure 3-7**) and exploitation (LV1: permuted $p < 0.001$, **Figure 3-S6**), higher optimal choice percentage was associated with higher level of BOLD signal variability across the totality of trials in each condition. In terms of the spatial pattern, these results differed from those observed in the Behavior PLS with IQR BOLD modulation. The analysis with IQR BOLD level in exploration showed significant effects in the middle temporal gyrus, inferior frontal gyrus and frontal pole, as well as in the angular gyrus (**Table 3-S7**). The brain pattern in exploitation was similar, though more frontally distributed, and included the inferior frontal gyrus, frontal pole and orbitofrontal cortex, as well as the angular gyrus and precentral gyrus (**Table 3-S8**).

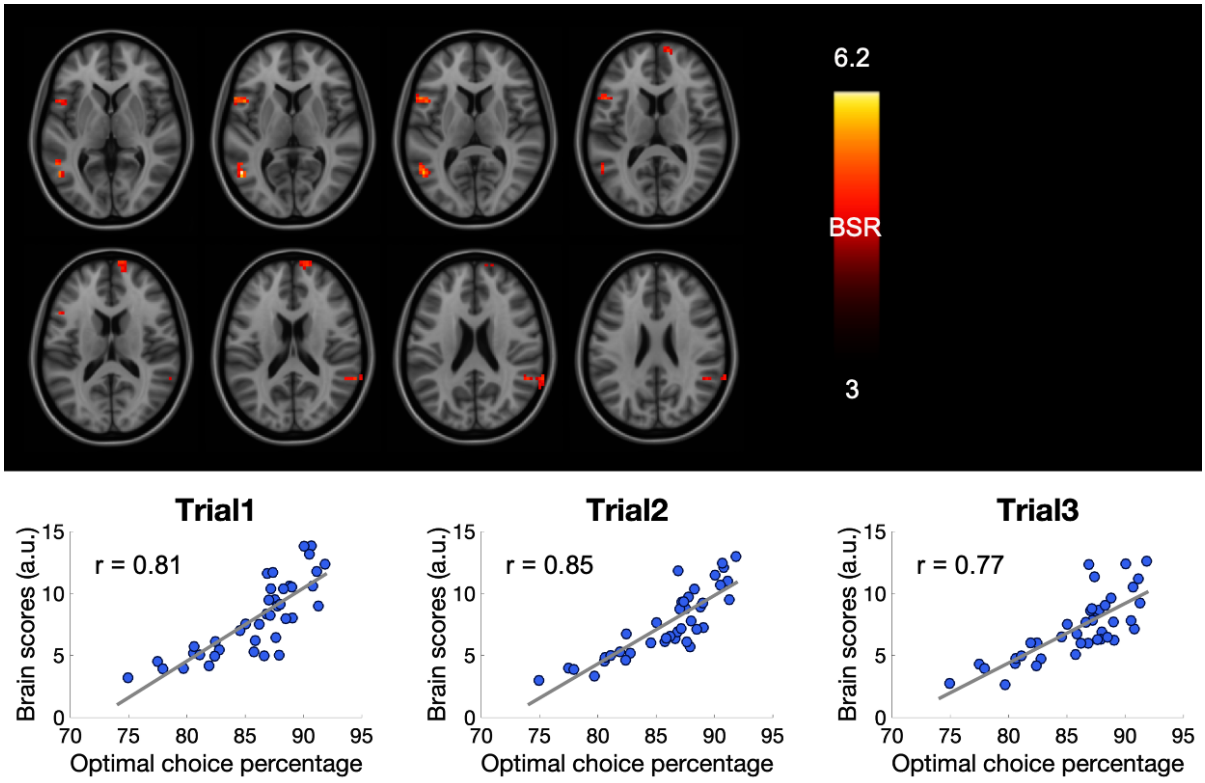


Figure 3-7. Results of behavior PLS analysis with level of IQR BOLD in exploration trials 1, 2, 3 and optimal choice percentage. *Top panel – axial brain view. MNI coordinates of the first slice: z = 3. Each next slice increases z coordinate in increments of 3. BSR – bootstrap ratio. Bottom panel – correlation of IQR BOLD level (expressed as brain scores) with optimal choice percentage. Bootstrapped CI for correlations: trial 1 – [0.73, 0.91], trial 2 – [0.76,0.91], trial 3 – [0.71, 0.90].*

3.3.6 Stronger modulation of IQR BOLD in the direction of uncertainty change reflects higher switch percentage in exploration and longer periods of staying in the same mode in exploitation

We further examined the link between IQR BOLD modulation in exploration and exploitation and switch percentage as a measure of flexibility. In a constantly changing environment (as is the case in our task), it is important to constantly update information about which bandit currently provides a good payout. Not changing between exploration and exploitation modes often enough (low switch percentage) can thus be detrimental for performance, since a change of the best-paying bandit can easily be missed. We therefore expected higher levels of switch percentage to correlate positively with stronger modulation of IQR BOLD – a decrease of IQR BOLD in exploration and an increase in exploitation.

Behavior PLS analysis showed a significant relationship between IQR BOLD modulation in both exploration and exploitation and switch percentage (LV1: permuted p = 0.005, **Figure 3-8**). As expected, participants who switched more, decreased IQR BOLD more in exploration. However, contrary to our expectations, IQR BOLD increased more in those who switched *less* in exploitation. Spatially, the effects were located in multiple relatively small clusters, including the posterior cingulate cortex, frontal pole, middle and superior frontal gyri, the insula, as well as a number of occipital and parietal brain regions (**Table 3-S9**).

We then performed several additional analyses to further probe why a more modest – and not stronger – increase of IQR BOLD in exploitation is related to better and more flexible performance.

First, we examined the relationships between behavioral variables more closely. Switch percentage was determined by how long participants remained in one mode and correlated negatively with the median continuous exploration sequence length (r = -0.45, p = 0.003, **Figure 3-S7**) and median continuous exploitation sequence length (r = -0.86, p = 4.08e-13). Despite the latter relationship being stronger, a regression analysis revealed that exploration sequence length had a unique explanatory effect on the switch percentage (t(37) = -3.10, p = 0.003, R^2 = 0.05) in addition to the effect of exploitation sequence length (t(37) = -10.66, p = 7.67e-13, R^2 = 0.60).

3.3.7 A more variable brain system underpins switching out of exploitation

Further, we tested whether the lack of IQR BOLD modulation in exploitation in participants who switch more could be related to displaying generally higher levels of BOLD signal variability. For this purpose, we examined the relationship between the level of IQR BOLD in trials 1, 2, and 3 of exploitation sequences and switch percentage in a subsequent Behavioral PLS analysis. Higher switch percentage was associated with higher IQR BOLD level (LV1: permuted p < 0.001, **Figure 3-9**) in the posterior cingulate and insular cortices, as well as in the precentral and angular gyri (**Table 3-S11**). Topographically, these effects didn't match the effects observed in the PLS analysis with IQR BOLD

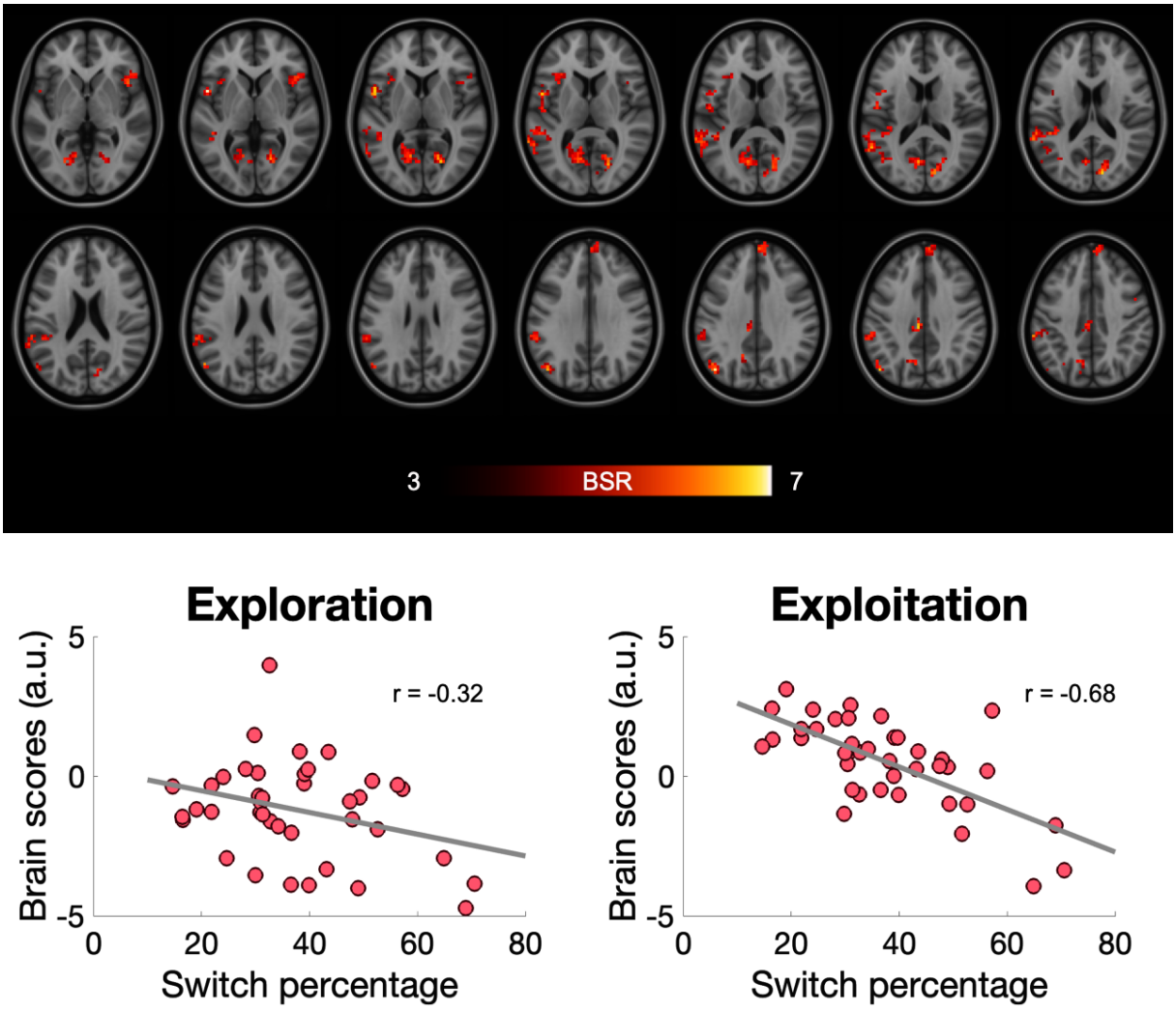


Figure 3-8. Results of behavior PLS analysis with modulation of IQR BOLD in the first 3 exploration and exploitation trials and switch percentage. *Top panel – axial brain view. MNI coordinates of the first slice: z = 0. Each next slice increases z coordinate in increments of 3. BSR – bootstrap ratio. Bottom panel – correlation of IQR BOLD modulation (expressed as brain scores) with switch percentage. Bootstrapped CI for correlations: Exploration – [-0.67, -0.19], Exploitation – [-0.89, -0.71].*

modulation and switch percentage. However, there was indeed a negative correlation between the brain scores from the PLS analysis with IQR BOLD modulation on exploitation trials and brain scores from the analysis with IQR BOLD level on exploit trial 1 (Spearman's rank correlation: rho = -0.70, p = 0.000001), trial 2 (Spearman's rank correlation: rho = -0.52, p = 0.0006), and trial 3 (Spearman's rank correlation: rho = -0.34, p = 0.03) (**Figure 3-10**). These latter results suggest that although participants who switched more increased IQR BOLD less during exploitation, they nevertheless exhibited a generally more variable neural system from moment to moment.

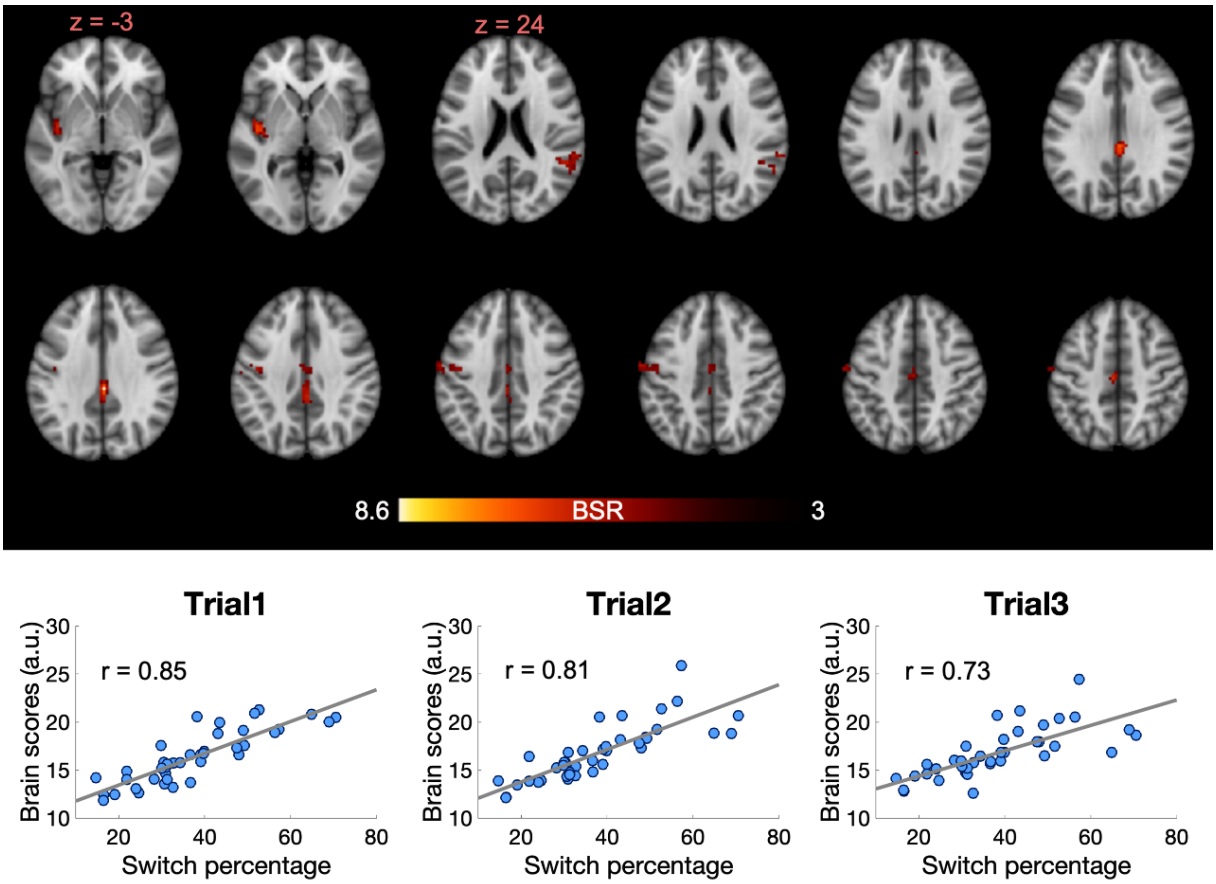


Figure 3-9. Results of behavior PLS analysis with level of IQR BOLD in exploitation trials 1, 2, 3 and switch percentage. *Top panel – axial brain view. MNI z coordinates are indicated. Each next slice increases z coordinate in increments of 3. BSR – bootstrap ratio. Bottom panel – correlation of IQR BOLD level (expressed as brain scores) with switch percentage. Bootstrapped CI for correlations: trial 1 – [0.78, 0.92], trial 2 – [0.76,0.92], trial 3 – [0.68, 0.91].*

3.4 Discussion

In the current study, we presented evidence that uncertainty drives BOLD signal variability, providing a potential neural mechanism underlying flexible switching between exploration and exploitation modes. Specifically, we demonstrate that the level of IQR BOLD parametrically decreased as uncertainty decreased during exploration trials, and increased with growing uncertainty during exploitation trials. We further show that IQR BOLD was most strongly related to posterior estimation uncertainty (posterior sigma), a proxy for the standard deviation of the posterior belief distribution about the estimated value of the reward options. Moreover, our results suggest that more flexible performance was associated with a stronger IQR BOLD decrease during exploration and that higher levels of IQR BOLD in the beginning of the exploitation period could provide the mechanism allowing one to shift out of exploitation mode more quickly, thus leading to more flexible and successful performance.

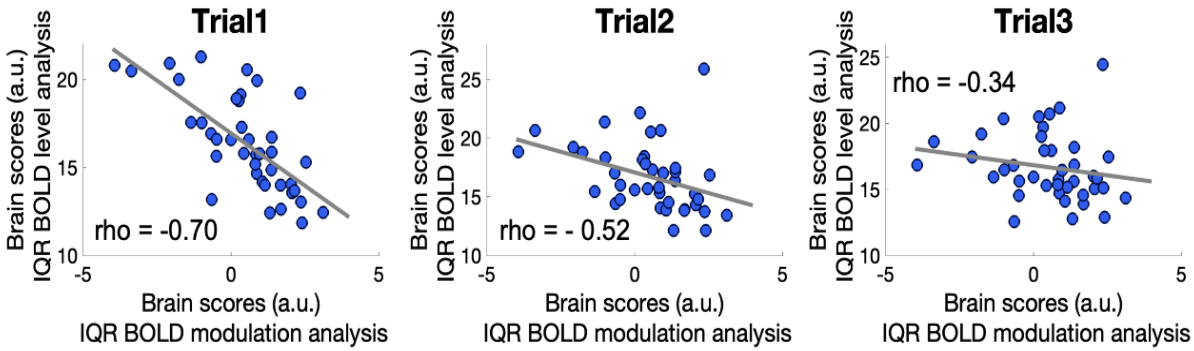


Figure 3-10. Relationship between IQR BOLD modulation and level in exploitation. *Correlations between (a) brain scores extracted from a behavior PLS analysis with IQR BOLD modulation in exploitation and switch percentage and (b) brain scores extracted from a behavior PLS analysis with IQR BOLD level at exploit trial 1, 2, 3 and switch percentage.*

3.4.1 Uncertainty-driven BOLD signal variability as a mechanism to balance exploration and exploitation in a changing environment

In line with existing literature, we show that the change of IQR BOLD levels mirrors the change of uncertainty, which might reflect an adaptation of the neural system to a dynamic, uncertain environment (Garrett, Samanez-Larkin, et al., 2013; Grady & Garrett, 2018). Compared to deterministic environments, uncertainty about the options grows fast and previously learned information quickly becomes obsolete in a changing environment (Behrens et al., 2007; Bruckner et al., 2022; Courville et al., 2006). When one option is exploited, other options are not sampled, thus resulting in rapidly growing uncertainty about their value estimates. In our task, this process is even more extreme, since information and reward feedback are completely separated and no information is presented on exploitation trials. The values of all options are thus forgotten during exploitation, resulting in an increase of uncertainty about all options on each successive exploitation trial. Conversely, choosing an *explore* response in our task provides the only way to decrease uncertainty through exploration (Blanchard & Gershman, 2018; Gershman, 2018) in order to create and maintain an up-to-date picture of the world (Pearson et al., 2011).

Though previous literature linked uncertainty to brain signal variability (Waschke et al., 2021), our study is the first to examine the relationship between different types of uncertainty (prior and posterior estimation uncertainty, choice entropy) and BOLD signal variability. Both prior estimation uncertainty (Daw et al., 2006) and choice uncertainty (Muller et al., 2019) have been shown to play a significant role in exploration-exploitation behavior. Our results nevertheless indicate that BOLD signal variability was most robustly linked to posterior estimation uncertainty. While choice uncertainty additionally reflected value discriminability and prior estimation uncertainty shared influences of both exploration and exploitation modes on switch trials, it was posterior estimation uncertainty that reflected uncertainty change effects unique to each mode, which were also seen in the IQR BOLD trial-sequence analysis.

Previous research suggested that brain signal variability (and BOLD signal variability in particular) might provide a neural mechanism for flexibly adapting behavior to an uncertain, changing environment (Waschke et al., 2021). While higher BOLD signal variability was previously observed for more feature-rich stimuli, which can be thought of as having higher levels of perceptual uncertainty (Garrett et al., 2020), and tasks endowed with more uncertainty compared to less uncertain tasks (Grady & Garrett, 2018), our study is the first to demonstrate parametric changes of BOLD signal variability following changes in uncertainty in the context of exploration-exploitation decision-making. Importantly, while uncertainty levels were inferred from the task design in these studies, our study is the first to quantify uncertainty based on model-derived estimates. Changes of BOLD signal variability with uncertainty observed in our study might indicate a neural adaptation to a situation when the beliefs about the environment become more uncertain (as is the case when uncertainty increases during exploitation), on the one hand, vs. a scenario when the picture of the world becomes more precise (as is the case on exploration trials). In the former case, the agent should be prepared to face a greater number of possible "states" of the world and react to them accordingly, while, in the latter case, instead of spending resources on high preparedness, the agent can focus on choosing the best action according to the known state of the environment (Garrett, Kovacevic, et al., 2013; Grady & Garrett, 2018).

3.4.2 BOLD signal variability differentially relates to behavior in exploration and exploitation

Our study highlights an overall important role of BOLD signal variability for good performance (both flexibly switching between exploration and exploitation and choosing the highest-paying bandit. Good behavioral performance was associated with high levels of BOLD signal variability in both exploration and exploitation, despite them being thought of as opposite functions, requiring flexibility to decide what to explore and focus to concentrate on exploitation. Though this may sound surprising at first, previous literature has already demonstrated similar effects. Higher levels of BOLD signal variability were associated with better behavioral performance in both a switching task, which required cognitive flexibility, and a distractor inhibition task, which required cognitive stability (Armbruster-Genç et al., 2016). Our results lend further support to the interpretation that participants who have higher levels of BOLD signal variability are likely to perform better overall (Garrett, Samanez-Larkin, et al., 2013).

Despite this similarity, our study also highlights how modulation of BOLD signal variability differentially relates to behavior on exploration vs. exploitation trials. Based on the results of task PLS analyses, one might conclude that neural mechanisms behind exploration and exploitation are nothing more than the opposites of each other (BOLD variability increasing as uncertainty increased during exploitation and decreasing as uncertainty decreased during exploration). However, the results of behavior PLS analyses indicate that the neural mechanisms driving exploration and exploitation might be more complex than could be accounted for by treating the two response modes as mere opposites.

For exploration, a decrease of IQR BOLD following a decrease in uncertainty during the first three exploration trials might indicate that the decision process became easier with each successive trial, as the beliefs about the reward structure became more precise. A negative association between switch percentage and IQR BOLD modulation in exploration (which corresponds to a positive association between higher switch percentage and stronger IQR BOLD modulation in the direction of uncertainty change), might also suggest that the easier it was to make a decision on each successive exploration trial, the easier it was to switch between exploration and exploitation modes, possibly because of a better understanding of the reward structure. Skowron et al. (2024) recently reported similar effects: BOLD signal variability decreased as beliefs about the underlying stimulus distribution became more precise with evidence observed on each successive trial. Moreover, participants who showed stronger reduction of BOLD signal variability also showed more accurate behavioral performance, possibly indicating more efficient learning (Skowron et al., 2024). Our results thus fit an interpretation of stronger decrease of variability with each successive exploration trial as a possible marker of a more efficient learning process.

For exploitation, we observed an increase of IQR BOLD following an increase in uncertainty in the first three exploitation trials. However, contrary to our expectations, it was a smaller, not larger, increase of IQR BOLD (a direction opposite to the direction of uncertainty change) that was associated with higher switch percentage (more flexible performance). The latter effect was explained by the length of continuous exploitation sequences – participants who increased IQR BOLD in exploitation less, spent less time continuously in exploitation and thus switched more. In addition, participants who switched more and increased IQR BOLD less during exploitation, nevertheless showed higher levels of BOLD signal variability on each of the three exploitation trials. Taken together, these results may point to a process similar to attractor dynamics. Behavioral stability and flexibility have been previously described as functions of attractor states (Durstewitz & Seamans, 2008; Ueltzhöffer et al., 2015). The authors characterize stability as a low-energy state, which is illustrated by a deep attractor basin and requires more effort to switch to a different mode of action. In contrast, flexibility is described as a high-energy state with a shallow attractor basin, allowing easy switching between different actions (Durstewitz & Seamans, 2008; Ueltzhöffer et al., 2015). Participants who showed lower increases of IQR BOLD during the first three exploitation trials and at the same time had higher levels of IQR BOLD on each of these trials might keep their neural system in a high-energy (i.e., high variability) state, in which switching to exploration may happen more easily. On the other hand, going deeply into an exploitation mode would require more energy to switch to exploration.

3.4.3 Topographic results reveal how flexible adaptation to the environment might support exploration-exploitation decision-making in a changing, uncertain environment

Studies utilizing BOLD signal variability to examine behavioral flexibility are rare (Armbruster-Genç et al., 2016; Grady & Garrett, 2018); our work represents the first study to link BOLD variability to reward-based learning and exploration-exploitation. Spatial distributions of the BOLD effects seen in our task

often encompassed large portions of the brain, revealing the unusual strength of the relationship between BOLD signal variability and uncertainty in the context of exploration-exploitation. Overall, our topographic results emphasize the particular importance of uncertainty processing and behavioral flexibility for balancing exploration and exploitation in a changing environment.

Note that analyses based on the mean BOLD signal and BOLD signal variability often produce complementary results in terms of the brain regions involved in a task (Garrett, Samanez-Larkin, et al., 2013). Therefore, a relationship between a brain region and a cognitive function should be demonstrated by variability-based studies to support the interpretation of the given region's function. While our study makes a significant first contribution to elucidating variability-based neural mechanisms behind exploration-exploitation decision-making, more neuroimaging research based on BOLD variability is needed in the domains of value-based decision-making and reinforcement learning (Waschke et al., 2021).

Behavioral flexibility

Especially frontally-projecting thalamic regions (as were also found in our data) have been shown to play a crucial role in behavioral flexibility (Shine et al., 2023), including switching between task-relevant rules (Marton et al., 2018; Rikhye et al., 2018) and options (Chakraborty et al., 2016), as well as discovering changes in the reward structure and using them to adapt behavior (Chakraborty et al., 2016). Moreover, frontally-projecting neurons in the thalamus have been reported to reflect performance variability by accounting for both the positive influence of reward-related exploration and the negative influence of memory fluctuations (Wang et al., 2020). As for BOLD variability, higher BOLD signal variability in the thalamus has been associated with lower error rates when task switching was required (Armbruster-Genç et al., 2016), thus showing its relevance for behavior. Moreover, thalamic BOLD signal variability could play a prominent role in in the functioning of the brain overall, as it was shown to be a key link between local regional variability and functional integration of the whole brain (Garrett et al., 2018). Our results show that the thalamus might support behavioral flexibility needed to successfully balance exploration and exploitation, as well as establish uncertainty-driven BOLD signal variability as a possible neural mechanism supporting this function in the context of exploration-exploitation decision-making.

Uncertainty processing

Effects linking IQR BOLD to uncertainty on the ExploreExploit task were observed in such brain regions as the thalamus and insula, which have been strongly implicated in general uncertainty-related processing during decision-making (Bach & Dolan, 2012; Morriss et al., 2019; Shine et al., 2023). For instance, parametric increases in thalamic mean BOLD signal tracked parametric increases in uncertainty (Kosciessa et al., 2021), while modulations of BOLD signal variability in the insula were related to task performance as participants decreased uncertainty about the underlying stimulus distribution (Skowron et al., 2024). Modulation of BOLD signal variability in response to perceptual uncertainty level in both regions also showed relevance for behavioral performance; participants who increased BOLD signal variability in response to an increase in uncertainty, performed best on a host of

different tasks (Garrett et al., 2020). In the context of exploration-exploitation research, increased mean BOLD signal in both regions was associated with higher uncertainty in exploration-exploitation tasks (Chakroun et al., 2020; Tomov et al., 2020) and increasing uncertainty during a series of exploitation trials (Chakroun et al., 2020). Furthermore, mean BOLD activity in both regions was reported to be higher in exploration compared to exploitation in a study that used a direct behavioral measure to unambiguously categorize exploration and exploitation trials (like our own) (Blanchard & Gershman, 2018). Our results thus extend the findings of existing exploration-exploitation studies by showing that BOLD signal variability could provide a neural mechanism supporting uncertainty processing in an exploration-exploitation task.

In addition, studies investigating adaptation of BOLD signal variability to different levels of uncertainty reported effects in several other brain regions which are also found in our results. For example, a study by Grady and Garrett (2018) examined BOLD signal variability as participants engaged in internally vs. externally focused tasks, hypothesizing that BOLD variability would be higher on externally oriented tasks as they entail more uncertainty reflecting the need to monitor and adapt to the external world. They found the hypothesized effect in a number of regions that also feature prominently in our results, such as frontal cortex (inferior, middle, and superior frontal gyri), posterior cingulate, posterior parietal cortex, precuneus, and insula. Another study examined BOLD signal variability as perceptual input was parametrically degraded by adding noise (Garrett et al., 2014) and found an inverted U-shape effect seen in a number of brain areas that overlap with our results, including inferior and middle frontal gyri, middle temporal gyrus, anterior cingulate, hippocampus, and thalamus. These brain regions demonstrated changes in BOLD signal variability related to changes in uncertainty in different types of tasks, such as perceptual processing (Garrett et al., 2014), externally vs. internally focused cognition (Grady & Garrett, 2018), and exploration-exploitation decision-making (current study). It is therefore possible that these regions are involved in supporting neural adaptations to changes in uncertainty levels in general.

3.4.4 Summary

In summary, we show that BOLD signal variability reflects uncertainty changes during exploration and exploitation. This might be a mechanism which underlies flexible switching between the two modes, necessary for good performance and, more generally, for successfully navigating a rapidly changing world. Our results highlight the importance of brain regions supporting flexible behavior for successfully adapting exploration and exploitation to a changing, uncertain environment.

3.5 References

Addicott, M. A., Pearson, J. M., Froeliger, B., Platt, M. L., & McClernon, F. J. (2014). Smoking automaticity and tolerance moderate brain activation during explore–exploit behavior. *Psychiatry Research: Neuroimaging*, *224*(3), 254–261. https://doi.org/10.1016/j.pscychresns.2014.10.014

Andersson, J. L. R., Skare, S., & Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *NeuroImage*, *20*(2), 870–888. https://doi.org/10.1016/s1053-8119(03)00336-7

Arco, J. E., González-García, C., Díaz-Gutiérrez, P., Ramírez, J., & Ruz, M. (2018). Influence of activation pattern estimates and statistical significance tests in fMRI decoding analysis. *Journal of Neuroscience Methods*, *308*, 248–260. https://doi.org/10.1016/j.jneumeth.2018.06.017

Armbruster-Genç, D. J. N., Ueltzhöffer, K., & Fiebach, C. J. (2016). Brain Signal Variability Differentially Affects Cognitive Flexibility and Cognitive Stability. *The Journal of Neuroscience*, *36*(14), 3978–3987. https://doi.org/10.1523/jneurosci.2517-14.2016

Avants, B. B., Tustison, N. J., Stauffer, M., Song, G., Wu, B., & Gee, J. C. (2014). The Insight ToolKit image registration framework. *Frontiers in Neuroinformatics*, *8*, 44. https://doi.org/10.3389/fninf.2014.00044

Bach, D. R., & Dolan, R. J. (2012). Knowing how much you don't know: a neural organization of uncertainty estimates. *Nature Reviews Neuroscience*, *13*(8), 572. https://doi.org/10.1038/nrn3289

Badre, D., Doll, B. B., Long, N. M., & Frank, M. J. (2012). Rostrolateral Prefrontal Cortex and Individual Differences in Uncertainty-Driven Exploration. *Neuron*, *73*(3), 595–607. https://doi.org/10.1016/j.neuron.2011.12.025

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. *ArXiv*. https://doi.org/10.48550/arxiv.1406.5823

Beckmann, C. F., & Smith, S. M. (2004). Probabilistic Independent Component Analysis for Functional Magnetic Resonance Imaging. *IEEE Transactions on Medical Imaging*, *23*(2), 137–152. https://doi.org/10.1109/tmi.2003.822821

Behrens, T. E. J., Johansen-Berg, H., Woolrich, M. W., Smith, S. M., Wheeler-Kingshott, C. A. M., Boulby, P. A., Barker, G. J., Sillery, E. L., Sheehan, K., Ciccarelli, O., Thompson, A. J., Brady, J. M., & Matthews, P. M. (2003). Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. *Nature Neuroscience*, *6*(7), 750–757. https://doi.org/10.1038/nn1075

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/10.1038/nn1954

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/10.3758/s13415-017-0556-2

Boer, L. de, Axelsson, J., Riklund, K., Nyberg, L., Dayan, P., Bäckman, L., & Guitart-Masip, M. (2017). Attenuation of dopamine-modulated prefrontal value signals underlies probabilistic reward learning deficits in old age. *ELife*, *6*, e26424. https://doi.org/10.7554/elife.26424

Bond, K., Dunovan, K., Porter, A., Rubin, J. E., & Verstynen, T. (2021). Dynamic decision policy reconfiguration under outcome uncertainty. *ELife*, *10*. https://doi.org/10.7554/elife.65540

Boorman, E. D., Behrens, T. E. J., Woolrich, M. W., & Rushworth, M. F. S. (2009). How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action. *Neuron*, *62*(5), 733–743. https://doi.org/10.1016/j.neuron.2009.05.014

Bruckner, R., Heekeren, H. R., & Nassar, M. R. (2022). *Understanding Learning Through Uncertainty and Bias*. https://doi.org/10.31234/osf.io/xjkbg

Castello, M. V. di O., Dobson, J. E., Sackett, T., Kodiweera, C., V., J., Haxby, Goncalves, M., Ghosh, S., & Halchenko, and Y. O. (2020). ReproNim/reproin 0.6.0. *Zenodo*. https://doi.org/10.5281/zenodo.3625000

Chakraborty, S., Kolling, N., Walton, M. E., & Mitchell, A. S. (2016). Critical role for the mediodorsal thalamus in permitting rapid reward-guided updating in stochastic reward environments. *ELife*, *5*, e13588. https://doi.org/10.7554/elife.13588

Chakroun, K., Mathar, D., Wiehler, A., Ganzer, F., & Peters, J. (2020). Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making. *ELife*, *9*, e51260. https://doi.org/10.7554/elife.51260

Cockburn, J., Man, V., Cunningham, W. A., & O'Doherty, J. P. (2022). Novelty and uncertainty regulate the balance between exploration and exploitation through distinct mechanisms in the human brain. *Neuron*, *110*(16), 2691-2702.e8. https://doi.org/10.1016/j.neuron.2022.05.025

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/10.1098/rstb.2007.2098

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*(7), 294–300. https://doi.org/10.1016/j.tics.2006.05.004

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. https://doi.org/10.1038/nature04766

Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P., & Meester, L. E. (2005). A Modern Introduction to Probability and Statistics, Understanding Why and How. *Springer Texts in Statistics*. https://doi.org/10.1007/1-84628-168-7

Dombrovski, A. Y., Luna, B., & Hallquist, M. N. (2020). Differential reinforcement encoding along the hippocampal long axis helps resolve the explore–exploit dilemma. *Nature Communications*, *11*(1), 5407. https://doi.org/10.1038/s41467-020-18864-0

Doya, K. (2008). Modulators of decision making. *Nature Neuroscience*, *11*(4), 410–416. https://doi.org/10.1038/nn2077

Durstewitz, D., & Seamans, J. K. (2008). The Dual-State Theory of Prefrontal Cortex Dopamine Function with Relevance to Catechol-O-Methyltransferase Genotypes and Schizophrenia. *Biological Psychiatry*, *64*(9), 739–749. https://doi.org/10.1016/j.biopsych.2008.05.015

Efron, B., & Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, *1*(1). https://doi.org/10.1214/ss/1177013815

Findling, C., & Wyart, V. (2021). Computation noise in human learning and decision-making: origin, impact, function. *Current Opinion in Behavioral Sciences*, *38*, 124–132. https://doi.org/10.1016/j.cobeha.2021.02.018

Garrett, D. D., Epp, S. M., Kleemeyer, M., Lindenberger, U., & Polk, T. A. (2020). Higher performers upregulate brain signal variability in response to more feature-rich visual input. *NeuroImage*, *217*, 116836. https://doi.org/10.1016/j.neuroimage.2020.116836

Garrett, D. D., Epp, S. M., Perry, A., & Lindenberger, U. (2018). Local temporal variability reflects functional integration in the human brain. *NeuroImage*, *183*(Neuroimage 172 2018), 776–787. https://doi.org/10.1016/j.neuroimage.2018.08.019

Garrett, D. D., Kovacevic, N., McIntosh, A. R., & Grady, C. L. (2010). Blood Oxygen Level-Dependent Signal Variability Is More than Just Noise. *The Journal of Neuroscience*, *30*(14), 4914–4921. https://doi.org/10.1523/jneurosci.5166-09.2010

Garrett, D. D., Kovacevic, N., McIntosh, A. R., & Grady, C. L. (2013). The Modulation of BOLD Variability between Cognitive States Varies by Age and Processing Speed. *Cerebral Cortex*, *23*(3), 684–693. https://doi.org/10.1093/cercor/bhs055

Garrett, D. D., McIntosh, A. R., & Grady, C. L. (2014). Brain Signal Variability is Parametrically Modifiable. *Cerebral Cortex*, *24*(11), 2931–2940. https://doi.org/10.1093/cercor/bht150

Garrett, D. D., Nagel, I. E., Preuschhof, C., Burzynska, A. Z., Marchner, J., Wiegert, S., Jungehülsing, G. J., Nyberg, L., Villringer, A., Li, S.-C., Heekeren, H. R., Bäckman, L., & Lindenberger, U. (2015). Amphetamine modulates brain signal variability and working memory in younger and older adults. *Proceedings of the National Academy of Sciences*, *112*(24), 7593–7598. https://doi.org/10.1073/pnas.1504090112

Garrett, D. D., Samanez-Larkin, G. R., MacDonald, S. W. S., Lindenberger, U., McIntosh, A. R., & Grady, C. L. (2013). Moment-to-moment brain signal variability: A next frontier in human brain mapping? *Neuroscience & Biobehavioral Reviews*, *37*(4), 610–624. https://doi.org/10.1016/j.neubiorev.2013.02.015

Gershman, S. J. (2018). Uncertainty and Exploration. *Decision*, *6*(3), 277–286. https://doi.org/10.1037/dec0000101

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., … Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, *3*(1), 160044. https://doi.org/10.1038/sdata.2016.44

Grady, C. L., & Garrett, D. D. (2018). Brain signal variability is modulated as a function of internal and external demand in younger and older adults. *NeuroImage*, *169*. https://doi.org/10.1016/j.neuroimage.2017.12.031

Guitart-Masip, M., Salami, A., Garrett, D., Rieckmann, A., Lindenberger, U., & Bäckman, L. (2016). BOLD Variability is Related to Dopaminergic Neurotransmission and Cognitive Aging. *Cerebral Cortex*, *26*(5), 2074–2083. https://doi.org/10.1093/cercor/bhv029

Haynes, J.-D. (2015). A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. *Neuron*, *87*(2), 257–270. https://doi.org/10.1016/j.neuron.2015.05.025

Hogeveen, J., Mullins, T. S., Romero, J. D., Eversole, E., Rogge-Obando, K., Mayer, A. R., & Costa, V. D. (2022). The neurocomputational bases of explore-exploit decision-making. *Neuron*, *110*(11), 1869-1879.e5. https://doi.org/10.1016/j.neuron.2022.03.014

Kloosterman, N. A., & Garrett, D. D. (n.d.). *Tracking of natural-scene feature richness in brain signal variability predicts memory and changes with age*.

Kosciessa, J. Q., Lindenberger, U., & Garrett, D. D. (2021). Thalamocortical excitability modulation guides human perception under uncertainty. *Nature Communications*, *12*(1), 2430. https://doi.org/10.1038/s41467-021-22511-7

Krishnan, A., Williams, L. J., McIntosh, A. R., & Abdi, H. (2011). Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review. *NeuroImage*, *56*(2), 455–475. https://doi.org/10.1016/j.neuroimage.2010.07.034

Lindenberger, U., & Lövdén, M. (2019). Brain Plasticity in Human Lifespan Development: The Exploration–Selection–Refinement Model. *Annual Review of Developmental Psychology*, *1*(1), 197–222. https://doi.org/10.1146/annurev-devpsych-121318-085229

Marton, T. F., Seifikar, H., Luongo, F. J., Lee, A. T., & Sohal, V. S. (2018). Roles of Prefrontal Cortex and Mediodorsal Thalamus in Task Engagement and Behavioral Flexibility. *The Journal of Neuroscience*, *38*(10), 2569–2578. https://doi.org/10.1523/jneurosci.1728-17.2018

McCarthy, P. (2023). FSLeyes (1.7.0). *Zenodo*. https://doi.org/10.5281/zenodo.8033457

McIntosh, A. R., Bookstein, F. L., Haxby, J. V., & Grady, C. L. (1996). Spatial Pattern Analysis of Functional Brain Images Using Partial Least Squares. *NeuroImage*, *3*(3), 143–157. https://doi.org/10.1006/nimg.1996.0016

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., Hausmann, D., Fiedler, K., & Gonzalez, C. (2015). Unpacking the Exploration–Exploitation Tradeoff: A Synthesis of Human and Animal Literatures. *Decision*, *2*(3), 191–215. https://doi.org/10.1037/dec0000033

Morriss, J., Gell, M., & Reekum, C. M. van. (2019). The uncertain brain: A co-ordinate based meta-analysis of the neural signatures supporting uncertainty during different contexts. *Neuroscience & Biobehavioral Reviews*, *96*, 241–249. https://doi.org/10.1016/j.neubiorev.2018.12.013

Muller, T. H., Mars, R. B., Behrens, T. E., & O'Reilly, J. X. (2019). Control of entropy in neural models of environmental state. *ELife*, *8*. https://doi.org/10.7554/elife.39404

Mumford, J. A., Davis, T., & Poldrack, R. A. (2014). The impact of study design on pattern estimation for single-trial multivariate pattern analysis. *NeuroImage*, *103*, 130–138. https://doi.org/10.1016/j.neuroimage.2014.09.026

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Orbán, G., Berkes, P., Fiser, J., & Lengyel, M. (2016). Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex. *Neuron*, *92*(2), 530–543. https://doi.org/10.1016/j.neuron.2016.09.038

Pearson, J. M., Heilbronner, S. R., Barack, D. L., Hayden, B. Y., & Platt, M. L. (2011). Posterior cingulate cortex: adapting behavior to a changing world. *Trends in Cognitive Sciences*, *15*(4), 143–151. https://doi.org/10.1016/j.tics.2011.02.002

Rikhye, R. V., Gilra, A., & Halassa, M. M. (2018). Thalamic regulation of switching between cortical representations enables cognitive flexibility. *Nature Neuroscience*, *21*(12), 1753–1763. https://doi.org/10.1038/s41593-018-0269-z

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shine, J. M., Lewis, L. D., Garrett, D. D., & Hwang, K. (2023). The impact of the human thalamus on brain-wide information processing. *Nature Reviews Neuroscience*, 1–15. https://doi.org/10.1038/s41583-023-00701-0

Skowron, A., Kosciessa, J. Q., Lorenz, R., Hertwig, R., Bos, W. van den, & Garrett, D. D. (2024). Neural variability compresses with increasing belief precision during Bayesian inference. *BioRxiv*, 2024.01.11.575180. https://doi.org/10.1101/2024.01.11.575180

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., Bannister, P. R., Luca, M. D., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., Stefano, N. D., Brady, J. M., & Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, *23*, S208–S219. https://doi.org/10.1016/j.neuroimage.2004.07.051

Stoffel, M. A., Nakagawa, S., & Schielzeth, H. (2021). partR2: partitioning R2 in generalized linear mixed models. *PeerJ*, *9*, e11414. https://doi.org/10.7717/peerj.11414

Tardiff, N., Medaglia, J. D., Bassett, D. S., & Thompson-Schill, S. L. (2021). The modulation of brain network integration and arousal during exploration. *NeuroImage*, *240*, 118369. https://doi.org/10.1016/j.neuroimage.2021.118369

Team, R. C. (2022). *R: A language and environment for statistical computing*. URL https://www.R-project.org/

Tomov, M. S., Truong, V. Q., Hundia, R. A., & Gershman, S. J. (2020). Dissociable neural correlates of uncertainty underlie different exploration strategies. *Nature Communications*, *11*(1), 2371. https://doi.org/10.1038/s41467-020-15766-z

Ueltzhöffer, K., Armbruster-Genç, D. J. N., & Fiebach, C. J. (2015). Stochastic Dynamics Underlying Cognitive Stability and Flexibility. *PLOS Computational Biology*, *11*(6), e1004331. https://doi.org/10.1371/journal.pcbi.1004331

Wang, J., Hosseini, E., Meirhaeghe, N., Akkad, A., & Jazayeri, M. (2020). Reinforcement regulates timing variability in thalamus. *ELife*, *9*, e55872. https://doi.org/10.7554/elife.55872

Waschke, L., Kloosterman, N. A., Obleser, J., & Garrett, D. D. (2021). Behavior needs neural variability. *Neuron*. https://doi.org/10.1016/j.neuron.2021.01.023

Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal Autocorrelation in Univariate Linear Modeling of FMRI Data. *NeuroImage*, *14*(6), 1370–1386. https://doi.org/10.1006/nimg.2001.0931

4. Gaze patterns as a real-time observed marker of exploration-exploitation decision-making in a reinforcement learning task

Abstract

In contrast to latent-level accounts of the decision-making process achieved via computational modeling, gaze analysis an observable (and grossly underutilized) measure of decision-making dynamics. Using eye movements as a marker of the underlying decision-making processes in combination with our newly designed ExploreExploit task, we investigate the utility of fixation-based dwell location patterns during decision-making in predicting exploration and exploitation responses and level of task performance. To our knowledge, the current study is the first to analyze gaze behavior in the context of exploration-exploitation decision-making in a reward-based reinforcement learning task. We demonstrate that gaze behavior during the decision-making period reflects both the expected value and uncertainty of the options and predicts the trial type (exploration or exploitation). Moreover, we show that trials with different numbers of dwell locations might provide complementary insights into the decision-making process, such as how confident participants are in their choice at the beginning of the trial and how they compare options under consideration. Our results also indicate that poorer task performance is associated with participants applying correct gaze strategies to an inaccurate mental image of the reward structure. Our results demonstrate the utility of eye tracking data as a real-time observable measure of the processes that lead to a decision to explore or exploit.

Keywords: exploration, exploitation, eye tracking, gaze patterns, value-based decision-making, reinforcement learning

4.1 Introduction

On the use of scan path analysis as a directly observable signature of decision-making

While computational modeling provides an indispensable window into the mechanics of decision-making processes at a *latent* level, eye movements continue to provide a *directly observable* signature of decision-making processes in real time (Huddleston et al., 2018; Spering, 2022). Although it remains dramatically underutilized in the field, gaze analysis has been used in a number of value-based research fields, such as consumer behavior (Gidlöf et al., 2013; Jacob & Karn, 2003), value-based decision-making (Gluth et al., 2020; Krajbich & Rangel, 2011; Thomas et al., 2021), decision-making under risk (Glöckner & Herbold, 2011; Zhou et al., 2016), and economic decision-making (Byrne et al., 2023; Krol & Krol, 2017; Polonio et al., 2015) to better understand how decisions evolve, to probe the role of item characteristics (e.g. value), and to predict choice. Among gaze analysis techniques, scan path analysis is particularly suited to capturing the development of a decision process over time (Byrne et al., 2023; Kümmerer & Bethge, 2021; Polonio et al., 2015) as it represents spatial positions of eye fixations in a temporal order (Byrne et al., 2023; Jacob & Karn, 2003). In economic decision-making studies, scan paths have been shown to differentiate between decision strategies that resulted in an optimal choice and those that did not (Polonio et al., 2015). Scan paths have also been used as a basis for machine learning analyses differentiating optimal and sub-optimal choice strategies (Byrne et al., 2023; Krol & Krol, 2017).

In addition, many-alternative (6 alternatives (Russo & Rosen, 1975), 9-36 alternatives (Thomas et al., 2021)) value-based decision-making studies demonstrated that, instead of looking at all items sequentially and then choosing the one with the highest value, participants selected the preferred item using a simpler kind of comparative processing, during which their gaze often alternated between two items (Russo & Rosen, 1975; Thomas et al., 2021). For example, Russo and Rosen (1975) describe patterns created by gaze as often alternating between two out of six options as “xyx” or “xyxy…”, revealing that subjects will shrink the pool of viable options and iterate between them until a decision is made. However, it is not yet clear how and why options enter the set of alternatives that are considered for a choice.

Beyond gaze patterning, the eye tracking literature also demonstrates that options that are fixated more often and for a longer total time are more likely to be chosen (Jacob & Karn, 2003; Krajbich et al., 2010; Thomas et al., 2021). Studies on value-based decision-making report more fixations on trials on which choices were more difficult (defined as value similarity between alternatives) (Callaway et al., 2021; Krajbich et al., 2010). In addition, value-based decision-making studies emphasized a strong role of the expected value of options in driving both fixation duration and probability of fixating on an item again (Gluth et al., 2020; Krajbich & Rangel, 2011; Thomas et al., 2021). Callaway and colleagues (2021) demonstrated that both expected value and uncertainty influenced gaze behavior; fixations were most often allocated to the option with the highest expected value and highest uncertainty during choices between three alternatives. While the last fixated item was shown to have higher probability of being

chosen in general, and especially if it had high expected value (Callaway et al., 2021; Krajbich et al., 2010; Thomas et al., 2021), such effects were not present for the first fixated item (Krajbich et al., 2010). The absence of the effect for the first fixated item is not surprising given that each trial in these paradigms begins with a need to collect sensory evidence about what options participants could choose from. Instead, the probability of choosing the first fixated item grew with the duration of the first fixation (Callaway et al., 2021; Krajbich et al., 2010; Thomas et al., 2021).

The majority of past research on eye movements in value-based decision-making has made use of task designs in which participants must first collect visual information to know which options are available to be chosen (e.g. Krajbich & Rangel, 2011). Consequently, in tasks with a small number of alternatives, each option presented on the current trial *must* be fixated on at least once; as such, weighing and selecting the options can be done only after the initial collection of visual information about which alternative options are available on the current trial (Spering, 2022). Moreover, if an item is fixated again, it is unclear whether that reflects the process of comparing and selecting alternatives or is caused by not having acquired enough sensory evidence during the first fixation of that item (Russo & Rosen, 1975). However, the necessity to look at each option at least once to know what one is choosing from could be eliminated by a task in which the *same* options are presented on each trial, obviating the need to collect any visual information prior to decision-making. Only in this experimental scenario can fixating on choice options be regarded as a valid signature of decision-making that is relatively independent of sensory evidence accumulation (Russo & Rosen, 1975; Spering, 2022).

Using scan path analysis to understand explore-exploit decision-making

No previous studies have analyzed gaze behavior in the context of exploration-exploitation decision-making. Here, we capitalize on the unique design elements of the ExploreExploit task (also utilized in Chapters 2 and 3) to do so; by not requiring participants to collect any visual information prior to making a choice, gaze behavior on ExploreExploit trials cannot be muddied by the *need* to look at choice options. We achieve this given that visual input during decision-making is static at all times (i.e., participants are shown three fixed geometric figures representing the bandits).

We first sought to establish that gaze reflected the decision-making processes behind responses in our task. We expected the chosen bandit to be fixated more often than unchosen ones, showing that the eye tracking data reflected choice in a model-independent way. We also expected that fixating on a higher-paying vs. more uncertain bandit would be associated with exploitation and exploration respectively, revealing that gaze reflects the decision-making processes formulated in our computational model (see Chapter 1).

We then applied the idea of the scan path as a temporal sequence of spatially localized dwell locations (Jacob & Karn, 2003) to investigate whether the *order* in which participants looked at the bandits during the decision-making period (dwell patterns) predicted the choice to explore or exploit. Our aim was to uncover how temporal ordering of dwell locations could provide a detailed signature of exploration-exploitation decision-making, beyond what is shown by more general measures such as fixation count

and duration. Dwell pattern analyses were anchored by the idea of an association between expected value and exploitation on one hand, and uncertainty and exploration on the other. Since our study is the first to investigate gaze in the exploration-exploitation domain, we formed several general hypotheses. First, we reasoned that dwelling on the highest-paying bandit should predict exploitation and dwelling on the most uncertain bandit should predict exploration. Since it is not necessary for our participants to collect visual evidence to make a choice, any particular trial could include any number of dwell locations, resulting in patterns of different lengths. The effect of fixating on the bandit with highest expected value predicting exploitation and on the bandit with highest uncertainty predicting exploration could be most evident on trials on which just one bandit was fixated, as this could correspond to an "easy choice" (cf. Krajbich et al., 2010). In patterns with multiple dwell locations, the start and the end bandit could be particularly informative for a choice (cf. Callaway et al., 2021). A focus on the start bandit might indicate that participants start the trial with a preference for one option, and this preference is confirmed in the course of the decision-making. A focus on the end bandit might suggest that participants were less sure at the beginning of the trial and formed a preference towards the end of the decision-making period. In addition, in patterns with multiple dwell locations, it may be expected that several bandits that are considered for an explore or exploit choice could be compared by alternating the gaze between them (e.g. the highest-paying or most uncertain bandit could be looked at interspersed with other bandits, taking a form of xyx, xyxy or xyxz) (cf. Russo & Rosen, 1975). Combined, such gaze analytic insights could provide novel elucidation of the real-time dynamics of exploration-exploitation decision-making.

4.2 Materials and Methods

The eye tracking data presented was collected as part of the fMRI study, previously described in Chapters 2 and 3. Please refer to these chapters for detailed information on participants, task design, and computational modeling. In the following, we briefly summarize the most important methods-related points of the current study and describe acquisition and analysis of the eye tracking data.

4.2.1 Participants

Fifty-one young adults were tested in the current study and 47 had usable behavioral data (see Chapter 2). The eye tracking data for four of these participants was not available due to hardware problems during data acquisition. In addition, we excluded five participants (as we did in Chapter 3) who had virtually no sequences of multiple exploration or exploitation trials in a row. The final sample thus consisted of 38 participants.

4.2.2 Task design

The ExploreExploit task is a 3-armed bandit task in which participants choose whether they want to explore or exploit one of the bandits on each trial (see **Figure 2-1** in Chapter 2 for details). The rewards behind each bandit change according to a random walk. Crucially, the feedback consists only of reward

after exploitation choices and only of information after exploration was chosen. There were 500 trials in the task.

4.2.3 Computational model

We developed a reinforcement learning computational model, which was shown to best reflect the behavior in our task (see *Computational modeling* section in the Methods of Chapter 2 for details). This model calculates the expected value (EV) for each bandit on each trial, as well as the prior uncertainty (prior sigma; before the choice is made) and posterior uncertainty (posterior sigma; after one bandit was explored or exploited) associated with it. Since we examine only the decision-making period (time starting with the end of feedback on a previous trial and ending with a button press; cf. **Figure 2-1**) in the current study, we use only prior sigma (in the following, referred to as sigma) as a measure of uncertainty.

4.2.4 Eye tracking data acquisition and preprocessing

The eye tracking data was collected using EyeLink 1000 eye tracker (SR Research Ltd., Mississauga, Ontario, Canada) at a sampling rate of 1000 Hz. A 5-point calibration and validation procedure was done prior to the first task block and repeated, if needed, prior to any other block.

The preprocessing of the eye tracking data was done using EyeLink software, the FieldTrip toolbox (Oostenveld et al., 2011) and custom MATLAB (version R2020a, https://de.mathworks.com/) scripts (Kloosterman & Garrett, n.d.). Fixations were defined as data points not classified as saccades or blinks by the EyeLink software. For each geometric figure that represented a bandit, we defined location boxes (i.e. regions of interest) in such a way that horizontal space between the figures was divided evenly between the boxes. Fixations within each box were assigned to the respective bandit and fixations outside of the location boxes were removed from subsequent analyses. We further excluded fixations with a duration shorter than 50 ms and longer than 800 ms (Kloosterman & Garrett, n.d.). In general, most participants' (19 subjects) data contained less than 10% of trials with no valid eye tracking data. Among included participants, the highest level of missingness was seen in 4 participants with 30%-40% missing data and one participant falling into the range of 50-60% (**Figure 4-S1**). Regardless, enough trials remained in all of these participants to model effects of interest in the current study.

Since the goal of this study was to investigate gaze in the time until decision to explore or exploit, the time in each trial began with the end of feedback on the previous trial and ended with the button press. This includes the decision-making period (Figure 1 in Chapter 2; the duration of the decision-making period was sampled from a 1000 to 2000 ms uniform distribution, but due to an issue with the experiment code, this time was increased to ca. 1175 - 2200 ms, which did not affect performance on the task) and the time from the response cue to the button press (max. 1500 ms). In the following, we refer to the combination of these trial parts as the "decision-making period," by which we mean the time until the

button press. When we speak of the "trial time" in the following text, it refers directly to the time of the decision-making period on that trial.

4.2.5 Eye tracking data analyses

Measures of interest

For the analyses, we combined consecutive fixations within the same location box (i.e., fixations on the same bandit) into dwell locations (also called gaze locations). Dwell location is thus equivalent to bandit type. We define dwell time as the cumulative time spent in one location box (i.e., fixating on that specific bandit). Dwell time is expressed as a fraction of the total decision-making period, within-trial.

To link gaze behavior to our computational behavioral modeling parameters, we first characterized each bandit on each trial based on its expected reward (Q) (to which we refer as expected value (EV) in the current study) and uncertainty (sigma) associated with it. Specifically, each bandit was assigned an EV rank and a sigma rank, which ranged from 1 to 3 (rank 1 – bandit with the highest EV/sigma, rank 2 – bandit with the middle EV/sigma, rank 3 – bandit with the lowest EV/uncertainty on the respective trial). We then used the logic of scan paths (the temporal sequence of spatial positions of fixations) to create "dwell patterns," defined as sequences of dwell locations based on the expected value (EV) or sigma ranks of the respective bandits. We separately analyzed trials with one, two, or three or more dwell locations (referred to as 1 dwell location, 2 dwell locations, and 3 dwell locations in the following text). For trials with 1 dwell location, the dwell pattern is equivalent to the rank of the fixated bandit (**Figure 4-1**). For trials with 2 dwell locations, the dwell pattern starts with one of the possible 3 ranks and ends

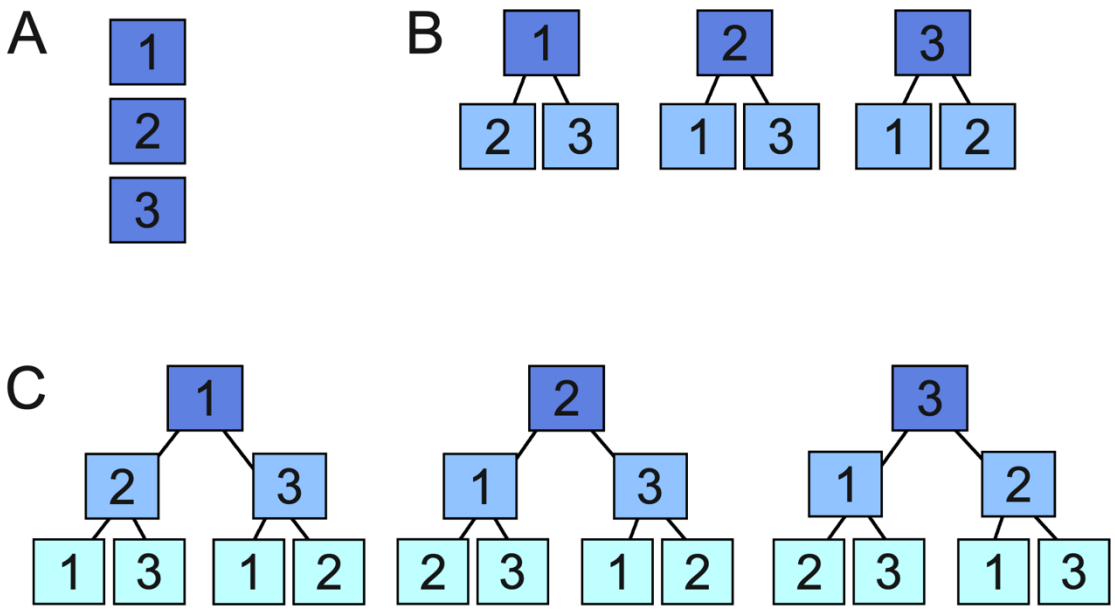


Figure 4-1. Structure of dwell patterns. *Dwell patterns with 1 (A), 2 (B), and 3 (C) dwell locations. Dwell patterns were based on expected value (EV) or sigma ranks. E.g. pattern 121 means that participant visited the location box of the bandit with rank 1, then that of the bandit with rank* 2, and then went back to bandit of rank 1 again.

with one of the other 2 ranks (resulting in 6 patterns in total). Finally, a pattern with 3 dwell locations starts with one of the 3 ranks, followed by one of the other 2 ranks, which is in turn followed by one of the other 2 ranks relative to the second position (resulting in 12 patterns in total). We refer to patterns with three dwell locations that start and end with the same rank as *xyx* patterns (e.g. 121, 131, as opposed to 123, 132).

Lastly, for dwell patterns with two or three locations, we additionally analyzed dwell time in each position in the dwell pattern, expressed as the fraction of total time in the sequence.

Statistical Analyses

For the analyses, we used a series of mixed-effects regression models implemented in the lme4 package (Bates et al., 2014) in R (version 4.3.2) (R Core Team, 2022). All such models include subject ID as a random intercept. For analyses with a dichotomous dependent variable (e.g., explore vs exploit trial type), we used a generalized linear mixed model with a logit link function (a mixed-effects logistic regression model) fit by maximum likelihood (REML; default option in the lme4 package (Bates et al., 2015) in R). In all logistic regression models with trial type as a dependent variable, exploitation was coded as 0 and exploration was coded as 1. We report results of logistic regression models as tables with standardized odds ratios (OR) and their respective 95% confidence intervals (95% CI). Results of mixed linear regression models are reported as tables with the regression weights (β) and their 95% CI. Results tables were produced using tab_model() function from sjPlot package (Lüdecke, 2023). We used the emmeans package (Lenth, 2022) to plot the effects of our statistical models.

Additionally, standardized odds ratios (OR) and their 95% CI are used as effect size measure in reported results of the logistic regression models (Chen et al., 2010; Ialongo, 2016; Peng et al., 2002). For a binary dependent variable and a binary independent variable (as is often the case in our models), OR can be interpreted as follows: OR larger than 1 indicates that the odds of the outcome coded as 1 (e.g. exploration) are higher than the odds of the outcome coded as 0 (e.g. exploitation) when the level of the binary predictor is 1 (e.g. xyx pattern) as opposed to 0 (e.g. non-xyx pattern) (Chen et al., 2010; Peng et al., 2002). In addition, for general linear mixed models we used marginal R^2 (R^2) as a measure of the effect size (Nakagawa & Schielzeth, 2013), as implemented in the partR2 package (Stoffel et al., 2021) in R.

Since participants don't have to collect visual information during the decision-making period and consequently don't have to look to any specific option on any particular trial (after button mappings are learned during piloting and in the first few exploration/exploitation trials), we first established the validity of the eye tracking data in relation to decision-making by analyzing whether the eye tracking data reflected choice. We used a mixed-effects logistic regression model with "fixated" (1 – the bandit/dwell location was fixated on the current trial, 0 – not fixated) as the dependent variable and "chosen" (1 – the bandit was chosen on the current trial, 0 – unchosen), trial type (exploration, exploitation), and their interaction as independent variables. This model tests whether the chosen bandit was looked at more often than the unchosen bandits and whether this was different between exploration and exploitation.

Next, we analyzed whether, on trials on which the chosen bandit was fixated, it was fixated longer than the unchosen bandits. For this, we used a linear mixed-effects regression model with dwell time as the dependent variable and chosen and trial type (defined as described above), as well as an interaction between them, as independent variables.

Next, we checked whether gaze behavior reflected key parameters of our computational model. First, we ran separate mixed-effects logistic regression models based on either expected value (EV) and uncertainty (sigma) ranks, with trial type as a dependent variable and EV/sigma rank (1 – rank 1, 2 – rank 2, 3 – rank 3), fixated (1 – the bandit was fixated on the current trial, 0 – not fixated), and their interaction as independent variables. Then, we investigated whether there was an interaction between gaze behavior driven by EV (expressed as fixating on the bandit with the highest expected value) and sigma (expressed as fixating on the bandit with the highest uncertainty) in determining the trial type. This corresponds to the role of value and uncertainty in determining the choice in the computational model, reflecting that the highest expected value and highest uncertainty contribute most to the probability of the bandit being chosen. For this analysis we used a mixed-effects logistic regression model with variables ev_rank1 (1 – bandit of EV rank 1 (highest expected value) was fixated on the current trial, 0 – not fixated), sigma_rank1 (same as for EV, but for sigma rank 1 (highest uncertainty)), and an interaction between them as independent variables predicting the trial type.

We then asked whether dwell patterns during the decision-making period predicted the trial type (exploration or exploitation). To this end, we used a series of mixed-effects logistic regression models to analyze patterns with 1, 2, and 3 dwell locations (found on trial with 1, 2, and 3 or more dwell locations, respectively) expressed as sequences of EV (expected value) or sigma (uncertainty) ranks of the respective bandits in the pattern. We analyzed patterns with max. 3 dwell locations because, in contrast to patterns with fewer dwell locations, such patterns could capture looking at all 3 bandits, while still providing a reasonable number of trials to analyze in our design. To elucidate whether looking at different ranks of EV and sigma could predict choosing exploration vs. exploitation, we first ran models for EV and sigma separately, and then followed with a combined model to examine potential interaction effects between EV and sigma for predicting trial type.

To better understand gaze patterns, we then investigated whether the dwell time (how long participants looked at a bandit) in a specific position in a pattern also predicted the trial type. Dwell time was the time spent in a given gaze location (bandit), defined as a fraction of total time of the dwell pattern. We investigated whether dwell time in each location differentially predicted trial type, using mixed-effects logistic regression models for patterns with 2 and 3 dwell locations. Since patterns with 1 dwell location have only one position, we did not include them in the analysis.

Finally, to probe the relationship between the presence of the dwell patterns and task performance, we correlated the frequency of dwell patterns with 1, 2, and 3 dwell locations in exploration and exploitation with behavioral measures of task performance, which included (1) optimal choice percentage and (2)

switch percentage. For patterns with 3 dwell locations, we limited the analysis to xyx patterns (see Results below).

4.3 Results

4.3.1 Fixations reflect choice

First, we examined whether the eye tracking data reflected choice. A logistic regression model predicting fixation (bandit fixated or not) showed a strong main effect of choice (OR = 14.53, 95% CI = [13.63, 15.50], p < 2e-16; all model results are reported in **Table 4-S1**); a bandit was more likely to be fixated if it was chosen on the current trial. Importantly there was a significant interaction between choice (chosen, unchosen) and the trial type, showing that exploration was more likely when an unchosen bandit was fixated and exploitation was more likely when a chosen bandit was fixated (OR = 0.54, 95% CI = [0.49, 0.60], p < 2e-16; **Figure 4-2A**).

Moreover, on those trials on which the chosen bandit was fixated, the dwell time on the chosen bandit was longer than on the unchosen bandits (β =0.28, 95% CI = [0.27, 0.29], p < 2e-16, R^2 = 0.11; **Table 4-S1**). Importantly, this model also showed a significant interaction between choice (chosen, unchosen) and trial type; dwell time on the chosen bandit was much longer on exploitation than on exploration trials, while there was only a modest difference for the unchosen bandits ((β =-0.04, 95% CI = [-0.05, -0.03], p = 8.89e-10, R^2 = 0.001; **Figure 4-2B**).

4.3.2 Fixations reflect computational modelling parameters

In the next step, we examined whether fixations were driven by both expected value (EV) and uncertainty (sigma), thus reflecting the key parameters of our computational model of behavior. First, we used two separate logistic regression models to examine whether fixating on different EV or sigma ranks differentially predicted the trial type, which was the case in both models.

The distribution of fixations to bandits of different EV and sigma ranks in exploration and exploitation is presented in **Figure 4-3A**. The EV-based model (**Table 4-S2**, **Figure 4-3B**) produced significant interactions between fixation (bandit fixated or not) and EV rank (fixated x rank 2: OR = 3.25, 95% CI = [2.93, 3.61], p < 2e-16; fixated x rank 3: OR = 3.29, 95% CI = [2.96, 3.66], p < 2e-16), showing that the slope of EV rank 1 (highest expected value) decreased when the bandit was fixated as opposed to not fixated, while the opposite was true for the slopes of EV ranks 2 and 3. Similarly, interactions between fixation (bandit fixated or not) and sigma ranks significantly predicted the trial type (fixated x rank 2: OR = 0.87, 95% CI = [0.79, 0.95], p = 0.03; fixated x rank 3: OR = 0.51, 95% CI = [0.47, 0.57], p < 2e-16). While the slopes for sigma ranks 1 and 2 increased from when the bandit was not fixated to when it was fixated, the opposite was true for rank 3 (**Table 4-S2**, **Figure 4-3B**).

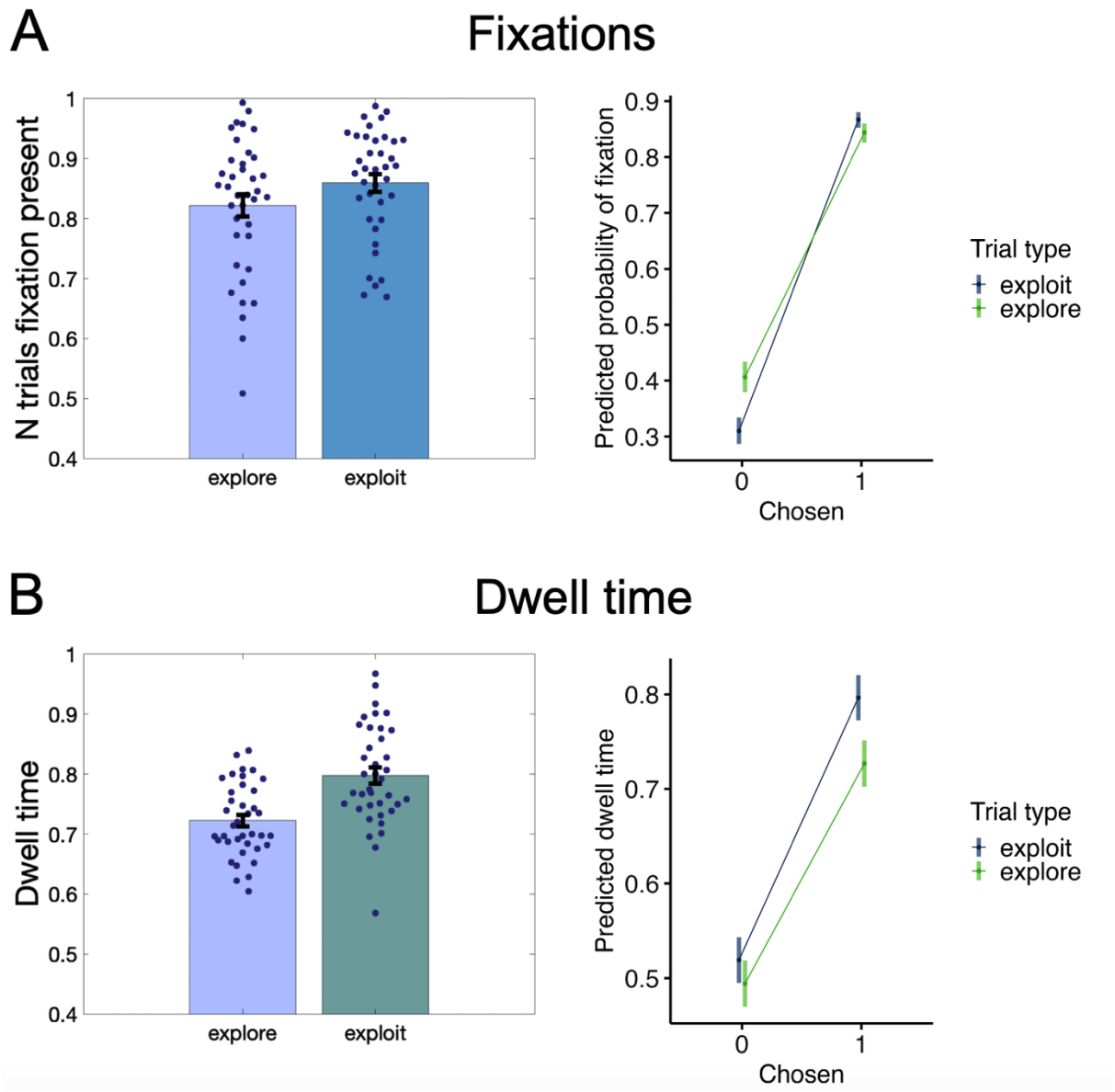


Figure 4-2. Fixations and dwell time reflect choice. *(A) Left – number of trials on which fixation on a chosen bandit was present, expressed as a fraction of total number of trials in each response condition. Right – interaction between choice (0 – unchosen, 1 – chosen) and trial type in a model predicting fixation (0 – not fixated, 1 – fixated). (B) Left – dwell time on the chosen bandit on trials on which chosen bandit was fixated, expressed as a fraction of total time in the decision-making period on the respective trial. Right – interaction between choice (0 – unchosen, 1 – chosen) and trial type in a model predicting dwell time. Dwell time on each bandit is expressed as a fraction of total time in the decision-making period on respective trial.*

Since the exploration and exploitation values of the bandits were driven most strongly by the highest expected value and the highest uncertainty in the computational model, we selected only EV rank 1 and sigma rank 1 for a combined EV and sigma logistic regression model. In this combined model, trial type was significantly predicted by an interaction between fixating on a bandit with EV rank 1 and fixating on a bandit with sigma rank 1 (OR = 1.39, 95% CI = [1.19, 1.63], p < 2e-16; **Table 4-S3**, **Figure 4-3C**),

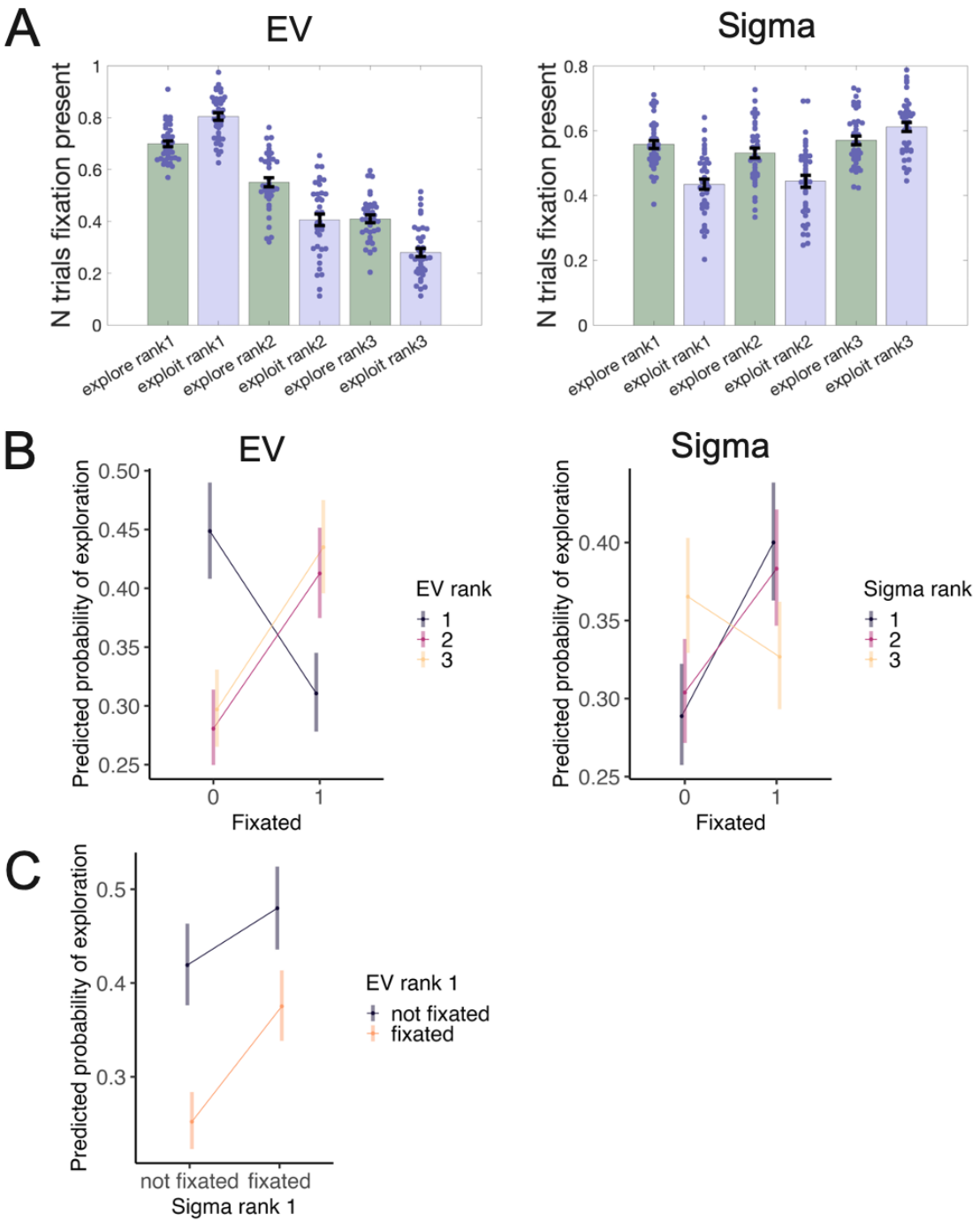


Figure 4-3. Gaze reflects computational model*. A – Number of trials on which fixation on a bandit with different EV ranks (left) or sigma ranks (right) was present, expressed as a fraction of total number of trials in each response condition. B – Bandits with different EV ranks (left) or sigma ranks (right) differentially predict trial type depending on whether they were fixated or not. C – interaction between fixating on the bandit with highest expected value (EV rank 1) and bandit with highest uncertainty (sigma rank 1) significantly predicts trial type. Predicted probability: 0 on the y-axis corresponds to 100% probability of exploitation (0% probability of exploration), 1 on the y-axis corresponds to 100% probability of exploration (0% probability of exploitation).*

indicating that exploitation was most probable when the bandit with EV rank 1 was fixated and the bandit with sigma rank 1 was not. Conversely, the highest probability of exploration occurred when the bandit with sigma rank 1 was fixated and the bandit with EV rank 1 was not.

4.3.3 Number of dwell locations differentially predicts trial type

The median number of dwell locations per trial ranged from 1 to 3 (**Figure 4-4**, left) and trials with more dwell locations were increasingly infrequent (**Figure 4-S2**). The number of dwell locations significantly predicted the trial type (OR = 1.34, 95% CI = [1.29, 1.38], p < 2e-16; **Table 4-S4**), showing that exploration became more likely as the number of dwell locations per trial increased (**Figure 4-4**, right). This result pointed to different base rates of exploration across trials, suggesting that trials with different number of dwell locations might capture different characteristics of exploration-exploitation decision-making. To capture different insights that trials with different number of dwell locations might provide into how expected value and uncertainty drive gaze behavior during exploration-exploitation decision-making, we analyzed them separately. These analyses are presented in the following sections.

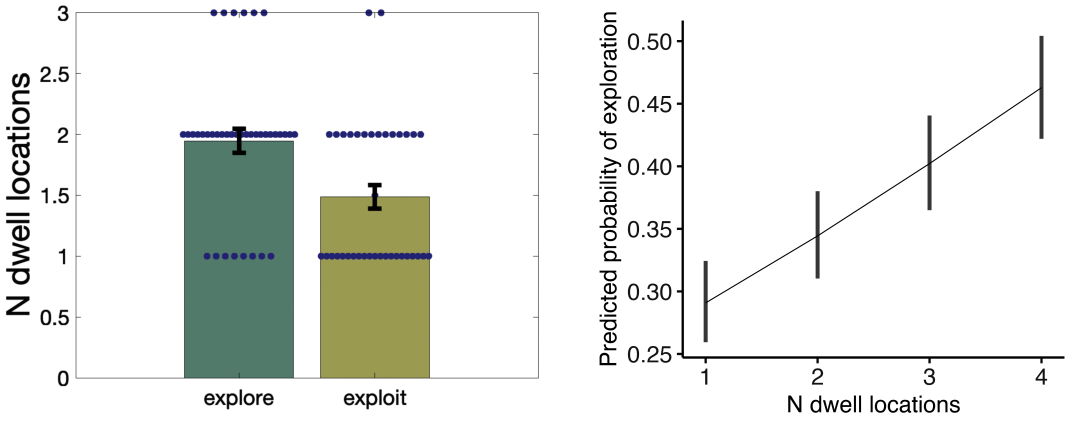


Figure 4-4. Trial type and number of dwell locations. *Left – subjects' median number of dwell locations on exploration and exploitation trials. Right – main effect of n locations in a model predicting trial type. Predicted probability: 0 on the y-axis corresponds to 100% probability of exploitation (0% probability of exploration); 1 on the y-axis corresponds to 100% probability of exploration (0% probability of exploitation).*

4.3.4 Dwell patterns with one dwell location

For separate EV and sigma models based on patterns with 1 dwell location, there was just one fixed-effects predictor, indicating the EV or sigma rank of the fixated bandit (1 – rank 1, 2 – rank 2, 3 – rank 3). A combined model included trial type as the dependent variable and ev_rank (1,2,3 – EV rank of the fixated bandit), sigma_rank (1,2,3 – sigma rank of the fixated bandit), and an interaction between them as independent variables. Dwell patterns thus correspond to the EV/sigma rank of the only fixated bandit (see **Figure 4-5A** for data distributions).

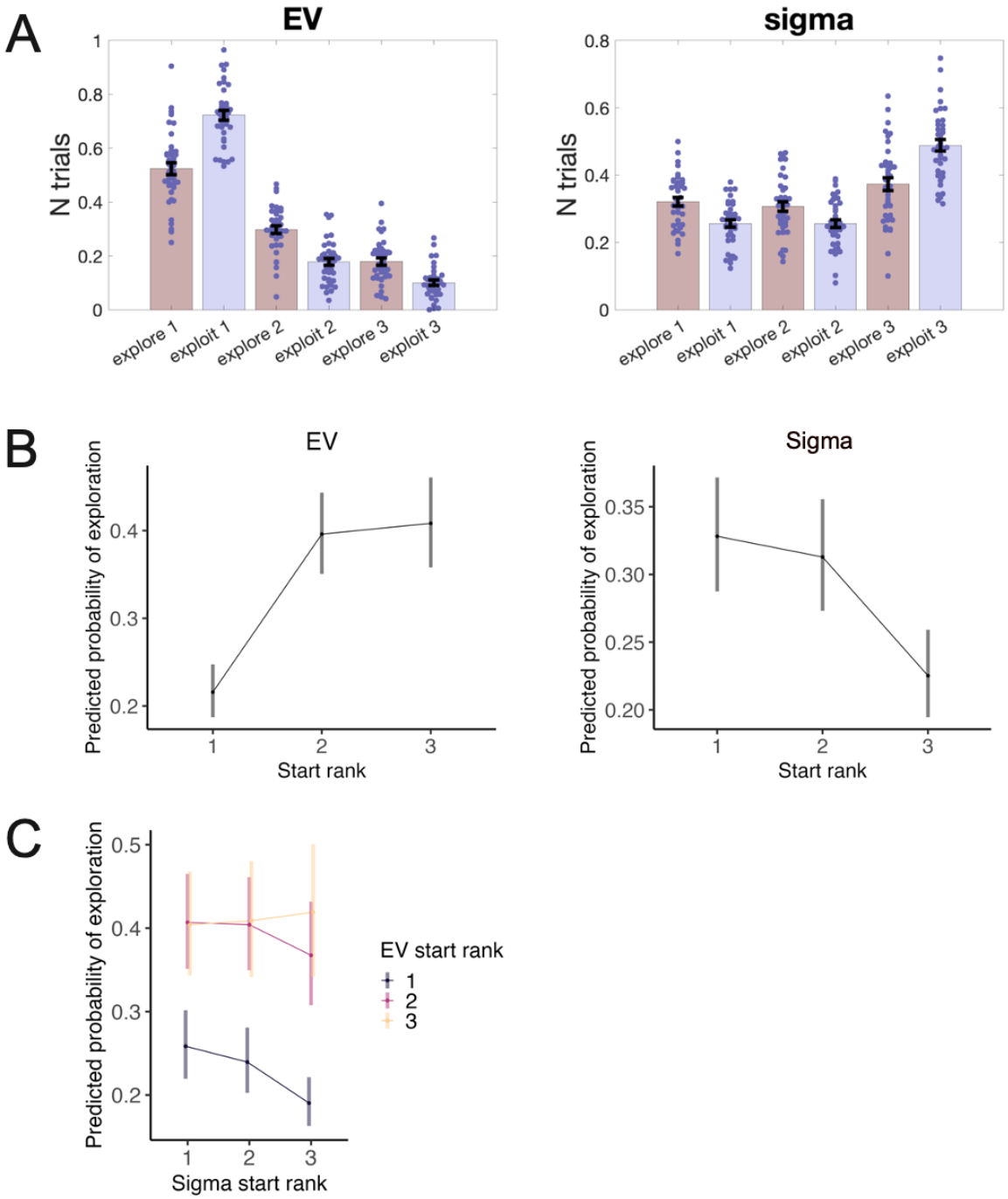


Figure 4-5. Dwell patterns with 1 dwell location. *A – distribution of dwell patterns on trials with 1 dwell location based on EV ranks (left) and sigma ranks (right), expressed as a fraction of total number of trials with 1 dwell location in each response condition. B – main effect of start rank (bandit) in the model based on EV ranks (left) and sigma ranks (right). Note that start bandit is the only fixated bandit on trials with 1 dwell location. C – interaction between patterns based on EV and sigma ranks predicts trial type. Predicted probability: 0 on the y-axis corresponds to 100% probability of exploitation, 0% probability of exploration, 1 on the y-axis corresponds to 100% probability of exploration, 0% probability of exploitation.*

For both EV and sigma, dwell patterns (ranks) differentially predicted the trial type (**Table 4-S5**, **Figure 4-5B**). When the fixated bandit had EV rank 1, the probability of exploitation was very high (close to 80%), while looking at either the EV rank 2 (OR = 2.38, 95% CI = [2.10, 2.70], p < 2e-16) or 3 bandit (OR = 2.51, 95% CI = [2.15, 2.92], p < 2e-16) significantly increased the probability of exploration. In a

model with sigma-based ranks, probability of exploitation was highest when the only fixated bandit was of rank 3 (i.e., when uncertainty was lowest; OR = 0.60, 95% CI = [0.53, 0.67], p < 2e-16), with a significant increase in exploration probability when bandits with sigma rank 1 or 2 were fixated. We also found a significant EV x sigma interaction (OR = 1.58, 95% CI = [1.07, 2.32], p = 0.02; **Table 4-S6**, **Figure 4-5C**). Plotting model results revealed that exploration probability for EV ranks 1 and 2 showed a similar negative trajectory over ranks of sigma (decreasing most at sigma rank 3), while for EV rank 3, a modestly increasing trajectory was noted across sigma ranks.

4.3.5 Dwell patterns with two dwell locations

For models based on patterns with 2 dwell locations, we examined the influence of the start and end bandit (bandit looked at first or last, respectively) on predicting response. Separate logistic regression models were used to avoid overparameterization. Each model had the same form as the model described above for patterns with 1 dwell location.

For both EV, sigma, and their combination (see **Figure 4-6A** for data distribution), logistic regression models predicting the trial type based on the start location (the rank of the first bandit in the pattern) produced no significant results (**Table 4-S7**, **Figure 4-S3**). In contrast, analyses of the end location (**Table 4-S8**, **Figure 4-6B**) revealed that EV rank 1 significantly decreased the chances of exploration compared to EV ranks 2 (OR = 1.68, 95% CI = [1.42, 1.99], p = 1.94e-09) and 3 (OR = 1.80, 95% CI = [1.49, 2.16], p = 5.66e-10). Similarly, sigma rank 1 in the end location significantly increased the probability of exploration compared to ranks 2 (OR = 0.79, 95% CI = [0.66, 0.94], p = 0.009) and 3 (OR = 0.66, 95% CI = [0.55, 0.79], p = 4.97e-06). A combined model with EV- and sigma-based ranks of the end location in a dwell pattern revealed a significant difference between the EV ranks (EV rank 1 vs 2: OR = 1.55, 95% CI = [1.12, 2.15], p = 0.009; EV rank 1 vs 3: OR = 1.83, 95% CI = [1.31, 2.56], p = 0.0004), but no significant main effects of sigma ranks or interactions between EV and sigma ranks (**Table 4-S9**, **Figure 4-6C**).

4.3.6 Dwell patterns with three dwell locations

On trials with 3 dwell locations, xyx patterns were far more prevalent than patterns including all 3 bandits (**Figure 4-7A**). We thus tested a model with predictors xyx (1 = current pattern is of type xyx; 0 = non-xyx pattern), start (1, 2, 3 - the rank of the start bandit), and interaction between them as independent variables. For a model combining EV and sigma, we combined the most prevalent EV-based and sigma-

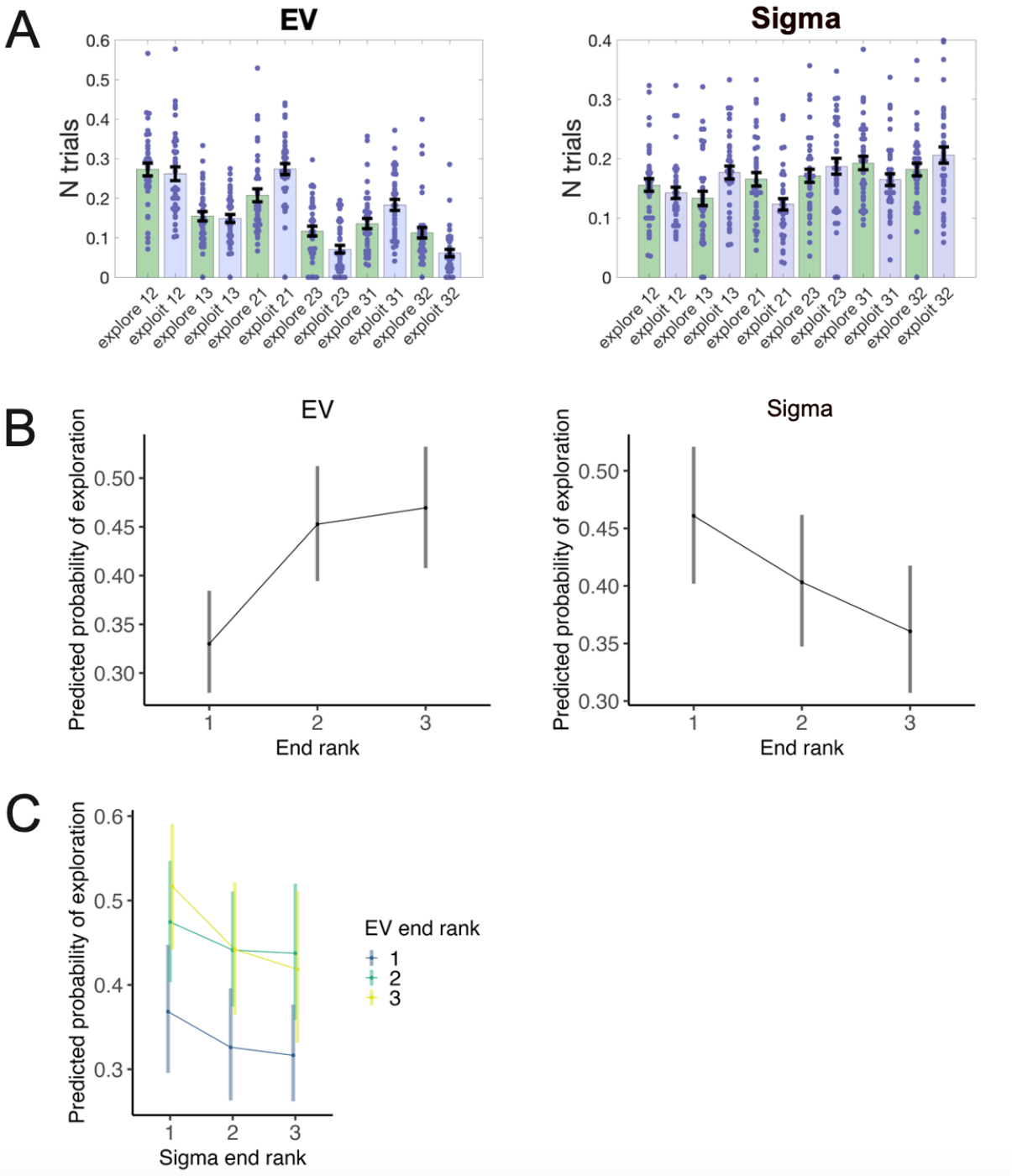


Figure 4-6. Dwell patterns with 2 dwell locations*. A – distribution of dwell patterns on trials with 2 dwell locations based on EV ranks (left) and sigma ranks (right), expressed as a fraction of total number of trials with 2 dwell locations in each response condition. B – main effect of end rank (bandit) in the model based on EV ranks (left) and sigma (right panel). C – no significant interaction between the end rank in patterns based on EV and sigma. Predicted probability: 0 on the y-axis corresponds to 100% probability of exploitation, 0% probability of exploration, 1 on the y-axis corresponds to 100% probability of exploration, 0% probability of exploitation.*

based xyx patterns into one variable each. These were patterns starting and ending with 1 for EV (121, 131) and patterns starting and ending with 3 for sigma (313, 323). We thus created predictors called ev_1y1 (a combination of EV-based 121 and 131 patterns) and sigma_3y3 (a combination of sigma-based 313 and 323 patterns). For both EV- and sigma-based variable, trials with 3+ dwell locations

on which any pattern that was part of the combination occurred were coded as 1, and 0 otherwise. The model also included an interaction term between ev_1y1 and sigma_3y3.

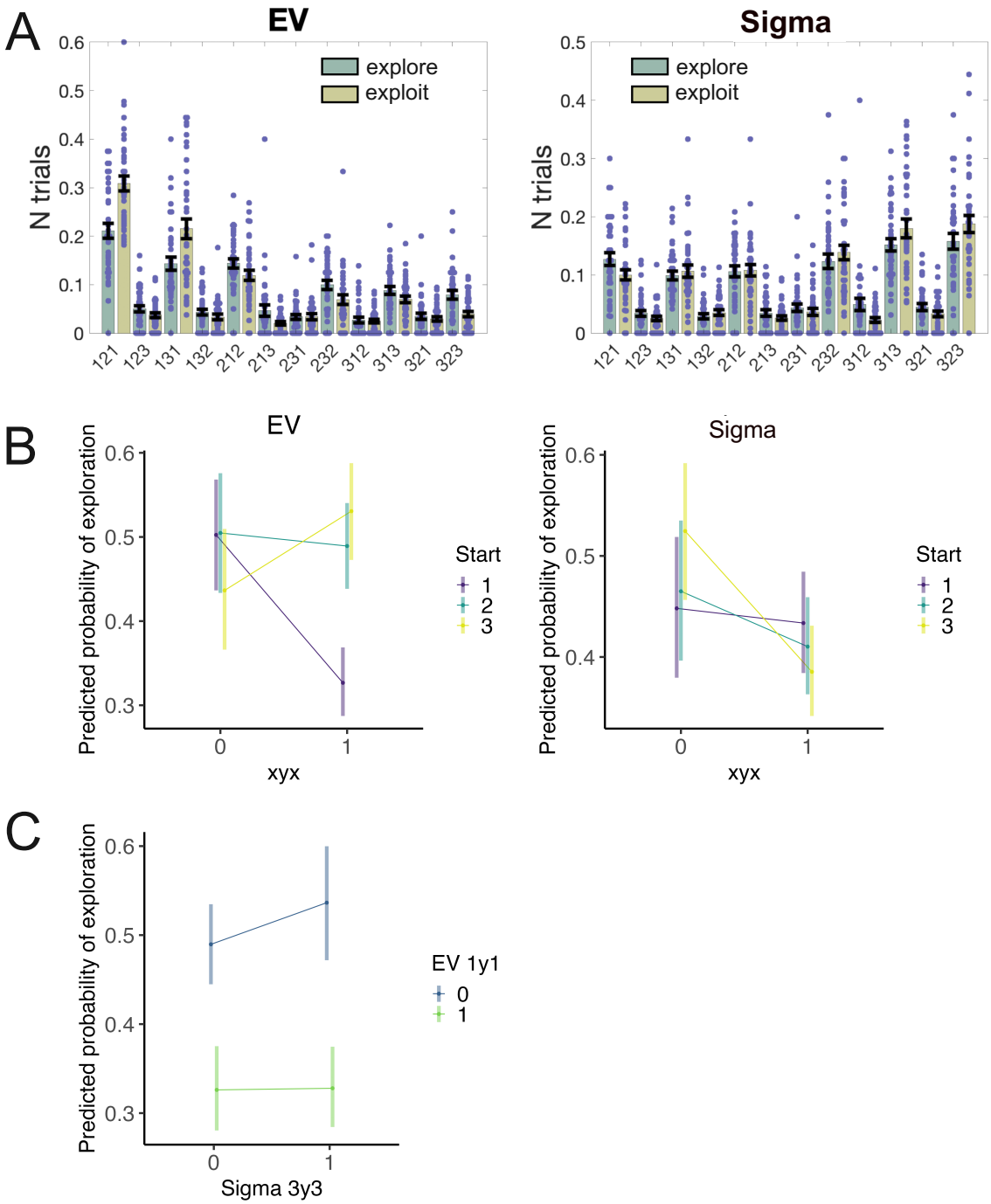


Figure 4-7. Dwell patterns with 3 dwell locations. *A – distribution of dwell patterns with 3 dwell locations based on EV ranks (left) and sigma ranks (right), expressed as the fraction of total number of trials with 3+ dwell locations in each response condition. B – interaction between xyx (0 – non-xyx pattern, 1 – xyx pattern) and the rank of the start bandit (which is the same as the rank of the end bandit in xyx patterns) in both the model based on EV ranks (left) and sigma ranks (right) significantly predicted trial type. C – no significant effect of interaction between most common xyx patterns based on EV and sigma ranks. For this analysis, most common xyx EV-based patterns (121 and 131) were combined into 1y1 predictor and most common xyx sigma-based patterns (313 and 323) were combined into 3y3 predictor. Predicted probability: 0 on the y-axis corresponds to 100% probability of exploitation, 0% probability of exploration, 1 on the y-axis corresponds to 100% probability of exploration, 0% probability of exploitation.*

Significant xyx by start interactions in the EV-based model (**Table 4-S10**, **Figure 4-7B**) indicated that while exploration probability was fairly similar for non-xyx patterns with different start ranks, exploration markedly decreased for xyx patterns that started/ended with 1, while it remained the same and even slightly increased for xyx patterns that started (and ended) with 2 (OR = 1.96, 95% CI = [1.37, 2.80], p = 0.0002) and 3 (OR = 3.04, 95% CI = [2.08, 4.44], p = 1.01e-08). Note that patterns starting with 1 and 3 showed opposite trajectories of exploration probability between non-xyx and xyx patterns. Consequently, the highest exploitation probability was predicted for xyx patterns starting with EV rank 1 and the highest exploration probability was predicted for xyx patterns starting with EV rank 3 (**Figure 4-7B**).

The only significant effect in the sigma-based model (**Table 4-S10**, **Figure 4-7B**) was an interaction showing a much stronger decrease in exploration probability between non-xyx and xyx patterns if the patterns started with rank 3 (OR = 0.60, 95% CI = [0.42, 0.87], p = 0.007), while there was only a slight decrease for patterns starting with ranks 1 and 2. Exploration probability was predicted to be highest for non-xyx patterns starting with sigma rank 3 and lowest for xyx patterns starting (and ending) with sigma rank 3.

A model combining the most prevalent xyx-type patterns based on EV ranks (patterns 121 and 131) and sigma ranks (patterns 313 and 323) showed a significant effect of EV patterns in predicting the trial type, but no significant effect of sigma patterns or interaction between EV and sigma patterns (**Table 4-S11**, **Figure 4-7C**).

4.3.7 Dwell time in each position in a dwell pattern

Since our results above showed that models combining the expected value (EV) and sigma patterns mostly produced significant results only for expected value and not for sigma or their interactions, we used only EV-based patterns for the following analyses of dwell time.

We used mixed-effects logistic regression models for patterns with 2 and 3 dwell locations to predict trial type from dwell time in each position in a pattern. Since patterns with 1 dwell location only have one position, these patterns were not analyzed. For patterns with 2 dwell locations, we ran a model with position (1 – start position in the pattern, 2 – end position in the pattern), dwell time (time spent in the position), and an interaction between them as independent variables. As only the end bandit significantly predicted response type in patterns with 2 dwell locations, we also included EV rank of the end bandit as a predictor. For patterns with 3 dwell locations, the model included position (1, 2, 3 – start, middle, or end position in the pattern, respectively), dwell time, an interaction between position and dwell time, as well as xyx (1 – pattern is of type xyx, 0 – non-xyx pattern) as predictors. The xyx predictor was based on EV rank and was included in the model due to its effect in predicting trial type in patterns with 3 dwell locations shown above.

For patterns with 2 dwell locations, a logistic regression model showed that an interaction between dwell time and position (OR = 1.78, 95% CI = [1.14, 2.78], p = 0.01) significantly predicted the trial type, after accounting for the effect of the EV rank of the end bandit. The longer participants looked at the first bandit, the more the chances of exploration decreased, while the longer they looked at the second bandit, the more chances of exploration increased (**Table 4-S12**, **Figure 4-S4A**).

Similarly, a model predicting the trial type from the dwell time spent in each position in patterns with 3 dwell locations showed a significant dwell time x position 2 interaction (OR = 2.96, 95% CI = [1.70, 5.13], p = 0.0001), after accounting for the effect of whether it was an xyx pattern or not. As participants spent more time in the first dwell location, exploitation became ever more likely, while spending more time in the second position increased the chances of exploration (there was no significant interaction effect between dwell time and position 3; **Table 4-S13**, **Figure 4-S4B**).

4.3.8 Correlations between frequency of dwell patterns and behavioral performance

We then examined possible relationships between frequency of the EV-based dwell patterns with 1, 2, and 3 dwell locations (limited to xyx patterns for the latter) in exploration and exploitation trials and task performance (optimal choice percentage and switch percentage). All correlation results (uncorrected) are listed in **Tables 4-S14 – 4-S16** for dwell patterns with 1-3 dwell locations, respectively.

For patterns with 2 dwell locations, there was a significant negative correlation (**Figure 4-8**) between switch percentage and the frequency of patterns 31 (Pearson's r = -0.35, p = 0.03; Spearman's rho = -0.38, p = 0.02) and 32 (Pearson's r = -0.37, p = 0.02; Spearman's rho = -0.31, p = 0.06) in exploration. In addition, the presence of pattern 31 correlated positively with exploitation percentage (explore 31: Pearson's r = 0.35, p = 0.03; Spearman's rho = 0.35, p = 0.03), while correlation of pattern 32 and exploitation percentage approached significance (explore 32: Pearson's r = 0.29, p = 0.08; Spearman's rho = 0.27, p = 0.11). Together these results indicate that the more participants first looked at the bandit of EV rank 3 but *ended* on either rank 1 or 2, the more they exploited and the less likely they were to switch.

For patterns with 3 dwell locations (only xyx patterns were examined), there was a significant negative correlation (**Figure 4-9**) between the frequency 323 pattern in exploration and optimal choice percentage (Pearson's r = -0.31, p = 0.07; Spearman's rho = -0.40, p = 0.02) as well as switch percentage (Pearson's r = -0.38, p = 0.03; Spearman's rho = -0.39, p = 0.02). Additionally, there was a significant positive correlation with exploitation percentage (Pearson's r = 0.38, p = 0.03; Spearman's rho = 0.35, p = 0.04).

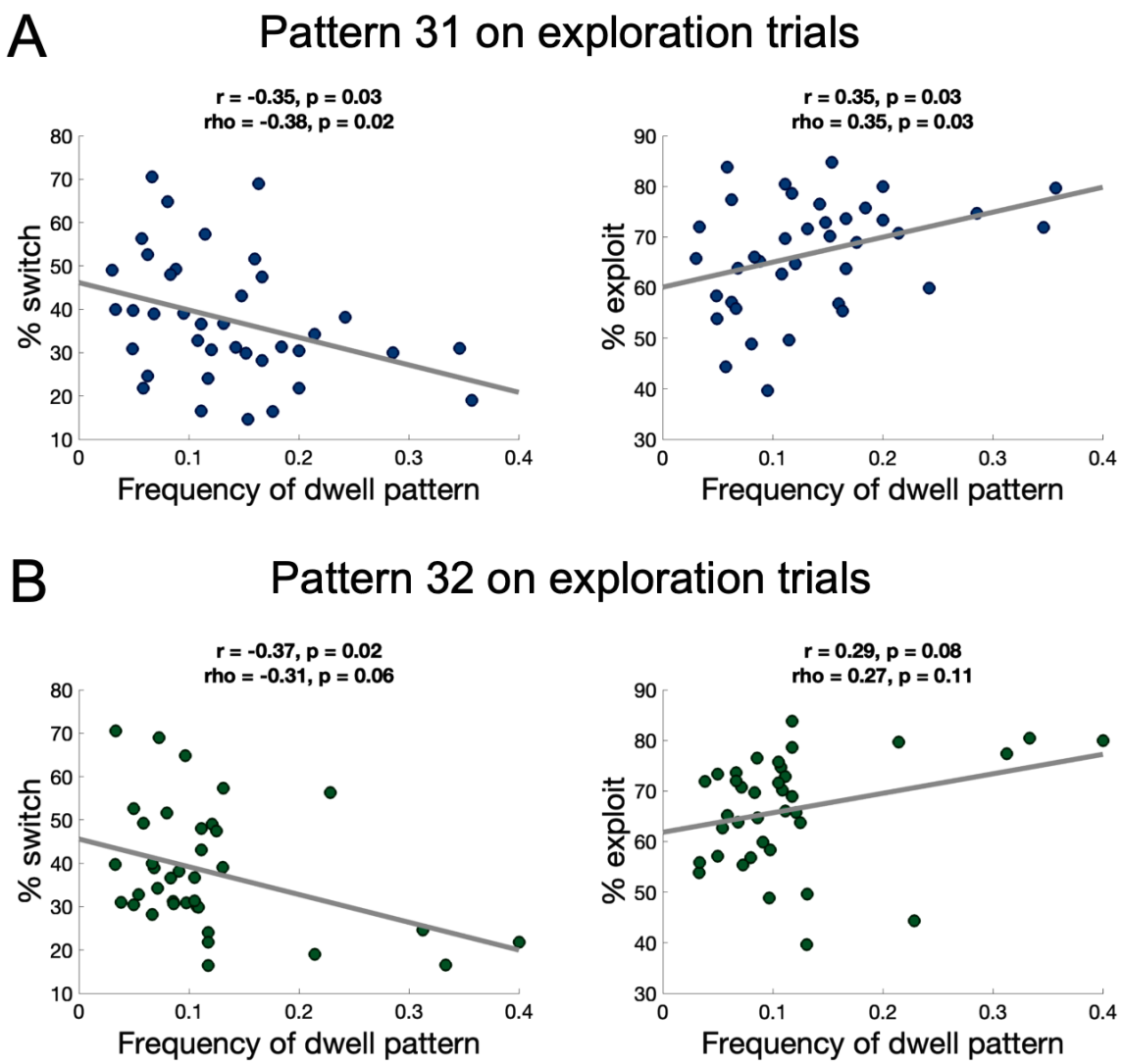


Figure 4-8. Correlations between patterns with 2 dwell locations and behavior. *Correlations between frequency of EV-based patterns 31 (A) and 32 (B) on exploration trials with switch percentage (left) and exploitation percentage (right). Frequency of dwell patterns is expressed as a fraction of total number of trials with 2 dwell locations in exploration and exploitation (separately). r – Pearson's correlation coefficient, rho – Spearman's correlation coefficient, p – p-value.*

These results suggest that as the proportion of EV-based pattern 323 in exploration increased, participants switched less, spending more time exploiting the non-optimal bandit during that time, thus making fewer optimal choices overall.

4.4 Discussion

In the current study, we investigated how gaze during the decision-making period could provide insights into the process behind a decision to explore or exploit. For this purpose, we analyzed fixation-based dwell location patterns defined by the expected value (EV) or uncertainty (sigma) ranks of the bandits in the ExploreExploit task. Our results show the great utility of dwell patterns during the decision-making

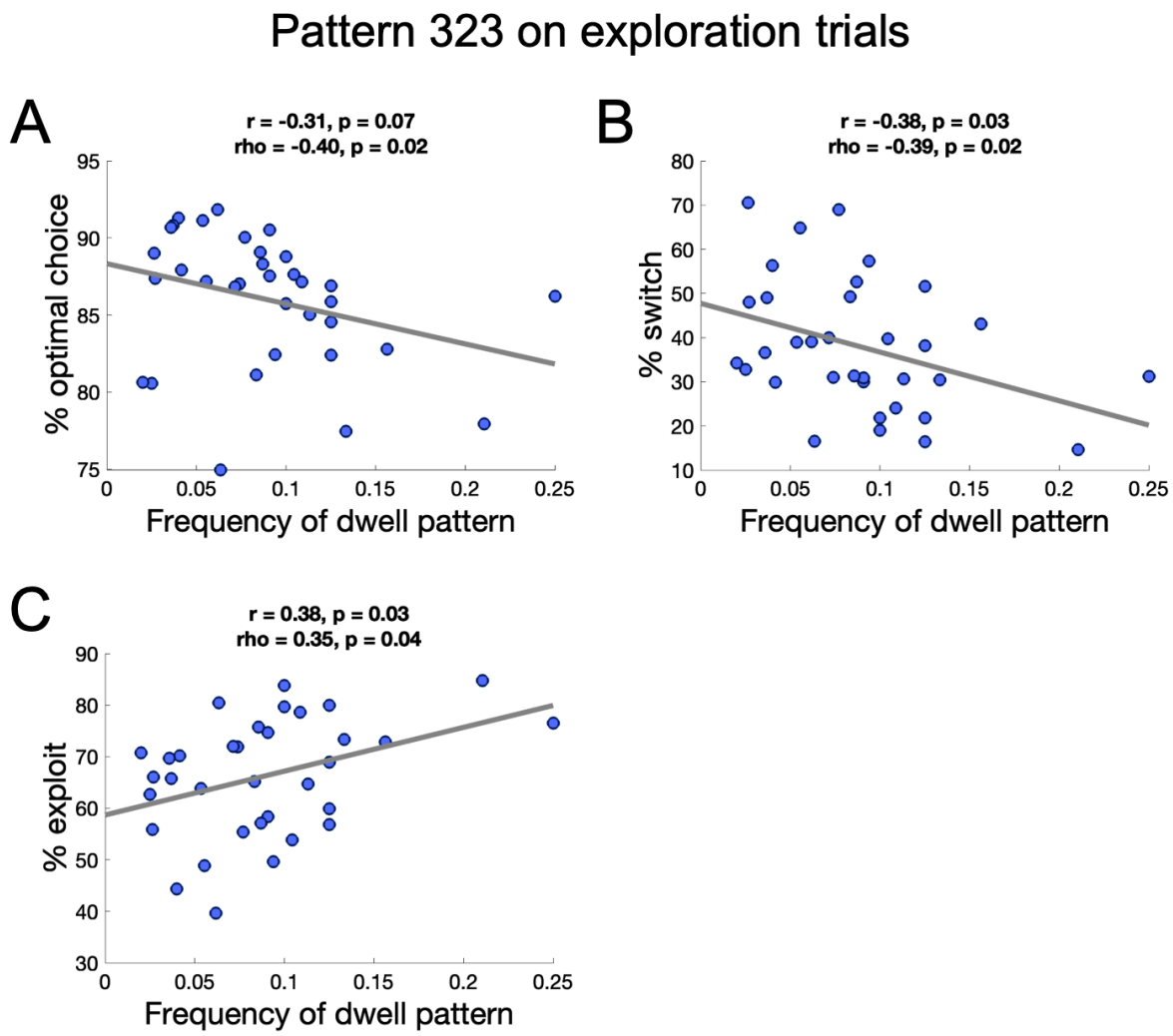


Figure 4-9. Correlations between patterns with 3 dwell locations and behavior. *Correlations between frequency of EV-based pattern 323 on exploration trials and optimal choice percentage (A), switch percentage (B), and exploitation percentage (C). Frequency of dwell patterns is expressed as a fraction of total number of trials with 3+ dwell locations in each response condition. r – Pearson's correlation coefficient, rho – Spearman's correlation coefficient, p – p-value.*

period for predicting a response (exploration or exploitation) and their association with task performance (optimal choice percentage).

4.4.1 Using gaze as a veridical signature of the decision-making process

Commonly used paradigms in value-based decision-making studies that have been combined with eye tracking often require participants to collect visual information about the available options before making a choice (Gluth et al., 2020; Krajbich & Rangel, 2011; Thomas et al., 2021). Not only does this information-gathering process at least partially determine the eye movements in the beginning of the trial (Spering, 2022), it cannot easily be separated from the processes of evaluation and option comparison throughout the trial. This inherent combination of perceptual information-collection and decision-making elements is reflected in the use of evidence accumulation models in such designs

(Krajbich & Rangel, 2011; Thomas et al., 2021). Furthermore, work on how gaze allocation is related to the subjective value of fixated and not-fixated items is still ongoing and presents a complex picture, highlighting the difficulty of interpreting the underlying decision-making mechanisms related to observed gaze behavior. For example, while it was traditionally assumed that the value of fixated items increased, Sepulveda and colleagues (2020) showed that participants fixated more on the item they intended to choose, regardless of whether they had to choose the item they preferred most or the one they preferred least. Their results indicate that allocation of visual attention does not lead to increased value *per se*, but instead modulates goal-directed evidence accumulation (Sepulveda et al., 2020).

In contrast to such paradigms, the advantage of the ExploreExploit task is that there is no need to look at the options prior to choice because they are always fixed. This feature makes it particularly well suited to be paired with gaze data related to the underlying decision-making process. Accordingly, the presence of strong effects of the chosen bandit – both in how often and for how long it was fixated – lends particularly convincing evidence that the gaze data in our task reflects choice rather than evidence accumulation, and can thus be considered a viable signature of the underlying decision-making process.

4.4.2 Gaze behavior reflects parameters of our computational model of exploration-exploitation decision-making

Beyond previous research showing that both EV and uncertainty drive fixations in a value-based decision-making task (Callaway et al., 2021), we show that participants' decision to explore or exploit was determined by fixating on bandits with the highest expected value (EV) and bandits with the highest uncertainty (sigma). The particular importance of high EV in predicting exploitation trials and of high sigma in predicting exploration trials was highlighted by both the interaction analysis and separate analyses of EV and sigma ranks predicting the trial type. These results are in line with the general conceptualization of exploitation as a behavior focused on reward maximization and of exploration as a behavior driven by uncertainty reduction (Blanchard & Gershman, 2018; Sutton & Barto, 1998), as is also implemented in the reinforcement learning model used in the current study.

Similarly, the results of dwell pattern analyses with 1, 2, and 3 dwell locations unanimously point to the unique ways that one may fixate on high-reward options during exploitation and on bandits with high uncertainty during exploration. When examining patterns with 2 and 3 dwell locations, there were significant effects of EV, but no significant effects of sigma or interactions. In contrast, there was an interaction between EV and sigma in a combined model for patterns with 1 dwell location. This interaction highlighted the divergence of trajectories of EV ranks for the lowest uncertainty bandit (sigma rank 3) in predicting the trial type, suggesting that EV might play the primary role in determining the trial type, while sigma could play a more subtle role in driving the decisions. The relative importance of the expected value might indicate that gaining information during exploration serves the purpose of gaining reward during exploitation, thus highlighting that two modes are connected and part of the same behavioral complex that allows participants to optimally function in their environment.

Taken together, these findings demonstrate that gaze behavior during the decision-making period provides an observed marker of the decision-making process behind the choice to explore or exploit, in line with how this process is formulated in our computational model.

4.4.3 Trials with different numbers of dwell locations provide complementary insights into the explore-exploit decision-making process

Our results suggest that patterns with 1, 2, and 3 dwell locations might highlight qualitatively different features of exploration-exploitation decision-making. Specifically, trials with just 1 dwell location – the most prevalent type of trial for most participants – seem to reflect decisions with a strong focus on a particular option, as indicated by the probability of exploitation of ~80% on trials on which the bandit with the highest EV was fixated. This is consistent with results showing that in a value-based choice task with three alternatives, the number of fixations was lower when the difference in value between alternatives was higher (and the choice was consequently easier) (Callaway et al., 2021; Krajbich et al., 2010) and might reflect that participants were relatively confident in their choice already at the beginning of the decision-making period. On trials with two dwell locations, only the end location (bandit) plays a key role in predicting the trial type, reminiscent of the choice bias towards the last fixated alternative shown by previous value-based decision-making studies (Callaway et al., 2021; Krajbich et al., 2010; Thomas et al., 2021). Such trials might reflect decisions that started with less confidence in which option should be chosen and a preference was formed by the end of the trial. Patterns with three dwell locations highlight a particular importance of the xyx pattern type (pattern in which the first and the last bandit are the same) for predicting the decision to explore or exploit. This result could indicate that these trials might have needed more deliberation to make a decision, which was achieved using repeated gaze allocation as a mechanism supporting the comparison of specific alternatives (Russo & Rosen, 1975; Thomas et al., 2021).

Although examining three dwell locations allows one to capture whether participants looked at all available bandits, participants indeed did *not* look at all three locations prior to choice on the majority of trials. This is a result that is derivable due to the design of the ExploreExploit task, which does not require participants to collect visual information about which options are available to make a choice. As most paradigms in the field would do, requiring evidence accumulation of choice options on a given trial would necessarily force fixations towards all available options, thus disallowing subjects to behaviorally reveal that they are *not* considering a given option.

All types of examined patterns emphasize the importance of fixating on the bandit with the highest expected value for predicting an exploitation response and on the importance of fixating on the bandit with the highest uncertainty for predicting an exploration response. However, the interaction between xyx pattern type and EV (or sigma) rank in patterns with 3 dwell locations reveals a novel and nuanced picture. Beginning and ending a pattern with a bandit with the highest EV (patterns 121 and 131) might represent a gaze behavior that is particularly strongly associated with exploitation. At the same time,

the sigma-based model highlights the differential role of starting a pattern with the bandit with the lowest uncertainty (sigma rank 3) for deciding whether to explore or exploit, based on whether it's an xyx pattern or not. These results suggests that, while the expected value seems to be the driving force behind choosing the response, uncertainty might have a more subtle role in exploration-exploitation decision-making. Previous research also showed that in large item sets in which not every item could be fixated, participants did not look at items sequentially, but instead compared options by shifting their gaze back and forth between them (Thomas et al., 2021). Similarly, Russo and Rosen (1975) suggested that participants use binary processing (reflected in fixation patterns of type xyx or xyxy…) as a strategy to compare alternatives in a multi-alternative value-based choice task. Such alternating gaze behavior possibly allows participants to focus their attention and decision-making resources on the options they are currently considering and might serve to arrive at a decision more efficiently by considering pairs of competitor options and eliminating one of them. This could be a possible mechanism explaining higher prevalence of xyx patterns, especially patterns beginning and ending with EV rank 1 (highest expected value) in our data, as well as a particularly important role of EV rank 1 patterns for predicting exploitation responses. Of note, evaluating more dwell locations could reveal further patterns than were possible in our current design (such as xyxy); as such, future studies might extend the duration of the decision-making period to allow for more dwell locations to be visited.

Our results indicate that the temporal order of dwell locations during a decision-making period provides a detailed decision-making signature (beyond more general measures such as fixation count and dwell time), which allows one to predict the trial response and shows an association with task performance. Our study allows future research to produce specific hypotheses as to how gaze can provide insights into the decision-making process behind exploration and exploitation, as well as to potentially use gaze signatures for an real-time prediction of behavior (Deng et al., 2020; cf. Ji et al., 2004)

4.4.4 Participants who perform worse use typical gaze strategies but apply them to an incorrect model of the reward structure

Correlating frequency of dwell location patterns with behavioral performance revealed that patterns emphasizing dwelling on options with the lowest expected value (EV) on exploration trials were associated with exploiting more, switching less, and choosing the optimal bandit less often. Both patterns with 2 (31 and 32) and 3 (323) dwell locations support the interpretation that worse performers may often apply typical gaze strategies to an incorrect mental picture of the reward structure. Although it is reassuring that all significant correlations point in the same direction of interpretation, emphasizing that dwelling on bandits in the lower range of the reward structure is detrimental to performance, caution is warranted in interpreting these findings, since these dwell patterns were observed on a relatively small number of trials (relative to other xyx patterns, especially those beginning with one) and correlations were exploratory. Replication studies are needed to confirm their robustness. Of note, behavioral performance in our task was very high and trials on which the aforementioned gaze patterns occurred were genrally rare. To capture the whole performance spectrum better, future studies could contrast more and less difficult task blocks (e.g. by manipulating bandit similarity) and thus shed more light on

the association between gaze patterns including lower EV ranks (e.g. 323) and worse task performance (lower optimal choice percentage).

4.4.5 Summary

Our study provides a first account of the association between gaze during explore-exploit decision-making and how it reflects task performance. Our results emphasize the importance of looking at the bandit with the highest expected value for exploitation and of looking at the bandit with the highest uncertainty for exploration. We also find that the expected value of the bandits could be the driving force behind gaze behavior, while uncertainty might have a more subtle role. We further show that trials with different numbers of dwell locations (fixated bandits) provide complementary insights into exploration-exploitation decision-making, highlighting how option features might have been considered in the decision-making process. Crucially, our findings reveal that it is both *where* (EV/sigma ranks of the bandit) and *how* (dwell pattern type) participants look at options that determine their decision to explore or exploit. Lastly, we show that participants who perform worse, use the same forms of gaze strategy as better-performing participants, but appear to have an erroneous understanding of the reward structure prior to choice. Our study offers a key starting-point for future studies to formulate specific hypotheses regarding gaze behavior during exploration-exploitation decision-making, while providing directly observable insights that neither button presses nor computational modeling alone can provide.

4.5 References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. *ArXiv*. https://doi.org/10.48550/arxiv.1406.5823

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

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/10.3758/s13415-017-0556-2

Byrne, S. A., Reynolds, A. P. F., Biliotti, C., Bargagli-Stoffi, F. J., Polonio, L., & Riccaboni, M. (2023). Predicting choice behaviour in economic games using gaze data encoded as scanpath images. *Scientific Reports*, *13*(1), 4722. https://doi.org/10.1038/s41598-023-31536-5

Callaway, F., Rangel, A., & Griffiths, T. L. (2021). Fixation patterns in simple choice reflect optimal information sampling. *PLoS Computational Biology*, *17*(3), e1008863. https://doi.org/10.1371/journal.pcbi.1008863

Chen, H., Cohen, P., & Chen, S. (2010). How Big is a Big Odds Ratio? Interpreting the Magnitudes of Odds Ratios in Epidemiological Studies. *Communications in Statistics - Simulation and Computation*, *39*(4), 860–864. https://doi.org/10.1080/03610911003650383

Deng, Q., Wang, J., Hillebrand, K., Benjamin, C. R., & Söffker, D. (2020). Prediction Performance of Lane Changing Behaviors: A Study of Combining Environmental and Eye-Tracking Data in a Driving Simulator. *IEEE Transactions on Intelligent Transportation Systems*, *21*(8), 3561–3570. https://doi.org/10.1109/tits.2019.2937287

Gidlöf, K., Wallin, A., Dewhurst, R., & Holmqvist, K. (2013). Using Eye Tracking to Trace a Cognitive Process: Gaze Behaviour During Decision Making in a Natural Environment. *Journal of Eye Movement Research*, *6*(1). https://doi.org/10.16910/jemr.6.1.3

Glöckner, A., & Herbold, A. (2011). An eye-tracking study on information processing in risky decisions: Evidence for compensatory strategies based on automatic processes. *Journal of Behavioral Decision Making*, *24*(1), 71–98. https://doi.org/10.1002/bdm.684

Gluth, S., Kern, N., Kortmann, M., & Vitali, C. L. (2020). Value-based attention but not divisive normalization influences decisions with multiple alternatives. *Nature Human Behaviour*, *4*(6), 634–645. https://doi.org/10.1038/s41562-020-0822-0

Huddleston, P. T., Behe, B. K., Driesener, C., & Minahan, S. (2018). Inside-outside: Using eye-tracking to investigate search-choice processes in the retail environment. *Journal of Retailing and Consumer Services*, *43*, 85–93. https://doi.org/10.1016/j.jretconser.2018.03.006

Ialongo, C. (2016). Understanding the effect size and its measures. *Biochemia Medica*, *26*(2), 150–163. https://doi.org/10.11613/bm.2016.015

Jacob, R. J. K., & Karn, K. S. (2003). The Mind's Eye. *Section 4: Eye Movements in Human—Computer Interaction*, *Vision Research391999*, 573–605. https://doi.org/10.1016/b978-044451020-4/50031-1

Ji, Q., Zhu, Z., & Lan, P. (2004). Real-Time Nonintrusive Monitoring and Prediction of Driver Fatigue. *IEEE Transactions on Vehicular Technology*, *53*(4), 1052–1068. https://doi.org/10.1109/tvt.2004.830974

Kloosterman, N. A., & Garrett, D. D. (n.d.). *Tracking of natural-scene feature richness in brain signal variability predicts memory and changes with age*.

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292–1298. https://doi.org/10.1038/nn.2635

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, *108*(33), 13852–13857. https://doi.org/10.1073/pnas.1101328108

Krol, M., & Krol, M. (2017). A novel approach to studying strategic decisions with eye-tracking and machine learning. *Judgment and Decision Making*, *12*(6), 596–609. https://doi.org/10.1017/s1930297500006720

Kümmerer, M., & Bethge, M. (2021). State-of-the-Art in Human Scanpath Prediction. *ArXiv*. https://doi.org/10.48550/arxiv.2102.12239

Lenth, R. V. (2022). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. https://CRAN.R-project.org/package=emmeans

Lüdecke, D. (2023). *_sjPlot: Data Visualization for Statistics in Social Science_. R package version 2.8.15*. https://CRAN.R-project.org/package=sjPlot

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience*, *2011*, 156869. https://doi.org/10.1155/2011/156869

Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research*, *96*(1), 3–14. https://doi.org/10.1080/00220670209598786

Polonio, L., Guida, S. D., & Coricelli, G. (2015). Strategic sophistication and attention in games: An eye-tracking study. *Games and Economic Behavior*, *94*, 80–96. https://doi.org/10.1016/j.geb.2015.09.003

Russo, J. E., & Rosen, L. D. (1975). An eye fixation analysis of multialternative choice. *Memory & Cognition*, *3*(3), 267–276. https://doi.org/10.3758/bf03212910

Sepulveda, P., Usher, M., Davies, N., Benson, A. A., Ortoleva, P., & Martino, B. D. (2020). Visual attention modulates the integration of goal-relevant evidence and not value. *ELife*, *9*, e60705. https://doi.org/10.7554/elife.60705

Spering, M. (2022). Eye Movements as a Window into Decision-Making. *Annual Review of Vision Science*, *8*(1), 427–448. https://doi.org/10.1146/annurev-vision-100720-125029

Stoffel, M. A., Nakagawa, S., & Schielzeth, H. (2021). partR2: partitioning R2 in generalized linear mixed models. *PeerJ*, *9*, e11414. https://doi.org/10.7717/peerj.11414

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning* (Vol. 135). MIT press.

Team, R. C. (2022). *R: A language and environment for statistical computing*. URL https://www.R-project.org/

Thomas, A. W., Molter, F., & Krajbich, I. (2021). Uncovering the computational mechanisms underlying many-alternative choice. *ELife*, *10*, e57012. https://doi.org/10.7554/elife.57012

Zhou, L., Zhang, Y., Wang, Z., Rao, L., Wang, W., Li, S., Li, X., & Liang, Z. (2016). A Scanpath Analysis of the Risky Decision-Making Process. *Journal of Behavioral Decision Making*, *29*(2–3), 169–182. https://doi.org/10.1002/bdm.1943

5. General Discussion

The aim of this dissertation was to examine behavioral, computational, neural, and physiological mechanisms of human exploration-exploitation behavior. In Chapter 2 of this dissertation, I presented a newly designed ExploreExploit task, which combines the benefits of previously used paradigms while avoiding their drawbacks. Using the ExploreExploit task and a computational model that best reflected behavior in this task, I examined fMRI and eye tracking data to shed light on neural and physiological mechanisms underlying exploration-exploitation behavior. In Chapter 3, I showed that BOLD signal variability could function as a neural mechanism that allows to flexibly adapt exploring and exploiting to a changing environment. In Chapter 4, I demonstrated that different gaze patterns during the decision-making period predicted response and provided complementary insights into the decision-making process that lead to a decision to explore or exploit. In this chapter, I summarize contributions of this dissertation to the field of exploration-exploitation research, discuss limitations and outline suggestions for future studies.

5.1 Main contributions of this dissertation to the field of exploration-exploitation research

First, this dissertation expands the toolkit of exploration-exploitation research with the ExploreExploit task. Making use of the key distinction in motivation for exploratory and exploitative actions (receiving information vs. reward, respectively (Blanchard & Gershman, 2018)), the ExploreExploit task allows participants to indicate directly whether they explore or exploit on a given trial, thus ensuring valid trial categorization. In addition, this task possesses a flexible reward structure (Daw et al., 2006) that allows for the manipulation of multiple features of the reward environment (such as reward range, magnitude, probability, wins and losses, volatility, etc.) and for one to adjust the task structure itself (increase or decrease number of bandits, incorporate different task horizons, novel options, etc.). At the same time, our task allows participants to explore or exploit on any trial as they see fit (unlike paradigms forcing exploration on certain trials by providing a novel option (e.g. Hogeveen et al., 2022) or by manipulating uncertainty levels of the options (e.g. Wilson et al., 2014)). The ExploreExploit task thus captures natural exploration-exploitation behavior. In addition, the best-fitting computational model presented in this dissertation adapts commonly used reinforcement learning models (Daw et al., 2006; Sutton & Barto, 1998) to a task in which information and reward feedback are strictly separated, thus laying the grounds for computational modeling of similar tasks. All in all, the ExploreExploit task not only captures key behavioral features of exploration and exploitation, but also does so in multiple experimental settings, proving itself a robust paradigm capable of producing replicable results.

Further, this dissertation demonstrates that uncertainty-driven BOLD signal variability could potentially be a neural mechanism behind flexibly adapting exploration-exploitation behavior to a changing environment. Exploration-exploitation studies typically examined different types of uncertainty only in the context of their relation to different exploration types, such as relative and total uncertainty

representing directed and random exploration (Gershman, 2018; Tomov et al., 2020), or studies concentrate only on one uncertainty measure (e.g. choice entropy (Muller et al., 2019)). This dissertation is the first to systematically examine the role of different types of uncertainty (prior and posterior estimation uncertainty, choice uncertainty) in flexibly adapting exploration-exploitation behavior to a changing environment. I show that BOLD signal variability was most robustly related to posterior estimation uncertainty, which reflects uncertainty about the reward structure participants have after they make a choice and receive feedback about it, and was the only uncertainty type that reflected uncertainty changes unique to exploration and exploitation modes. This dissertation thus demonstrates a strong link between uncertainty changes and BOLD signal variability during exploration-exploitation decision-making, in line with the idea that higher uncertainty levels in the environment require a more variable neural system to be prepared to adapt to a larger number of possible "states" of the world (Grady & Garrett, 2018; Waschke et al., 2021). Furthermore, my results highlight brain regions (like the thalamus and the insula) that are likely involved in uncertainty processing (Bach & Dolan, 2012) and form the neural basis for flexibly adapting behavior to changes in the environment (Shine et al., 2016), thus emphasizing the importance of these functions for exploration-exploitation decision-making and an important role of BOLD signal variability in supporting these functions.

Both higher levels (Armbruster-Genç et al., 2016; Garrett et al., 2011; Grady & Garrett, 2018) and stronger modulation (Garrett et al., 2013, 2015, 2020; Skowron et al., 2024) of BOLD signal variability have been shown to be beneficial for performance in a vast array of tasks. My results contribute to a more nuanced understanding of this relationship. In my task, a higher level of BOLD signal variability in both exploration and exploitation was associated with better performance, as was shown in multiple previous studies (CITE). However, stronger variability modulation only during exploration was associated with better and more flexible performance. Contrary to our expectation, during exploitation, less strong modulation of BOLD variability was related to better and more flexible performance. This finding might suggest that stronger BOLD variability modulation is not beneficial for behavioral performance per se, but instead the strength of the modulation adapts to meet the task demands. Specifically, while stronger modulation of BOLD variability in exploration possibly reflected a more efficient learning process (Skowron et al., 2024), stronger modulation of variability in exploitation was related to staying longer (possibly too long) in exploitation mode. Though participants who showed more flexible behavior modulated variability in exploitation less, they nevertheless had higher levels of brain signal variability overall, potentially indicating a more variable neural system that allowed them to switch out of exploitation mode faster. As has been shown in a study that pharmacologically modulated BOLD variability in younger and older participants, the relationship between the increase of variability and performance might not be linear and pharmacologically increasing some participants' BOLD signal variability too much might have had an adverse effect on performance (Garrett et al., 2015). This dissertation thus contributes to the understanding of a complex relationship of BOLD signal variability changes and behavior.

Further, this dissertation is the first to examine the mechanisms of exploration-exploitation decision-making using gaze data. In line with the existing literature reporting that scan paths (fixation-based gaze

patterns) during the decision-making phase can differentiate between optimal and sub-optimal responses (Byrne et al., 2023; Krol & Krol, 2017; Polonio et al., 2015), I show that gaze patterns during the decision-making period robustly predicted whether participants would explore or exploit. To date, eye tracking studies have only used pupillometry to examine the role of uncertainty and expected value in exploration-exploitation decision-making, showing that higher levels of total uncertainty were associated with higher pupil diameter (Fan et al., 2023) and that pupil dilation prior to choice correlated with the value belief of the chosen option (Slooten et al., 2018). Extending this literature, this dissertation sheds light on how the expected value and uncertainty of the options might drive a decision to explore or exploit, revealed through eye tracking; gaze patterns indeed reflected the prominent influence of the expected value in exploitation and uncertainty in exploration (as is formulated in the computational model used in this study). Furthermore, my results show that participants who performed worse, applied typical gaze patterns to options in the lower part of the reward spectrum. This dissertation emphasizes the potential of the gaze data during the decision-making period to predict subsequent exploration or exploitation response based on the expected value and uncertainty of the options, as well as the relationship between gaze patterns during the decision-making period and behavioral performance.

Crucially, in contrast to experimental paradigms used in eye tracking studies on value-based decision-making (Byrne et al., 2023; Callaway et al., 2021; Polonio et al., 2015; Thomas et al., 2021), the ExploreExploit task does not require participants to collect visual information about the options to make a choice (and thus does not require to look at all options on each trial). The gaze data can thus reflect which options are considered for a choice on a given trial and provide a veridical signature of the decision-making process (Spering, 2022; Thomas et al., 2021). This is another useful feature of the ExploreExploit task that makes it a valuable tool for eye tracking research in exploration-exploitation domain. This dissertation shows that the number of bandits that were looked at during the decision-making period might reflect different characteristics of the decision-making process, such as being relatively confident in the choice at the beginning of the decision-making period or rather forming a decision by the end of it. This result was most prominent when only one bandit was looked at and it had the highest expected value (in this case the probability of the response being exploitation reached 80%). In addition, when more deliberation is needed, comparing options in pairs might be the strategy allowing to arrive at a decision most efficiently (Russo & Rosen, 1975; Thomas et al., 2021). Taken together, these results extend our understanding of the decision-making process in an exploration-exploitation task.

5.2 Limitations

The ExploreExploit task contains multiple important features that allow for a thorough investigation of exploration-exploitation behavior. However, no task design is without limitations. In the following, I discuss them and make suggestions for how they could be addressed in future research.

For example, the use of optimal choice percentage as a measure of task performance is necessary and valuable in our study, because, in contrast to the obtained reward percentage, it can indicate that

participants (1) understood how to do the task, (2) diligently did the task, and (3) learned and could follow the reward structure. See *Optimal choice as a measure of task performance* in Supplementary Methods of Chapter 2 for a detailed discussion. However, using optimal choice to assess task performance also entails certain costs, as outlined below.

First, optimal choice percentage does not reflect the trade-off between exploration and exploitation (reward percentage would reflect the trade-off, but it would punish exploration very strictly due to no reward on exploration trials). It is thus possible that, even with the smallest number of exploitation trials, participants could obtain a very high percentage of optimal choice, if they exploited the highest-paying bandit on those trials. In this case, one might think that such data sets reflect suboptimal decision-making processes and don't possess sufficient quality to be included in the analysis. However, in addition to benefits of optimal choice as a measure of task performance discussed in Chapter 2, it should be pointed out that (1) allowing participants to explore and exploit as they see fit (i.e., natural exploration-exploitation behavior) was one of the key objectives of this task, and (2) to ensure the quality of the data, only data sets with at least 15% of exploration and exploitation trials were included in the analysis. Indeed, participants showed a wide range of exploration-exploitation ratios (see *Results* in Chapter 2). While the majority of the data sets had a high exploitation percentage, there were some that had a majority of exploration trials, and others with a very high switch percentage (often switching between exploration and exploitation). The presence of such unusual exploration-exploitation ratios might even provide an incentive for further research to examine *why* participants showed a specific behavior (e.g. high exploration or switch rates), and *how* these behaviors could be elicited through manipulations of the reward environment.

For example, one might speculate that participants who switched more often (alternating often between exploration and exploitation) perceived the reward structure as strongly volatile, preferred to receive information about changes in rewards more often, and thus opted for not spending multiple trials exploiting to avoid losing track of the reward structure. It is also possible, that participants who showed high exploration rates needed more information to form proper estimates of the reward structure and feel confident enough to switch to exploitation. Personality traits, such as higher risk aversion or lower uncertainty tolerance, might potentially also contribute to higher exploration or switch percentage in the data. Though the task instructions aimed to give the same weight to obtaining the most reward and exploiting the highest-paying bandit (making optimal choices), it cannot be excluded that some participants had high exploration rates because they overweighted the importance of finding the best-paying bandit. Importantly, all data sets included in the analysis showed that participants could learn and follow the reward structure well. Future studies could thus specifically investigate whether interindividual differences in exploration-exploitation and switch ratios are associated with subjective volatility perception or personality traits. Future studies could also use the ExploreExploit task to specifically investigate the possibility to increase/decrease within-person exploration and switch rates by producing reward structures with multiple levels of volatility. Of note, similar work has been done by Speekenbrink and Konstantinidis (Speekenbrink & Konstantinidis, 2015) using a multi-armed bandit task and a computational model to categorize trials as exploration or exploitation. Their results suggested

that a more volatile reward structure might lead to increased switching between bandits (Speekenbrink & Konstantinidis, 2015).

Next, participants were encouraged to find the highest-paying bandit in the instructions. Though this instruction was necessary and beneficial for driving task performance and improving participants' ability to track the reward structure, doing so limited the scope of response strategies to finding the best option and prevented us from observing other strategies that might have been useful in our task. For example, instead of selecting exploration more often or longer in order to find the best-paying bandit and incurring a cost of missing out on reward on each exploration trial, finding a "good enough" option and exploiting it (so-called satisficing, (Simon, 1956)) might have been a more appropriate strategy. (See *Optimal choice as a measure of task performance* in Supplementary Methods of Chapter 2 for a detailed discussion.) In environments, in which choosing an option that does not provide the highest reward results in a fairly high percentage of the maximum reward nonetheless (as is the case in our task), a satisficing strategy might be preferred; it leads to an acceptable level of reward, while accounting for the trade-off between the benefits of further searching for a better option and the costs such search would bring (Gigerenzer & Selten, 2002; Lieder et al., 2017; Payne et al., 1988).

Future research could address this question by specifically designing a reward structure for the ExploreEpxloit task containing blocks in which the best response strategy would be to choose a good enough option (satisficing) and blocks in which the best strategy would be finding the highest-paying bandit. Notably, if only the similarity of rewards provided by different bandits is manipulated to achieve this, an unintended consequence might be that different options become indistinguishable, if their rewards are too similar, or the best option becomes too obvious, if rewards are too far apart. One could thus manipulate not only the reward participants obtain, but also the costs participants incur when not choosing the highest-paying bandit or when spending too much time searching for it. To this end, the number of bandits could be increased to produce a grid structure (for example, 4 x 7 grid (Lieder et al., 2017), 1 x 30 grid and 11 x 11 grid (Wu et al., 2018) have been used in the literature), which would already pose a greater strain on participants' resources. The costs of not following the right response strategy could be adjusted by introducing time pressure (e.g. in the form of a reward loss that increases the longer it takes to make a response), which would favor satisficing, or a punishment in the form of a reward loss that increases the further away the reward of a chosen bandit is from the highest-paying bandit. Importantly, the percentage of obtained reward in our data did not differentiate between participants who could and those who could not do the task (see *Optimal choice as a measure of task performance* in Supplementary Methods of Chapter 2). It is crucial to implement mechanisms to evaluate that participants understand the task and can follow the reward structure. Such a mechanism could be an extensive practice session followed by a test in which participants should reach a certain level of optimal choice responses, thus showing that they have a good understanding of the task and of the reward structure.

Lastly, participants received only reward as feedback on exploitation trials and only information as feedback on exploration trials (which is an essential feature of the task). A total separation of reward

and information feedback allows to delineate participants' motivation to obtain reward and thus choose exploitation vs. motivation to gather information and thus choose exploration. It is therefore beneficial and necessary in the experimental context, allowing for an unambiguous separation of exploration and exploitation trials. However, such a strict separation of reward and information does not reflect real-life decisions, in which feedback includes both information and reward. For example, when trying out a new restaurant (exploration), one still hopes to get (and does or does not get) tasty food (reward) and one learns about the quality of food at that place. On the other hand, when going to a favorite restaurant (exploitation), one can enjoy a good meal (reward), but also receives information whether the quality of food as es expected or different. In case of the ExploreExploit task, a complete separation of information and reward might have influenced participants' behavior in different ways. For example, participants seemed to use exploration most often to receive information about the highest-paying bandit; one-trial exploration sequences (the most prevalent length of exploration sequences) were indeed more often directed to the bandit that was exploited prior to that exploration trial. Because feedback on exploitation trials does not contain any information about the reward of the chosen option, the only way to check for a change in the rewards of the preferred bandit (currently exploited bandit) would be to choose that bandit on an exploration trial. Behavioral results suggest that such "checking" might have been a motivation for a high number of exploration trials directed to the bandit with the highest reward (which was exploited most often). Though checking the preferred (highest-paying) bandit reflects that exploration is driven by a combination of reward and information seeking (as is also formulated in our computational model), the need for checking is dictated by the task design. If participants knew how much reward they received when they exploited a bandit, the distribution of exploration trials to bandits with different reward ranks might favor the bandit with the second highest reward (based on a combination of highest reward and highest uncertainty).

5.3 Future directions

This dissertation presents several novel avenues for investigating behavioral, neural, and physiological mechanisms behind exploration-exploitation decision-making, providing a fruitful basis for future studies. In the following, I suggest analyses that could be done in the future to advance our understanding of exploration-exploitation behavior.

For example, to show that computational model-derived estimates of uncertainty and BOLD signal variability truly change in the same direction during exploration-exploitation decision-making, future studies could investigate whether a positive relationship between uncertainty and BOLD signal variability holds in a task in which uncertainty decreases with each exploration trial, but – in contrast to the ExploreExploit task – does not grow during exploitation trials. In changing environments, prior information about rewards is guaranteed to become more and more obsolete the longer the options are not explored (Cohen et al., 2007). In the current study, this is the case for all options during exploitation because the ExploreExploit task does not present information as feedback on exploitation trials. If BOLD signal variability indeed follows uncertainty levels during exploration-exploitation decision-making, the effects should be different in tasks with a static, deterministic environment. A task with a deterministic

reward structure, in which the best option remains best throughout the course of a set number of trials (cf. Navarro et al., 2016), requires a short initial period of exploration to learn about the rewards of all options and determine the highest-paying one (resolve initial uncertainty), followed by exploitation until the end of the block to collect as much reward as possible. In such a task, BOLD signal variability should decrease during initial exploration (as uncertainty decreases) and remain at a constantly low level during the following exploitation period (uncertainty level should remain constantly low since the environment is known and guaranteed not to change). Because uncertainty about the expected rewards doesn't grow to provide a new incentive for exploration, no further adaptation is necessary in the course of a block in a task with a static, deterministic reward environment. The reward structure of the ExploreExploit task could be modified to include blocks with deterministic rewards and blocks with changing rewards, and thus specifically examine whether BOLD signal variability is coupled to the uncertainty regardless of the environment. Based on the findings of this dissertation, one might expect deterministic blocks to produce weaker effects in brain regions related to behavioral flexibility and uncertainty processing, such as the thalamus and insula (Bach & Dolan, 2012; Shine et al., 2023), reflecting that these functions don't play a primary role in deterministic environments.

Another potential direction for future studies could be the combination of eye tracking and fMRI to investigate the mechanisms behind exploration-exploitation behavior from several angles simultaneously. For example, Liu et al. (2017) combined eye movements analysis (using it to visualize information accumulation supporting memory formation) with fMRI analysis of the hippocampus and oculomotor network activity. In the same fashion, future exploration-exploitation studies could combine gaze pattern analyses and variability-based fMRI analyses, focusing on key areas for uncertainty processing and behavioral flexibility (but also on frontal, temporal and parietal regions that prominently featured in the results of the current study). Such analyses could be particularly useful in combination with deterministic and non-stationary rewards, as the difference in the role of uncertainty in these environments might allow for key inferences about the role of behavioral flexibility for successful adaptation of exploration-exploitation behavior to the environment. Of note, a combined analysis of eye tracking and fMRI data is already possible with the dataset from the current study (except that it does not contain deterministic task blocks). However, though undoubtably informative, such analyses were outside the scope of my dissertation, in which I focused on establishing key links between exploration-exploitation behavior and BOLD signal variability (fMRI) and gaze data (eye tracking) alone.

5.4 Conclusion

Overall, this dissertation presents a comprehensive investigation of exploration-exploitation behavior in human participants. Addressing fundamental issues of task design at the outset, this dissertation ensures a new foundation for future work on the behavioral, computational, neural and physiological mechanisms underlying exploration-exploitation decision-making.

5.5 References

Armbruster-Genç, D. J. N., Ueltzhöffer, K., & Fiebach, C. J. (2016). Brain Signal Variability Differentially Affects Cognitive Flexibility and Cognitive Stability. *The Journal of Neuroscience*, *36*(14), 3978–3987. https://doi.org/10.1523/jneurosci.2517-14.2016

Bach, D. R., & Dolan, R. J. (2012). Knowing how much you don't know: a neural organization of uncertainty estimates. *Nature Reviews Neuroscience*, *13*(8), 572. https://doi.org/10.1038/nrn3289

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/10.3758/s13415-017-0556-2

Byrne, S. A., Reynolds, A. P. F., Biliotti, C., Bargagli-Stoffi, F. J., Polonio, L., & Riccaboni, M. (2023). Predicting choice behaviour in economic games using gaze data encoded as scanpath images. *Scientific Reports*, *13*(1), 4722. https://doi.org/10.1038/s41598-023-31536-5

Callaway, F., Rangel, A., & Griffiths, T. L. (2021). Fixation patterns in simple choice reflect optimal information sampling. *PLoS Computational Biology*, *17*(3), e1008863. https://doi.org/10.1371/journal.pcbi.1008863

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/10.1098/rstb.2007.2098

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. https://doi.org/10.1038/nature04766

Fan, H., Burke, T., Sambrano, D. C., Dial, E., Phelps, E. A., & Gershman, S. J. (2023). Pupil Size Encodes Uncertainty during Exploration. *Journal of Cognitive Neuroscience*, *35*(9), 1508–1520. https://doi.org/10.1162/jocn_a_02025

Garrett, D. D., Epp, S. M., Kleemeyer, M., Lindenberger, U., & Polk, T. A. (2020). Higher performers upregulate brain signal variability in response to more feature-rich visual input. *NeuroImage*, *217*, 116836. https://doi.org/10.1016/j.neuroimage.2020.116836

Garrett, D. D., Kovacevic, N., McIntosh, A. R., & Grady, C. L. (2011). The Importance of Being Variable. *The Journal of Neuroscience*, *31*(12), 4496–4503. https://doi.org/10.1523/jneurosci.5641-10.2011

Garrett, D. D., Kovacevic, N., McIntosh, A. R., & Grady, C. L. (2013). The Modulation of BOLD Variability between Cognitive States Varies by Age and Processing Speed. *Cerebral Cortex*, *23*(3), 684–693. https://doi.org/10.1093/cercor/bhs055

Garrett, D. D., Nagel, I. E., Preuschhof, C., Burzynska, A. Z., Marchner, J., Wiegert, S., Jungehülsing, G. J., Nyberg, L., Villringer, A., Li, S.-C., Heekeren, H. R., Bäckman, L., & Lindenberger, U. (2015). Amphetamine modulates brain signal variability and working memory in younger and older adults. *Proceedings of the National Academy of Sciences*, *112*(24), 7593–7598. https://doi.org/10.1073/pnas.1504090112

Gershman, S. J. (2018). Uncertainty and Exploration. *Decision*, *6*(3), 277–286. https://doi.org/10.1037/dec0000101

Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. MIT press.

Grady, C. L., & Garrett, D. D. (2018). Brain signal variability is modulated as a function of internal and external demand in younger and older adults. *NeuroImage*, *169*. https://doi.org/10.1016/j.neuroimage.2017.12.031

Hogeveen, J., Mullins, T. S., Romero, J. D., Eversole, E., Rogge-Obando, K., Mayer, A. R., & Costa, V. D. (2022). The neurocomputational bases of explore-exploit decision-making. *Neuron*, *110*(11), 1869-1879.e5. https://doi.org/10.1016/j.neuron.2022.03.014

Krol, M., & Krol, M. (2017). A novel approach to studying strategic decisions with eye-tracking and machine learning. *Judgment and Decision Making*, *12*(6), 596–609. https://doi.org/10.1017/s1930297500006720

Lieder, F., Krueger, P. M., & Griffiths, T. L. (2017). An automatic method for discovering rational heuristics for risky choice. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

Liu, Z.-X., Shen, K., Olsen, R. K., & Ryan, J. D. (2017). Visual Sampling Predicts Hippocampal Activity. *The Journal of Neuroscience*, *37*(3), 599–609. https://doi.org/10.1523/jneurosci.2610-16.2016

Muller, T. H., Mars, R. B., Behrens, T. E., & O'Reilly, J. X. (2019). Control of entropy in neural models of environmental state. *ELife*, *8*. https://doi.org/10.7554/elife.39404

Navarro, D. J., Newell, B. R., & Schulze, C. (2016). Learning and choosing in an uncertain world: An investigation of the explore–exploit dilemma in static and dynamic environments. *Cognitive Psychology*, *85*, 43–77. https://doi.org/10.1016/j.cogpsych.2016.01.001

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive Strategy Selection in Decision Making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 534–552. https://doi.org/10.1037/0278-7393.14.3.534

Polonio, L., Guida, S. D., & Coricelli, G. (2015). Strategic sophistication and attention in games: An eye-tracking study. *Games and Economic Behavior*, *94*, 80–96. https://doi.org/10.1016/j.geb.2015.09.003

Russo, J. E., & Rosen, L. D. (1975). An eye fixation analysis of multialternative choice. *Memory & Cognition*, *3*(3), 267–276. https://doi.org/10.3758/bf03212910

Shine, J. M., Bissett, P. G., Bell, P. T., Koyejo, O., Balsters, J. H., Gorgolewski, K. J., Moodie, C. A., & Poldrack, R. A. (2016). The Dynamics of Functional Brain Networks: Integrated Network States during Cognitive Task Performance. *Neuron*, *92*(2), 544–554. https://doi.org/10.1016/j.neuron.2016.09.018

Shine, J. M., Lewis, L. D., Garrett, D. D., & Hwang, K. (2023). The impact of the human thalamus on brain-wide information processing. *Nature Reviews Neuroscience*, 1–15. https://doi.org/10.1038/s41583-023-00701-0

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138. https://doi.org/10.1037/h0042769

Skowron, A., Kosciessa, J. Q., Lorenz, R., Hertwig, R., Bos, W. van den, & Garrett, D. D. (2024). Neural variability compresses with increasing belief precision during Bayesian inference. *BioRxiv*, 2024.01.11.575180. https://doi.org/10.1101/2024.01.11.575180

Slooten, J. C. V., Jahfari, S., Knapen, T., & Theeuwes, J. (2018). How pupil responses track value-based decision-making during and after reinforcement learning. *PLOS Computational Biology*, *14*(11), e1006632. https://doi.org/10.1371/journal.pcbi.1006632

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and Exploration in a Restless Bandit Problem. *Topics in Cognitive Science*, *7*(2), 351–367. https://doi.org/10.1111/tops.12145

Spering, M. (2022). Eye Movements as a Window into Decision-Making. *Annual Review of Vision Science*, *8*(1), 427–448. https://doi.org/10.1146/annurev-vision-100720-125029

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning* (Vol. 135). MIT press.

Thomas, A. W., Molter, F., & Krajbich, I. (2021). Uncovering the computational mechanisms underlying many-alternative choice. *ELife*, *10*, e57012. https://doi.org/10.7554/elife.57012

Tomov, M. S., Truong, V. Q., Hundia, R. A., & Gershman, S. J. (2020). Dissociable neural correlates of uncertainty underlie different exploration strategies. *Nature Communications*, *11*(1), 2371. https://doi.org/10.1038/s41467-020-15766-z

Waschke, L., Kloosterman, N. A., Obleser, J., & Garrett, D. D. (2021). Behavior needs neural variability. *Neuron*. https://doi.org/10.1016/j.neuron.2021.01.023

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. https://doi.org/10.1037/a0038199

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/10.1038/s41562-018-0467-4

Appendices

A. Supplementary Materials to Chapter 1

Supplementary Tables

Table 1-S1. Reference list of publications listed in Table 1.

Addicott, M. A., Pearson, J. M., Froeliger, B., Platt, M. L., & McClernon, F. J. (2014). Smoking automaticity and tolerance moderate brain activation during explore–exploit behavior. *Psychiatry Research: Neuroimaging*, *224*(3), 254-261.

Badre, D., Doll, B. B., Long, N. M., & Frank, M. J. (2012). Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. *Neuron*, *73*(3), 595-607.

Blanchard, T. C., & Gershman, S. J. (2018). Pure correlates of exploration and exploitation in the human brain. *Cognitive, Affective, & Behavioral Neuroscience*, *18*, 117-126.

Boorman, E. D., Behrens, T. E., Woolrich, M. W., & Rushworth, M. F. (2009). How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron*, *62*(5), 733-743.

Chakroun, K., Mathar, D., Wiehler, A., Ganzer, F., & Peters, J. (2020). Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making. *eLife*, *9*, e51260.

Cockburn, J., Man, V., Cunningham, W. A., & O'Doherty, J. P. (2022). Novelty and uncertainty regulate the balance between exploration and exploitation through distinct mechanisms in the human brain. *Neuron*, *110*(16), 2691-2702.

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.

Dombrovski, A. Y., Luna, B., & Hallquist, M. N. (2020). Differential reinforcement encoding along the hippocampal long axis helps resolve the explore–exploit dilemma. *Nature communications*, *11*(1), 5407.

Hogeveen, J., Mullins, T. S., Romero, J. D., Eversole, E., Rogge-Obando, K., Mayer, A. R., & Costa, V. D. (2022). The neurocomputational bases of explore-exploit decision-making. *Neuron*, *110*(11), 1869-1879.

Laureiro-Martínez, D., Canessa, N., Brusoni, S., Zollo, M., Hare, T., Alemanno, F., & Cappa, S. F. (2014). Frontopolar cortex and decision-making efficiency: comparing brain activity of experts with different professional background during an exploration-exploitation task. *Frontiers in human neuroscience*, *7*, 927.

Muller, T. H., Mars, R. B., Behrens, T. E., & O'Reilly, J. X. (2019). Control of entropy in neural models of environmental state. *eLife*, *8*, e39404.

Tardiff, N., Medaglia, J. D., Bassett, D. S., & Thompson-Schill, S. L. (2021). The modulation of brain network integration and arousal during exploration. *NeuroImage*, *240*, 118369.

Tomov, M. S., Truong, V. Q., Hundia, R. A., & Gershman, S. J. (2020). Dissociable neural correlates of uncertainty underlie different exploration strategies. *Nature communications*, *11*(1), 2371.

B. Supplementary Materials to Chapter 2

Supplementary Methods

Data exclusion criteria

Five out of 52 participants in the lab study were excluded: the data of one participant was incorrectly saved; one participant swapped the response buttons; three participants were identified during the experiment as not understanding the task. In addition, one participant did not understand the task in one block and one participant swapped the response buttons in one block. This block was excluded from the respective participant's data. The final sample thus comprised 47 participants (45 participants with five task blocks and two participants with four task blocks).

As the lab study was the first study using the ExploreExploit task, we used its data to aid defining the inclusion-exclusion criteria for further studies, especially those carried out online, which do not provide a possibility for the experimenter to interact with the subject.

We a-priori selected exploration/exploitation percentage and optimal choice percentage (defined as a percentage of trials on which the highest-paying bandit was exploited relative to the number of exploitation trials in the block) as measures of interest for exclusion criteria. Choosing (almost) exclusively exploitation or exploration indicates poor task comprehension or lack of motivation and makes statistical analyses difficult due to a big mismatch in trial counts. We set the cut-off for the minimum percentage of exploration and exploitation trials to 15%, to ensure that only data sets that include both behaviors are evaluated. Though optimal choice percentage, as a measure of task performance, was a-priori selected as an exclusion criterion, there were multiple possibilities as to where to set the cut-off (16% for six combinations of behavior and bandit, 33% for three bandits, or 50% for two behavior types). Plotting the data of excluded and included participants in the lab study (**Figure 2-S2**; presented below) revealed that, in contrast to other measures, the optimal choice percentage differentiated near perfectly between the included and excluded blocks at 50% of optimal choice. Two included blocks in the lab study fell slightly under the cut-off of 50%. These blocks were kept int the data; they occurred in otherwise well-performing subjects in a block with particularly frequent changes of the best bandit, which resulted in lower optimal choice percentage than what participants typically showed.

The exclusion criteria of at least 15% exploration/exploitation trials and min. 50% optimal choices were applied to the data from the online study. Out of 54 participants tested in the online study, two were excluded: one because of incorrectly saved data, another because of not meeting

exploration/exploitation and optimal choice criteria. In addition, single blocks were excluded in some data sets: two participants had three blocks excluded, four participants had two blocks excluded, and three participants had one block excluded.

The data for each block was evaluated separately, which allowed to exclude single blocks and keep the rest of the dataset. If more than three blocks had to be excluded, the data of the respective participant was excluded entirely.

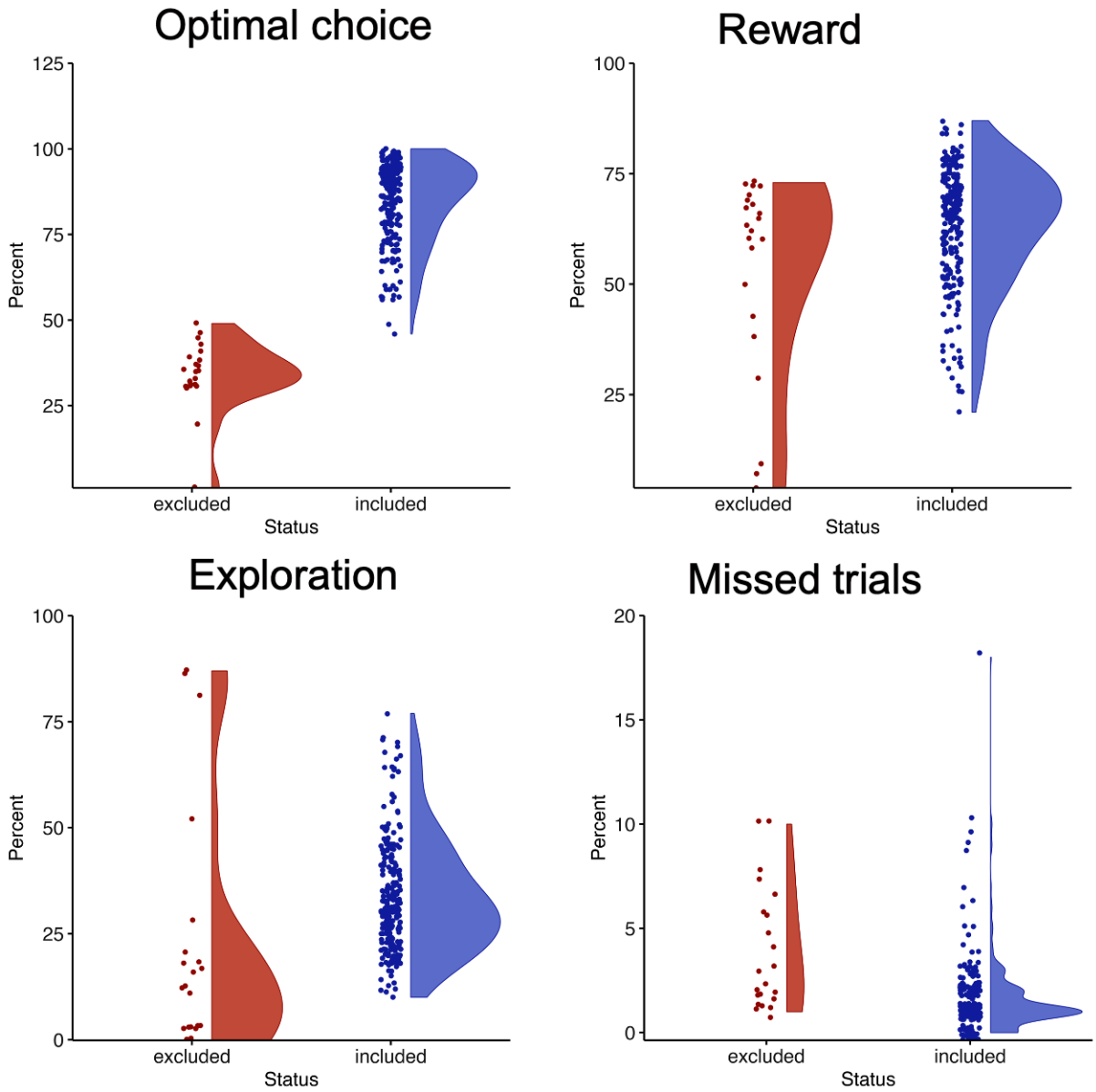


Figure 2-S2. Included vs. excluded blocks in the lab study. *Optimal choice percentage < 50% differentiated between included and excluded runs, while reward, exploration, and missed trials percentage did not. Exploration and reward percentage was calculated based on all trials right after the experimental block was finished.*

Optimal choice as a measure of task performance

Research has shown that participants adaptively choose decision strategies based on the characteristics of the task environment (Gigerenzer & Selten, 2002; Lieder et al., 2017; Lieder & Griffiths, 2015; Payne et al., 1988). Non-compensatory environments, in which the best option can be unambiguously identified, were associated with decision strategies that prioritize finding the best option, while compensatory environments encouraged a satisficing strategy – looking for the best option only until an option that exceeds certain criterion is found (Lieder et al., 2017; Lieder & Griffiths, 2015). In compensatory environments, a satisficing strategy (Simon, 1956) ensures a balance between invested effort (such as time and cognitive resources) and the result, allowing to earn – though not maximal – but high enough payoff (Lieder et al., 2017; Payne et al., 1988).

In the ExploreExploit task, a reward can be earned on every trial by exploiting one of the bandits. Since rewards for each bandit come from independent random walks, the best and second-best rewards on each trial are sometimes markedly different and sometimes quite similar. Choosing a reward that is further away from the maximal reward on one trial, can thus be compensated by choosing a reward that is much closer to the highest reward on another trial. It is thus possible to make random exploitation responses (even exclusively exploitation responses, which is certainly a proper task performance) and amass a considerable amount of maximal reward. Using reward earned as a measure of task performance would make it difficult to assess how well participants understood the task and how diligently they performed it. In line with previous studies in the field (Speekenbrink and Konstantinidis, 2015), we decided to use optimal choice, defined as exploiting the highest-paying bandit, as a measure of task performance. Not selecting the best option on one trial cannot be compensated by selecting it on another, providing a non-compensatory quality to optimal choice. To highlight the importance of this non-compensatory feature of the environment for performing the task, participants were encouraged to look for the best-paying bandit in the task instructions (cf. *Instructions* section in the Methods).

In line with these considerations, reward earned, calculated based on all valid trials (including both exploitation and exploration trials) did not differentiate between included and excluded blocks in the lab study, while optimal choice did (**Figure 2-S2**). In addition, reward percentage calculated based on exploitation trials only, and thus not confounded by the number of exploration trials, on which no reward could be earned, showed a markedly limited range, illustrating the compensatory quality of reward as a behavioral measure. Despite a strong correlation with optimal choice percentage (lab: $r = 0.90$, $p < 2.2e\text{-}16$; online: $r = 0.93$, $p < 2.2e\text{-}16$), the lower limit of reward percentage lied at more than 90% in the lab study and more than 85% in the online study (**Figure 2-S3**).

Computational models

This section describes computational models that were fit to the data from the ExploreExploit task. In total, there we tested 18 models, comprising UCB-type and discounting models (described in detail

below), their nested versions (leaving out certain parameters), and combination models (combining elements of UCB-type and discounting models). **Table 2-S2** (presented below) contains a summary of

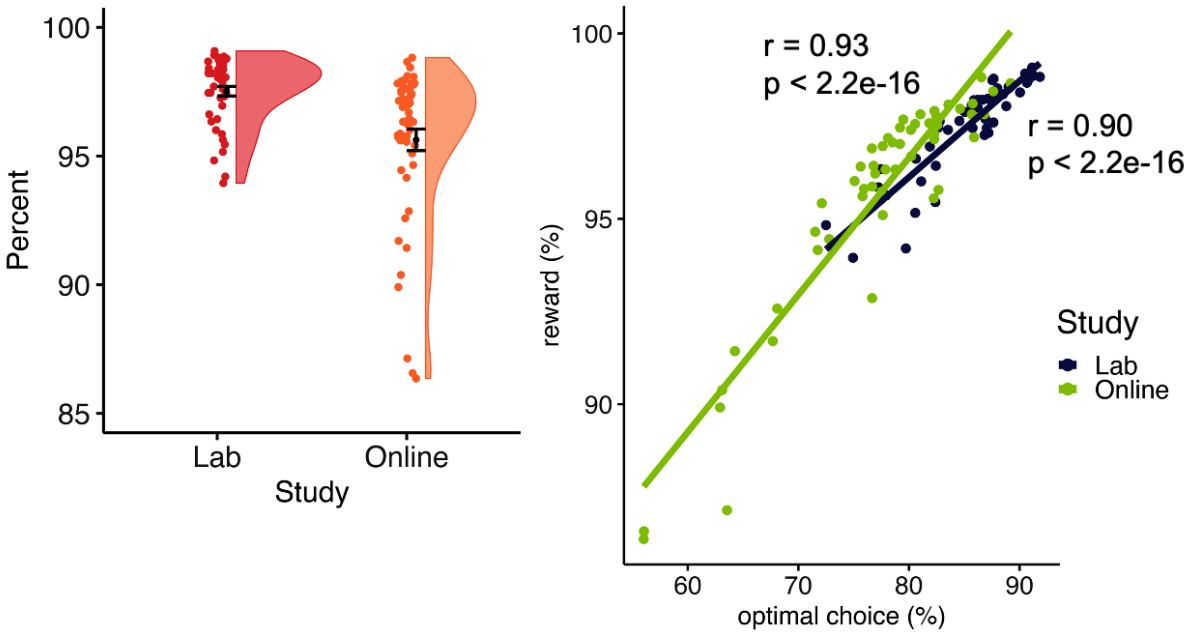


Figure 2-S3. Reward earned. *Left – Reward percentage. Right – Correlation between optimal choice and reward percentage. Reward percentage was calculated based on total reward available only on exploitation trials, to avoid a confounding influence of a higher number of exploration trials, on which no reward could be earned.*

parameters used in each model. The winning model (the model that best fit the data of both the lab and online studies; model 17 in **Table 2-S2**) is described in detail in *Computational modeling* section in the main text.

Expected value for exploitation

To reflect the difference in feedback between on exploration and exploitation trials (only information or reward, respectively), we modeled different expected values for exploring or exploiting each bandit. The expected value of exploitation ($V\_exploit_{i,t}$) for bandit *i* on trial *t* was defined as the expected reward of this bandit ($Q_{i,t}$):

$$V\_exploit_{i,t} = Q_{i,t}$$

Expected value for exploration: UCB-type models

We tested two strategies to computationally define the expected value of exploration ($V\_explore_{i,t}$) for the bandit *i* on trial *t*. In the first type of model, we capitalized on the Upper Confidence Bound (UCB) algorithm (Auer, 2002) that accounts for the influence of uncertainty about an option. Thus, the expected value of exploration ($V\_explore_{i,t}$) consisted of the sum of the expected reward value ($Q_{i,t}$) of bandit *i* on trial *t* and the expected uncertainty ($\sigma_{i,t}$) about this reward. In the full model of this type (model 5 in **Table**

2-S2), expected reward and uncertainty were weighted by parameters β_1 and β_2, respectively, which were estimated as free parameters:

$$V\_explore_{i,t} = \beta_1 * Q_{i,t} + \beta_2 * \sigma_{i,t}$$

Expected value for exploration: discounting models

The second way of modeling exploration consisted of applying a discounting factor to the sum of the expected reward value and expected uncertainty of a bandit *i* on trial *t* after it was explored, as well as a step-wise restoration of this value in the exploitation trials. In the full model, the sum of the expected reward and uncertainty was multiplied by either an exponential (model 15 in **Table 2-S2**) or hyperbolic (model 16 in **Table 2-S2**) discounting factor (Green and Myerson, 1996, 2004) on the trial following exploration, including the number of trials that the bandit has been explored ($n\_explore_t$) and a free parameter κ, that determines how strongly the value of exploring a certain bandit is discounted after it has been explored. The discounting factor was set to 0 if the bandit was not explored on a previous trial.

After the bandit was first exploited, a stepwise restoration factor was applied to bring the value of exploration, reduced by the discounting, back to its level before discounting. To model step-wise restoration factor, the sum of expected reward and uncertainty was multiplied by a free parameter γ, which determined how fast the value is restored to its original magnitude. In addition, γ was taken to the power of the number of trials that the bandit has been exploited for ($n\_exploit_t$). The step-wise restoration factor was then subtracted from the sum of expected reward and uncertainty. If a bandit was not exploited on the previous trial, the stepwise-restoration factor was set to 0.

Exploration value with exponential discounting (multiplicative factor) and stepwise restoration (subtraction factor):

$$V\_explore_{i,t} = \left(Q_{i,t} + \sigma_{i,t}\right) * \exp\left(-\kappa * n_{explore_t}\right) - \left(Q_{i,t} + \sigma_{i,t}\right) * \gamma^{n\_exploit_t}$$

Exploration value with hyperbolic discounting (multiplicative factor) and stepwise restoration (subtraction factor):

$$V\_explore_{i,t} = \left(Q_{i,t} + \sigma_{i,t}\right) * \frac{1}{1 + \kappa * n_{explore_t}} - \left(Q_{i,t} + \sigma_{i,t}\right) * \gamma^{n\_exploit_t}$$

Learning after the bandit was explored

After a bandit was explored, we modeled a learning process to incorporate information (r_t – observed reward on trial *t*) received as feedback into the existing idea about the rewards of the respective bandit. In addition, uncertainty about the expected reward of that bandit decreased. In all but one model, the

expected reward and uncertainty about it were updated with a temporal difference (TD) learning model (Sutton, 1988; Sutton and Barto, 1998) with a learning rate α (free parameter):

$$Q_{i,t+1} = Q_{i,t} + (r_t - Q_{i,t}) * \alpha$$

$$\sigma_{i,t+1} = \sqrt{\sigma_{i,t}^2 - \alpha * \sigma_{i,t}^2}$$

To test for differences in updating of reward and uncertainty, we implemented models with separate learning rates α_1 and α_2 for the expected reward und uncertainty about it, respectively:

$$Q_{i,t+1} = Q_{i,t} + (r_t - Q_{i,t}) * \alpha_1$$

$$\sigma_{i,t+1} = \sqrt{\sigma_{i,t}^2 - \alpha_2 * \sigma_{i,t}^2}$$

We tested an alternative learning function – the Kalman filter (Kalman, 1960; Daw et al., 2006) to dynamically update both the expected mean and the standard deviation of the reward values (model 1 in **Table S2-2**):

$$Q_{i,t+1} = Q_{i,t} + (r_t - Q_{i,t}) * K_t$$

$$\sigma_{i,t+1} = \sqrt{\sigma_{i,t}^2 - K_t * \sigma_{i,t}^2}$$

where K_t is the learning factor (also known as "Kalman gain"), which is updated on each trial t:

$$K_t = \frac{\sigma_{i,t}^2}{\sigma_{i,t}^2 + s_0^2}$$

where s_0^2 is the variance with which individual rewards were sampled from around the mean (cf. *Stimuli*).

Forgetting the after the bandit was exploited or not explored

The reward and uncertainty values of bandits that were exploited or not explored were forgotten with a forgetting rate λ (free parameter):

$$Q_{i,t+1} = \lambda * Q_{i,t} + (1 - \lambda) * Q_0$$

$$\sigma_{i,t+1} = \lambda * \sigma_{i,t} + (1 - \lambda) * \sigma_0$$

We also implemented models with separate rates λ_1 and λ_2 for forgetting of the expected reward and uncertainty, respectively:

$$Q_{i,t+1} = \lambda_1 * Q_{i,t} + (1 - \lambda_1) * Q_0$$

$$\sigma_{i,t+1} = \lambda_2 * \sigma_{i,t} + (1 - \lambda_2) * \sigma_0$$

We defined forgetting as returning to the starting value of Q_0 and σ_0.

Choice function

Six expected values, which corresponded to 6 response possibilities (exploring or exploiting 3 bandits), were passed to the softmax choice rule to determine the probability of each response ($P_{i,a,t}$ – probability of applying action *a* to bandit *i* on trial *t*). A trial outcome was then chosen according to the probabilities returned by the softmax algorithm, using inverse temperature (τ, free parameter) to determined stochasticity of the choice: the lower the inverse temperature, the more stochastically the response will be chosen (the less it is driven by the largest expected value):

$$P_{i,a,t} = \frac{\exp(\tau * V_{i,a,t})}{\sum_{j,x} \exp(\tau * V_{j,x,t})}$$

where *j* denotes all other bandits and *x* denotes all other actions.

Table 2-S2. Models included in the model comparison and their parameters. *Bold – best-fitting model in both lab and online study, defined by a UCB-type exploration value and separate forgetting rates for expected reward and uncertainty. KF – Kalman filter; TD – temporal difference learning; β_1, β_2 – weights for reward and uncertainty in UCB-type expected value for exploration; α – learning rate; α_1, α_2 – separate learning rates for reward and uncertainty, respectively; λ – forgetting rate; λ_1, λ_2 – separate forgetting rates for reward and uncertainty, respectively; exp – exponential discounting; hyp – hyperbolic discounting; k – discounting rate parameter; restore – stepwise restoration factor.*

Model	Expected reward & uncertainty	β_1, β_2 weights	Learning	Forgetting	Discounting	Restoration
mod1	KF	β_1, β_2	-	λ	-	-
mod2	TD+σ	β_1, β_2	α	λ	-	-
mod3	TD+σ	-	α	λ	-	-
mod4	TD+σ	β_1, β_2	α	-	-	-
mod5	TD+σ	β_1, β_2	α_1, α_2	λ_1, λ_2	-	-

mod6	TD+σ	-	α	-	-	-
mod7	TD+σ	-	α	λ	exp, no k	-
mod8	TD+σ	-	α	λ	hyp, no k	-
mod9	TD+σ	-	α	λ	exp, k	-
mod10	TD+σ	-	α	λ	hyp, k	-
mod11	TD+σ	β_1, β_2	α	λ	exp, no k	-
mod12	TD+σ	β_1, β_2	α	λ	hyp, no k	-
mod13	TD+σ	-	α	-	exp, no k	restore
mod14	TD+σ	-	α	-	hyp, no k	restore
mod15	TD+σ	-	α	-	exp, k	restore
mod16	TD+σ	-	α	-	hyp, k	restore
mod17	**TD+σ**	**β_1, β_2**	**α**	**λ_1, λ_2**	**-**	**-**
mod18	TD+σ	β_1, β_2	α_1, α_2	λ	-	-

References

Auer, P. (2002). Using Confidence Bounds for Exploitation-Exploration Trade-offs. *Journal of Machine Learning Research*, *3*, 397–422.

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. https://doi.org/10.1038/nature04766

Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. MIT press.

Green, L., & Myerson, J. (1996). Exponential Versus Hyperbolic Discounting of Delayed Outcomes: Risk and Waiting Time. *Integrative and Comparative Biology*, *36*(4), 496–505. https://doi.org/10.1093/icb/36.4.496

Green, L., & Myerson, J. (2004). A Discounting Framework for Choice With Delayed and Probabilistic Rewards. *Psychological Bulletin*, *130*(5), 769. https://doi.org/10.1037/0033-2909.130.5.769

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, *82*(1), 35–45. https://doi.org/10.1115/1.3662552

Lieder, F., & Griffiths, T. L. (2015). When to use which heuristic: A rational solution to the strategy selection problem. *CogSci*.

Lieder, F., Krueger, P. M., & Griffiths, T. L. (2017). An automatic method for discovering rational heuristics for risky choice. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive Strategy Selection in Decision Making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 534–552. https://doi.org/10.1037/0278-7393.14.3.534

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138. https://doi.org/10.1037/h0042769

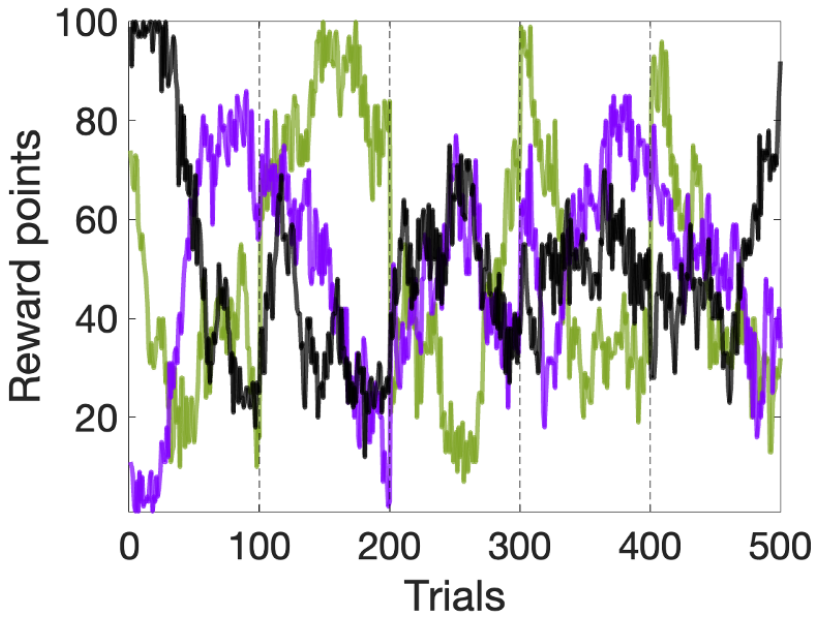
Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and Exploration in a Restless Bandit Problem. *Topics in Cognitive Science*, *7*(2), 351–367. https://doi.org/10.1111/tops.12145

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*(1), 9–44. https://doi.org/10.1007/bf00115009

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning* (Vol. 135). MIT press.

Supplementary Figures

Figure 2-S1. Example of a reward structure. Colors denote bandits, dashed line separates task blocks. Rewards for each bandit and each block were independent.



Figures 2-S2 – 2-S3 are presented in Supplementary Methods.

Figure 2-S4. Missed trials. Left – Percent missed trials. Right – Distribution of missed trials throughout the block (fraction of number total trials in the respective position).

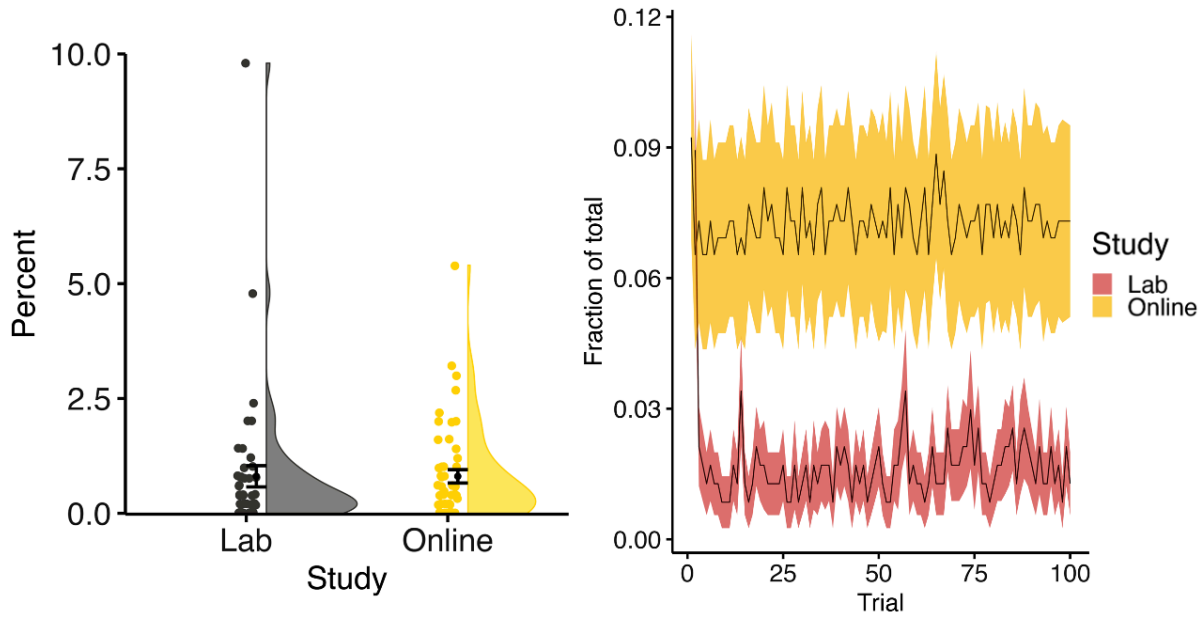
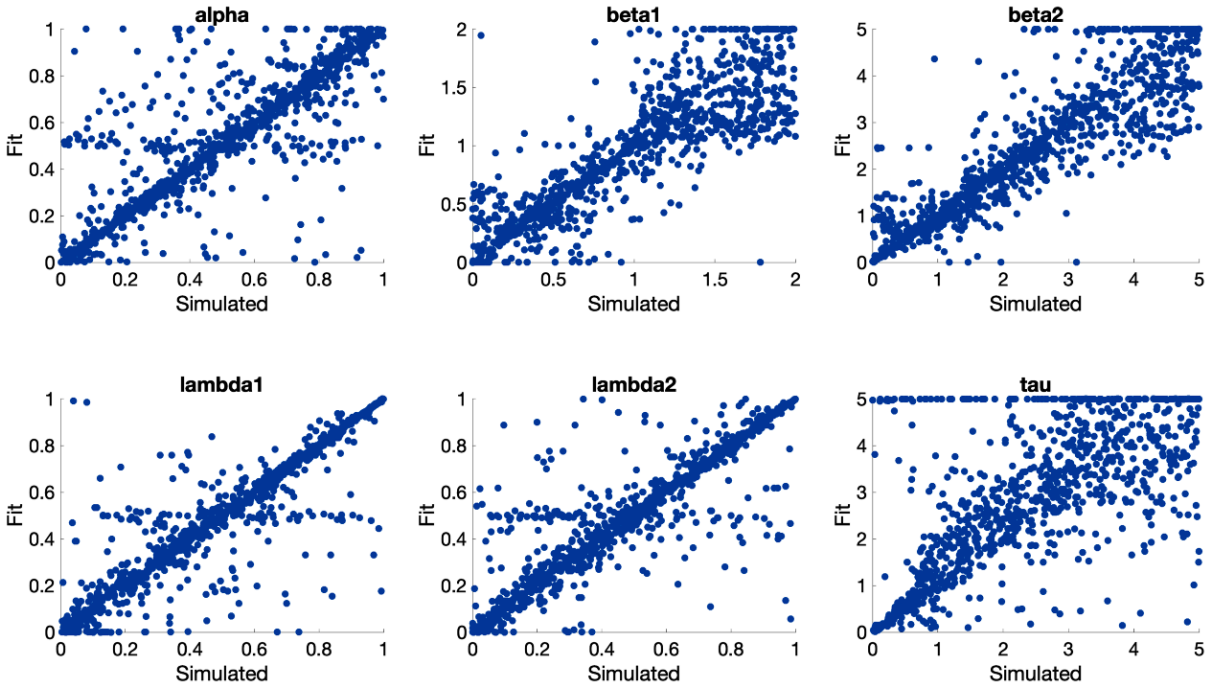


Figure 2-S5. Parameter recovery: simulated vs. fitted parameter values for the winning model. Alpha – learning rate (α), beta1 – weight for reward in exploration (β_1), beta2 – weight for uncertainty in exploration (β_2), lambda1 – forgetting rate for reward (λ_1), lambda2 – forgetting rate for uncertainty (λ_2), tau – inverse temperature (τ).



Supplementary Tables

Table 2-S1. R packages used for data analyses.

Package	Functionality	Citation
lmer4	Linear (mixed) models	(Bates et al., 2014)
afex	Repeated-measures ANOVA	(Singmann et al., 2022)
emmeans	Follow-up comparisons	(Lenth, 2022)
effectsize	η^2	(Ben-Shachar et al., 2020)
partR2	R^2	(Stoffel et al., 2021)
rstatix	Identify outliers	(Kassambara, 2021)
Hmisc	Correlation matrix	(Harrell, 2023)
tidyverse	Data manipulation	(Wickham et al., 2019)
ggplot2	Plotting	(Wickham, 2016)
viridis	Plotting	(Garnier et al., 2023)
ggcorrplot	Plotting	(Kassambara, 2023)

References

Bates D, Mächler M, Bolker B, Walker S (2014) Fitting Linear Mixed-Effects Models using lme4. Arxiv.

Ben-Shachar MS, Lüdecke D, Makowski D (2020) {e}ffectsize: Estimation of Effect Size Indices and Standardized Parameters. Journal of Open Source Software 5:2815 Available at: https://doi.org/10.21105/joss.02815.

Garnier S, Ross N, Rudis R, Camargo AP, Sciaini M, Scherer C (2023) viridis(Lite) - Colorblind-Friendly Color Maps for R. viridis package version 0.6.4. Available at: https://sjmgarnier.github.io/viridis/.

Harrell FE (2023) Hmisc: Harrell Miscellaneous. R package version 5.1-1. Available at: <https://CRAN.R-project.org/package=Hmisc>;.

Kassambara A (2021) Pipe-Friendly Framework for Basic Statistical Tests. Available at: https://CRAN.R-project.org/package=rstatix.

Kassambara A (2023) ggcorrplot: Visualization of a Correlation Matrix using “ggplot2”. R package version 0.1.4.1. Available at: https://CRAN.R-project.org/package=ggcorrplot.

Lenth RV (2022) emmeans: Estimated Marginal Means, aka Least-Squares Means. Available at: https://CRAN.R-project.org/package=emmeans.

Singmann H, Bolker B, Westfall J, Aust F, Ben-Shachar MS (2022) afex: Analysis of Factorial Experiments. Available at: https://CRAN.R-project.org/package=afex.

Stoffel MA, Nakagawa S, Schielzeth H (2021) partR2: partitioning R2 in generalized linear mixed models. Peerj 9:e11414.

Wickham H (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. Available at: https://ggplot2.tidyverse.org.

Wickham H et al. (2019) Welcome to the Tidyverse. J Open Source Softw 4:1686.

Table 2-S2 is presented in Supplementary Methods.

Table S2-3. Value ranges used for estimating free parameters.

Free parameter	Value range
α, α_1, α_2	0 - 1
β_1	0 - 2
β_2	0 - 5
λ, λ_1, λ_2	0 - 1
t	0.001 - 5
g	0 - 1
k	0 - 2

C. Supplementary Materials to Chapter 3

Supplementary Figures

Figure 3-S1. Results of task PLS analysis with IQR BOLD sequences of 3 exploration and exploitation trials with the BSR threshold raised to 5. Axial brain view. MNI coordinates of the first slice: z = -17. Each next slice increases z coordinate in increments of 3. BSR – bootstrap ratio.

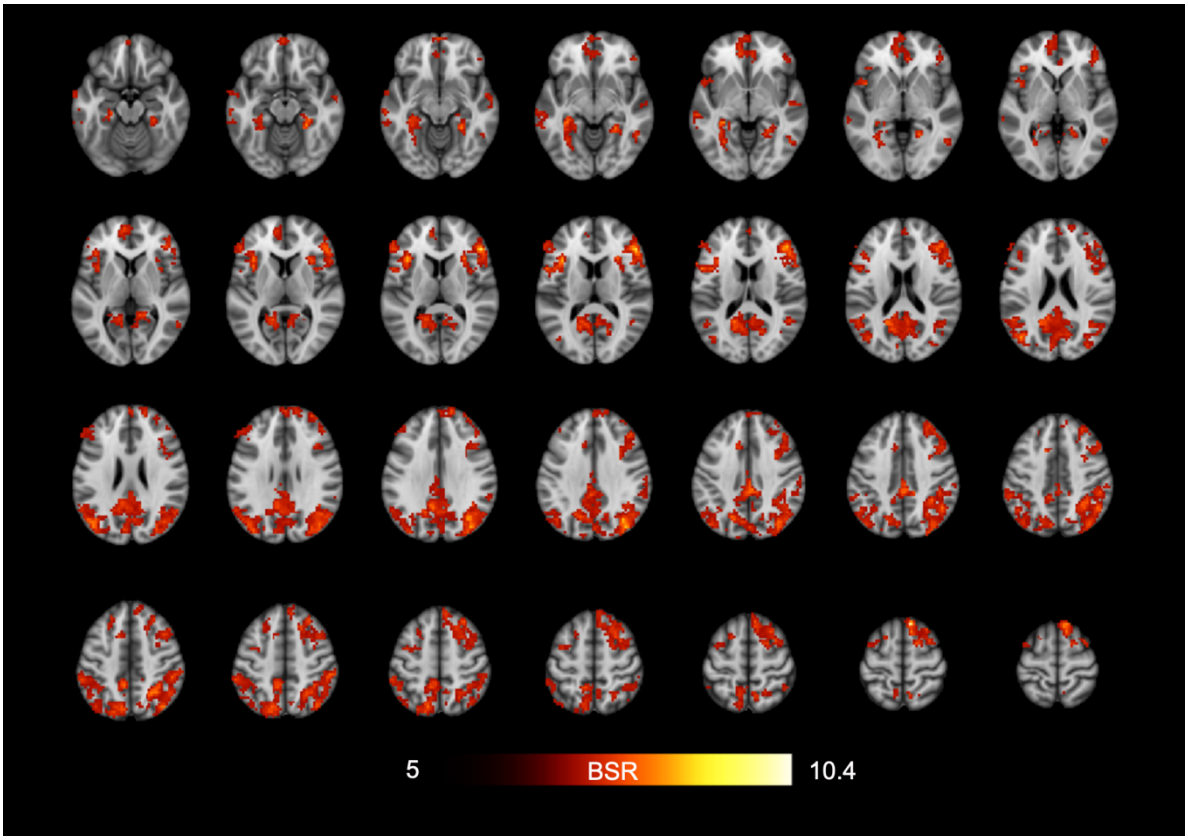


Figure 3-S2. Results of task PLS analysis with IQR BOLD in sequences of 5 exploitation trials. Left – IQR BOLD levels (expressed as brain scores) in a sequence of 5 exploitation trials increase from trial 1 to 3 and remain largely the same through trials 4 and 5. Error bars – SEM. Right – axial brain view. MNI coordinates of the first slice: z = -23. Each next slice increases z coordinate in increments of 3. BSR – bootstrap ratio. (Figure on the next page)

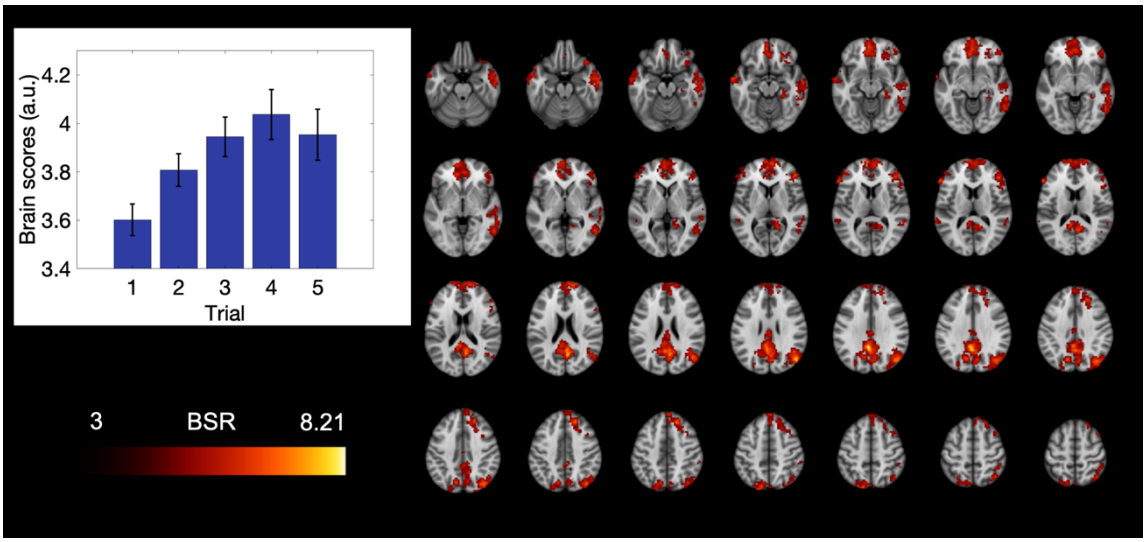


Figure 3-S3. Results of task PLS analysis with IQR BOLD at low, medium and high levels of sigma and entropy in exploration trials. Top – axial brain view. MNI coordinates of the first slice: z = -24. Each next slice increases z coordinate in increments of 3. BSR – bootstrap ratio. Bottom – IQR BOLD levels (expressed as brain scores) at low, medium and high levels of sigma (green) and entropy (blue). Error bars – SEM.

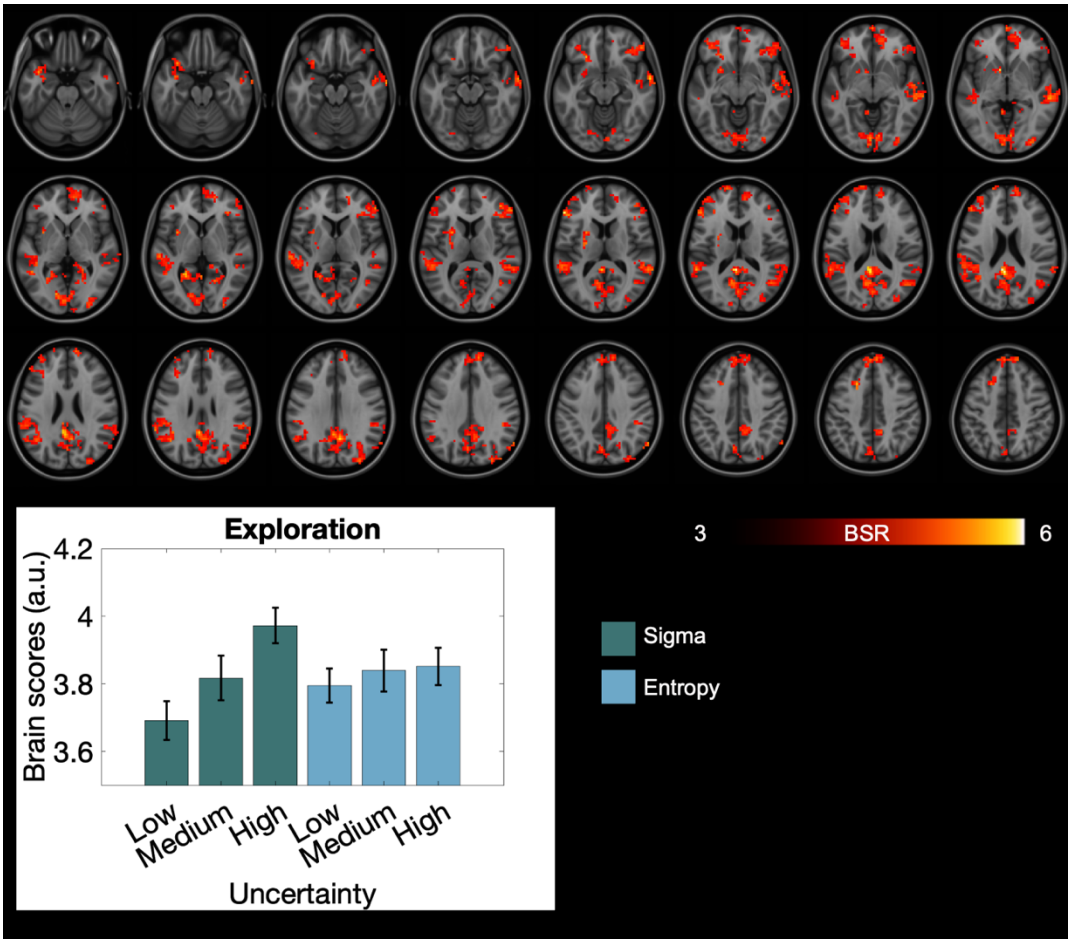


Figure 3-S4. Results of task PLS analysis with IQR BOLD in switch and noswitch exploration and exploitation trials with the BSR threshold raised to 6. Axial brain view. MNI coordinates of the first slice: z = -32. Each next slice increases z coordinate in increments of 3. BSR – bootstrap ratio.

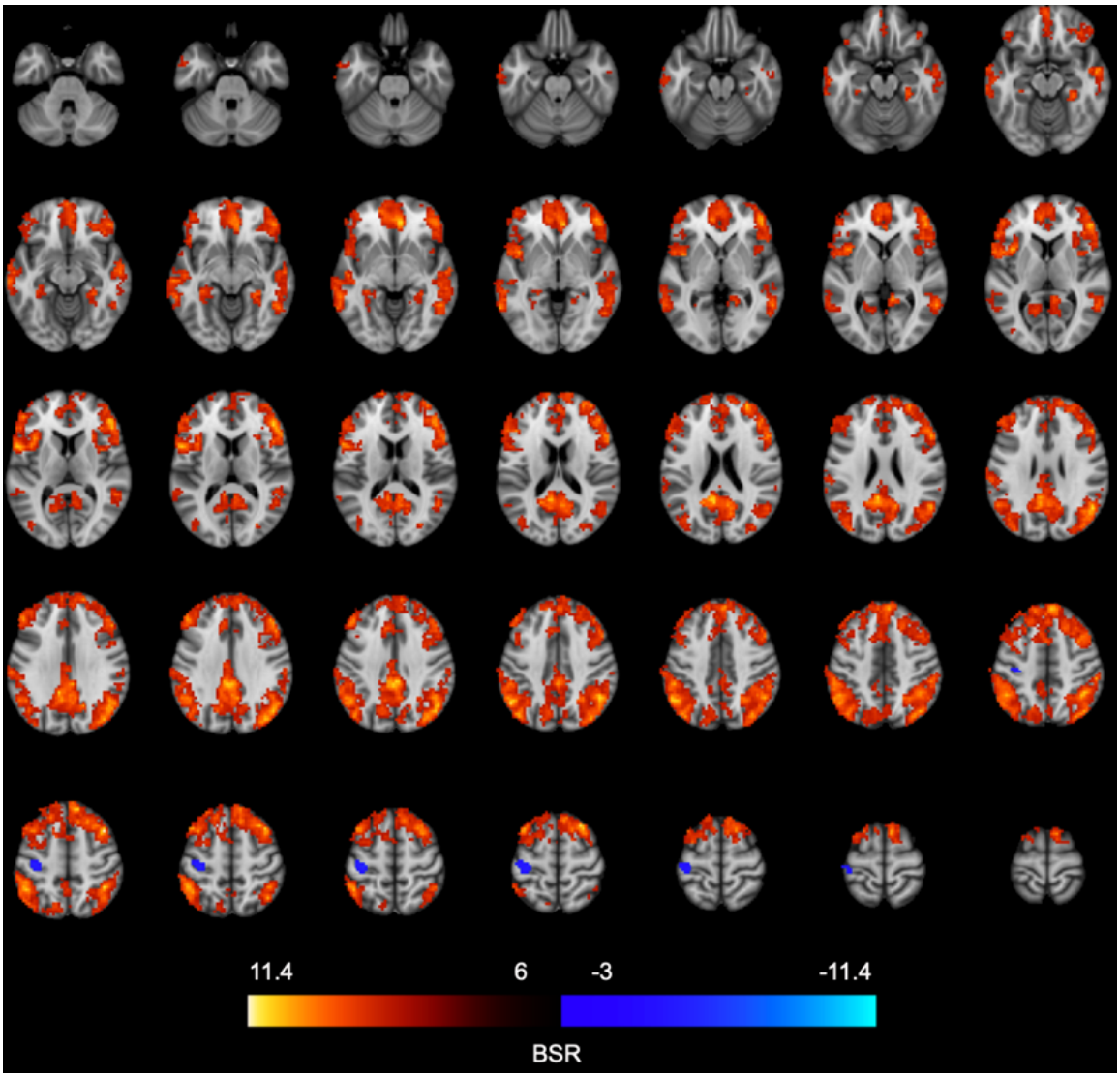


Figure 3-S5. Results of behavior PLS analysis with modulation of IQR BOLD in the first 3 exploration and exploitation trials and optimal choice percentage. Top panel – axial brain view. MNI coordinates of the first slice: z = 14. Each next slice increases z coordinate in increments of 3. BSR – bootstrap ratio. Bottom panel – correlation of IQR BOLD modulation (expressed as brain scores) with optimal choice percentage. Note that these results were not significant at the latent level. Bootstrapped CI for correlations: Exploration – [-0.83, -0.51], Exploitation – [-0.85, -0.50]. (Figure on the next page)

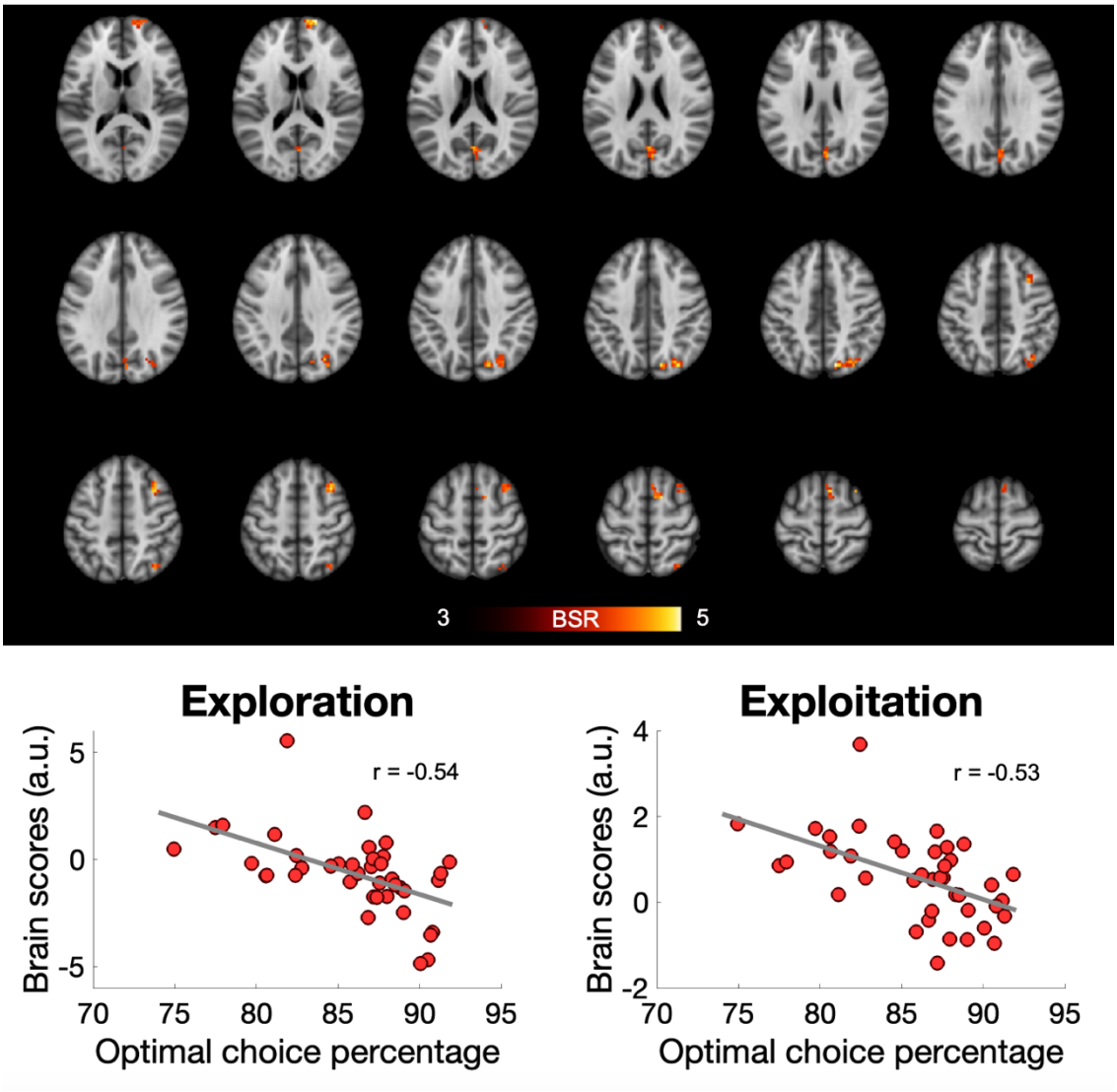


Figure 3-S6. Results of behavior PLS analysis with level of IQR BOLD in exploitation trials 1, 2, 3 and optimal choice percentage. Top panel – axial brain view. MNI z coordinates are indicated. Each next slice increases z coordinate in increments of 3. BSR – bootstrap ratio. Bottom panel – correlation of IQR BOLD level (expressed as brain scores) with optimal choice percentage. Bootstrapped CI for correlations: trial 1 – [0.78, 0.92], trial 2 – [0.77,0.91], trial 3 – [0.76, 0.91]. (Figure on the next page)

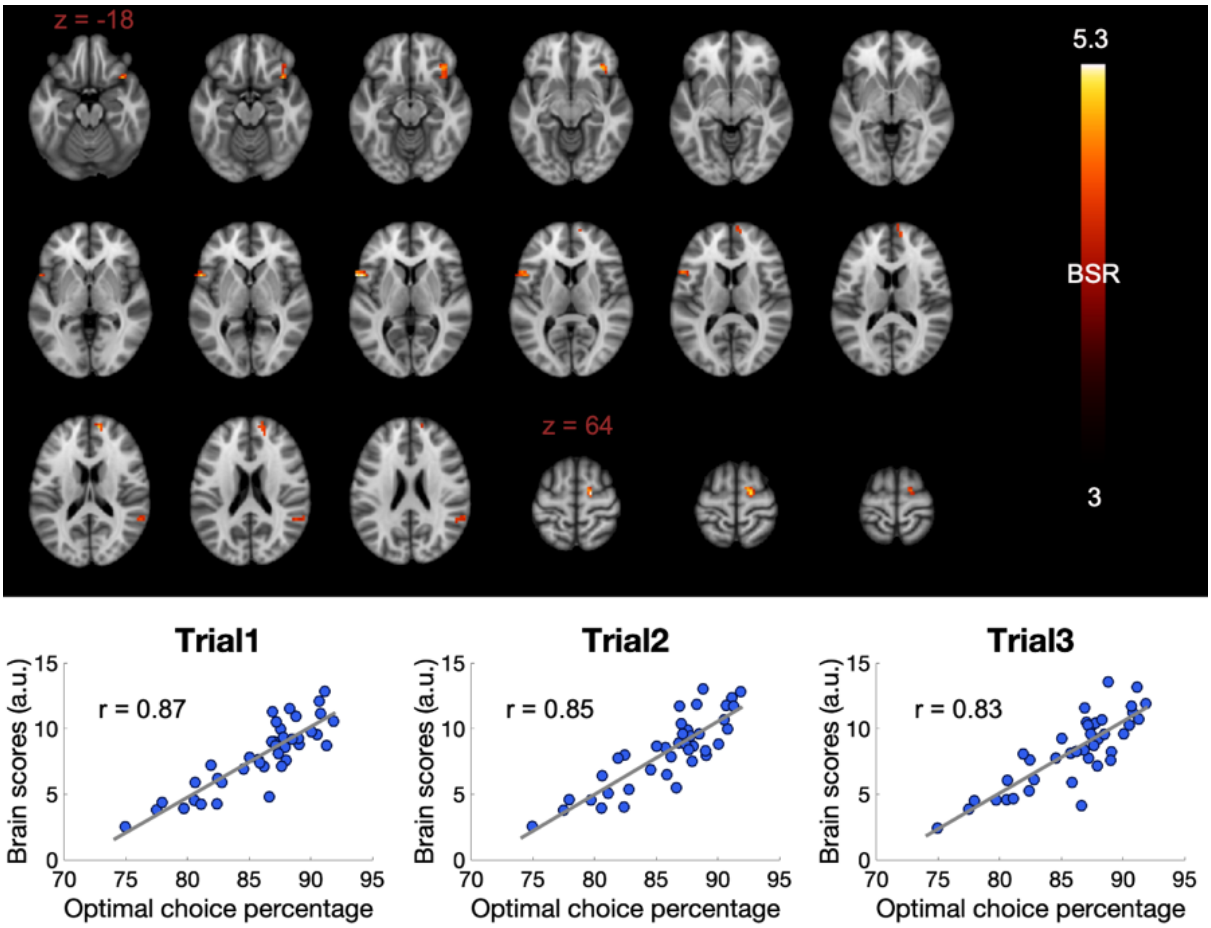
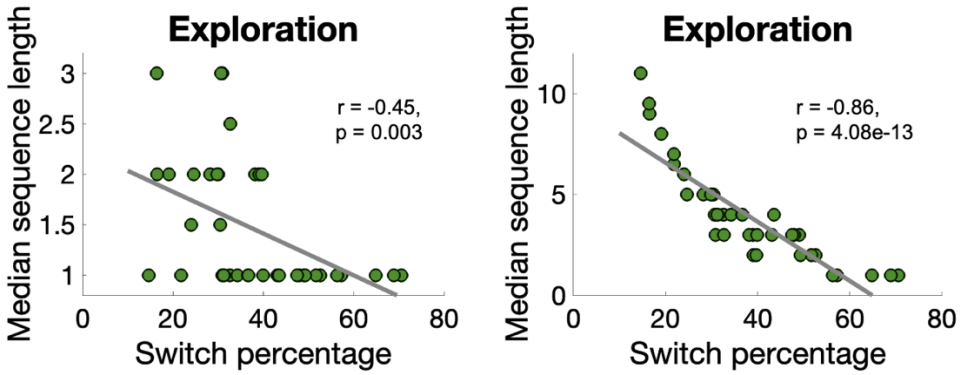


Figure 3-S7. Correlation between switch percentage and median continuous sequence length in exploration and exploitation.



Supplementary Tables

Table 3-S1. **Cluster peak coordinates from task PLS analysis with sequences of 3 continuous exploration and exploitation trials.** The upper part of the table refers to clusters observed with a BSR threshold of +/- 3. The bottom part of the table presents results with a BSR threshold increased to 5 in order to reveal clusters that were otherwise part of one big cluster. BSR – bootstrap ratio.

BSR [-3 3]					
Anatomical Region	MNI Coordinates			BSR	Cluster Size (voxels)
	X	Y	Z		
R Superior Frontal Gyrus	12	27	60	10.29	11893
R Lateral Occipital Cortex	42	-78	33	9.34	12435
R Putamen	30	-18	0	4.80	33
R Thalamus *	9	-9	9	4.42	45
L Thalamus *	-3	-6	3	4.42	45
BSR [-3 5]					
Anatomical Region	MNI Coordinates			BSR	Cluster Size (voxels)
	X	Y	Z		
R Superior Frontal Gyrus	12	27	60	10.29	940
R Inferior Frontal Gyrus	54	33	9	10.03	255
R Lateral Occipital Cortex	42	-78	33	9.34	1090
L Frontal Operculum Cortex	-33	18	12	9.04	293
L Lateral Occipital Cortex	-39	-72	21	8.67	649
R Temporal Fusiform Cortex	30	-39	-15	8.33	1626
L Hippocampus	-24	-36	-6	8.30	172
R Frontal Operculum Cortex	33	21	9	7.85	63
R Frontal Pole	18	60	30	7.28	109
L Paracingulate Gyrus	-9	12	42	7.23	39
L Inferior Frontal Gyrus	-51	33	9	7.16	63
L Frontal Pole	-3	57	3	7.15	231
L Middle Temporal Gyrus	-57	-27	-6	7.07	65
R Precuneus Cortex	12	-57	60	7.01	28
R Middle Temporal Gyrus	60	-57	0	6.80	40
L Inferior Temporal Gyrus	-60	-21	-24	6.49	43
L Superior Frontal Gyrus	-15	12	57	6.44	49
L Precentral Gyrus	-33	-3	51	6.35	50
R Middle Temporal Gyrus	66	-9	-12	6.22	30

Note: * original cluster peak coordinates (x,y,z - 0 0 3) were shifted to obtain a label.

Table 3-S2. Cluster peak coordinates from task PLS analysis with sequences of 5 continuous exploitation trials. BSR – bootstrap ratio.

Anatomical Region	MNI Coordinates			BSR	Cluster Size (voxels)
	X	Y	Z		
Cingulate Gyrus, posterior division	0	-48	30	8.12	1333
R Lateral Occipital Cortex	54	-72	27	7.66	550
L Frontal Pole	-3	66	9	6.58	1464
R Frontal Pole	51	36	12	5.84	234
L Middle Temporal Gyrus	-60	-3	-15	5.67	151
R Middle Temporal Gyrus	66	-3	-24	5.64	704
L Inferior Frontal Gyrus	-54	30	12	5.59	64
R Hippocampus	33	-27	-12	4.75	39
R Frontal Orbital Cortex	42	24	-21	4.61	31
R Frontal Pole	27	36	-15	4.32	59
L Superior Temporal Gyrus	-60	-42	12	4.12	27
L Frontal Pole	-42	42	9	4.05	46

Table 3-S3. Cluster peak coordinates from task PLS analysis with low, medium and high levels of sigma and entropy in exploration. BSR – bootstrap ratio.

Anatomical Region	MNI Coordinates			BSR	Cluster Size (voxels)
	X	Y	Z		
L Cingulate Gyrus, posterior division	-3	-45	15	5.92	930
L Inferior Frontal Gyrus	-54	33	12	5.81	153
L Caudate	-9	9	-3	5.75	28
R Superior Temporal Gyrus	63	-3	-12	5.52	70
L Precuneous Cortex	-21	-51	3	5.35	43
L Superior Frontal Gyrus	-21	18	42	5.14	33
L Supramarginal Gyrus	-57	-45	15	5.08	464
R Middle Temporal Gyrus	60	-30	-3	5.07	127
L Frontal Orbital Cortex	-24	27	-12	5.05	86
L Temporal Pole	-39	6	-24	5.00	76
R Frontal Pole	42	42	9	4.93	106
R Lateral Occipital Cortex	60	-63	36	4.92	273
L Putamen	-24	6	9	4.83	52
R Frontal Pole	12	54	33	4.74	221
R Frontal Pole	51	39	-12	4.64	101
R Lateral Occipital Cortex	33	-87	30	4.59	125

R Lateral Occipital Cortex	33	-90	-3	4.59	60
Frontal Pole	0	66	-6	4.58	136
R Lingual Gyrus	18	-60	3	4.53	48
L Frontal Pole	-15	66	15	4.52	37
R Superior Temporal Gyrus	54	-3	-18	4.22	28

Table 3-S4. Cluster peak coordinates from task PLS analysis with low, medium and high levels of sigma and entropy in exploitation. BSR – bootstrap ratio.

Anatomical Region	MNI Coordinates			BSR	Cluster Size (voxels)
	X	Y	Z		
Cingulate Gyrus, posterior division	0.0	-48.0	33	7.63	2121
R Frontal Pole	6.0	66.0	21	7.10	2109
R Frontal Pole	36.0	39.0	-15	6.51	727
R Middle Temporal Gyrus	63.0	-6.0	-15	6.43	573
R Lateral Occipital Cortex	42.0	-75.0	39	6.15	672
R Frontal Pole	6.0	57.0	-6	5.80	369
L Inferior Frontal Gyrus	-54.0	24.0	18	5.41	898
L Superior Temporal Gyrus	-54.0	-42.0	3	5.33	110
R Hippocampus	24.0	-12.0	-18	5.28	49
L Temporal Pole	-45.0	18.0	-27	5.16	254
L Cingulate Gyrus, anterior division	-3.0	33.0	6	4.62	26
R Supramarginal Gyrus	63	-39	27	4.49	45
L Supramarginal Gyrus	-60	-33	42	4.25	65

Table 3-S5. **Cluster peak coordinates from task PLS analysis with switch and noswitch exploration and exploitation trials.** BSR – bootstrap ratio. The upper part of the table refers to clusters observed with a BSR threshold of +/- 3. The bottom part of the table presents results with a BSR threshold increased to 6 in order to reveal clusters that were otherwise part of one big cluster.

BSR [-3 3]					
Anatomical Region	MNI Coordinates			BSR	Cluster Size (voxels)
	X	Y	Z		
L Precentral Gyrus	-30	-24	57	-4.89	93
R Middle Frontal Gyrus	45	15	51	11.30	28913
L Lateral Occipital Cortex	-48	-60	39	10.99	4438
L Occipital Pole	-18	-99	-3	4.66	31
BSR [-3 6]					

Anatomical Region	MNI Coordinates			BSR	Cluster Size
	X	Y	Z		(voxels)
L Precentral Gyrus	-30	-24	57	-4.89	93
R Middle Frontal Gyrus	45	15	51	11.30	5437
L Lateral Occipital Cortex	-48	-60	39	10.99	1227
R Angular Gyrus	45	-48	51	10.77	1680
L Cingulate Gyrus, posterior division	-3	-48	24	10.55	1553
L Middle Temporal Gyrus	-60	-54	-3	9.55	437
R Middle Temporal Gyrus	60	-9	-15	9.51	66
R Lingual Gyrus	33	-39	-9	8.60	73
L Hippocampus	-27	-36	-9	8.22	70

Table 3-S6. Cluster peak coordinates from behavior PLS analysis with IQR BOLD change in the first 3 exploration and exploitation trials and optimal choice percentage. BSR – bootstrap ratio.

Anatomical Region	MNI Coordinates			BSR	Cluster Size
	X	Y	Z		(voxels)
R Frontal Pole	24	66	18	4.96	27
R Lateral Occipital Cortex	18	-75	45	4.93	81
R Middle Frontal Gyrus	39	15	54	4.82	40
R Superior Frontal Gyrus	18	6	60	4.81	25
R Precuneous Cortex	6	-66	21	4.43	47

Table 3-S7. Cluster peak coordinates from behavior PLS analysis with level of IQR BOLD in the first 3 exploration trials and optimal choice percentage. BSR – bootstrap ratio.

Anatomical Region	MNI Coordinates			BSR	Cluster Size
	X	Y	Z		(voxels)
L Middle Temporal Gyrus	-51	-60	6	6.16	36
L Inferior Frontal Gyrus	-54	15	9	4.99	35
R Frontal Pole	15	66	18	4.20	27
R Angular Gyrus	63	-48	18	4.12	26

Table S8. Cluster peak coordinates from behavior PLS analysis with level of IQR BOLD in the first 3 exploitation trials and optimal choice percentage. BSR – bootstrap ratio.

Anatomical Region	MNI Coordinates			BSR

	X	Y	Z		Cluster (voxels)	Size
R Precentral Gyrus	21	-9	63	5.25	31	
L Inferior Frontal Gyrus	-57	12	6	5.25	30	
R Frontal Pole	15	57	18	4.87	29	
R Frontal Orbital Cortex	42	18	-15	4.85	39	
R Angular Gyrus	63	-45	30	4.49	35	

Table 3-S9. Cluster peak coordinates from behavior PLS analysis with IQR BOLD change in the first 3 exploration and exploitation trials and switch percentage. BSR – bootstrap ratio.

Anatomical Region	MNI Coordinates			BSR	Cluster (voxels)	Size
	X	Y	Z			
L Precentral Gyrus	-51	6	3	7.04	57	
L Lateral Occipital Cortex	-45	-72	33	6.99	41	
R Cuneal Cortex	12	-81	18	6.49	101	
R Frontal Pole	9	54	42	6.47	76	
L Cingulate Gyrus, posterior division	-6	-27	36	6.46	46	
L Supramarginal Gyrus	-54	-36	45	6.32	276	
L Superior Frontal Gyrus	-15	0	63	6.27	98	
L Intracalcarine Cortex	-6	-69	15	6.03	171	
L Precuneous Cortex	-3	-72	48	5.25	67	
L Lateral Occipital Cortex	-45	-69	45	5.23	73	
R Frontal Operculum Cortex	45	15	0	5.21	47	
L Middle Temporal Gyrus	-54	-24	-12	5.07	52	
L Precentral Gyrus	-36	-3	54	4.61	29	
R Middle Frontal Gyrus	45	6	45	4.51	31	
L Insular Cortex	-30	24	9	4.47	26	
L Lateral Occipital Cortex	-15	-63	60	4.07	28	

Table 3-S10. **Median sequence length of exploration and exploitation trials (separately) predicts PLS results.** In separate regression models, exploitation brain scores from the behavior PLS analysis with IQR BOLD modulation and switch percentage were predicted by median continuous sequence length in exploration and exploitation. b – beta, SE – standard error, CI – confidence interval, t(df) – t-value, p – p-value, adj. R^2 – adjusted R^2 as a measure of effect size.

b	SE	CI 95%	t(38)	p	adj. R^2
Exploitation					

3463	916	[1608; 5317]	3.78	0.0005	0.25
Exploration					
9194	3655	[1795; 16593]	2.51	0.016	0.12

Table 3-S11. Cluster peak coordinates from behavior PLS analysis with level of IQR BOLD in the first 3 exploitation trials and switch percentage. BSR – bootstrap ratio.

Anatomical Region	MNI Coordinates			BSR	Cluster Size (voxels)
	X	Y	Z		
R Cingulate Gyrus, posterior division	6	-30	33	8.55	54
L Insular Cortex	-39	-12	0	5.79	32
Precentral Gyrus	0	-21	48	5.47	39
R Angular Gyrus	57	-45	21	4.64	32
L Precentral Gyrus	-39	-12	36	4.34	39

D. Supplementary Materials to Chapter 4

Supplementary Figures

Figure 4-S1. Number of trials without valid eye tracking data for each participant. Number of trials is expressed as a fraction of total available (max. 500) trials.

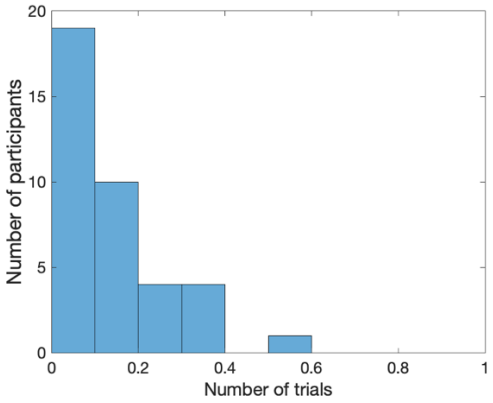


Figure 4-S2. Number of trials with different number of dwell locations for each participant. Number of trials is expressed as a fraction of total available (max. 500) trials. Category 3+ contains all trials with 3 and more dwell locations.

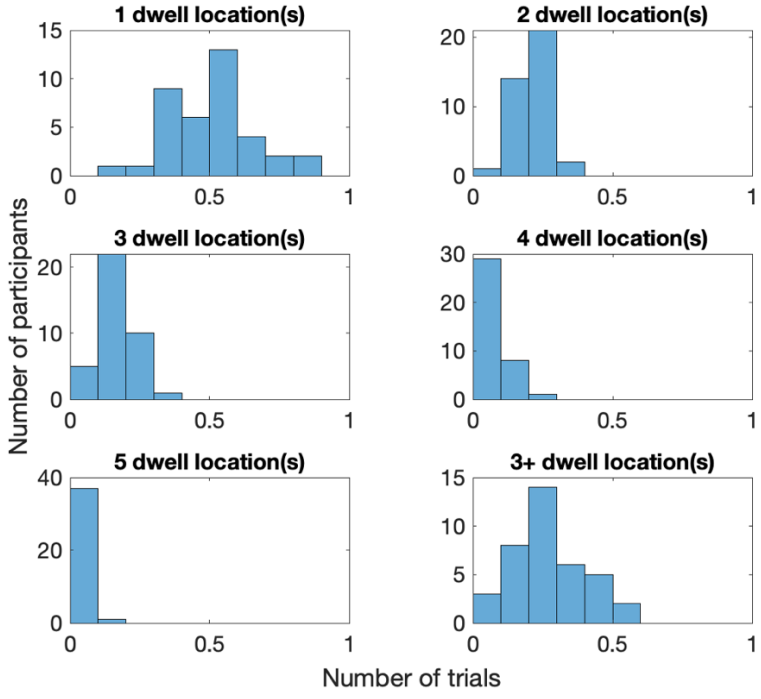


Figure 4-S3. Start rank in patterns with 2 dwell locations. For dwell patterns with 2 dwell locations, there was no significant effect of start rank (bandit) in the model based on either EV ranks (left) or sigma ranks (right). Predicted probability: 0 on the y-axis corresponds to 100% probability of exploitation, 0% probability of exploration, 1 on the y-axis corresponds to 100% probability of exploration, 0% probability of exploitation.

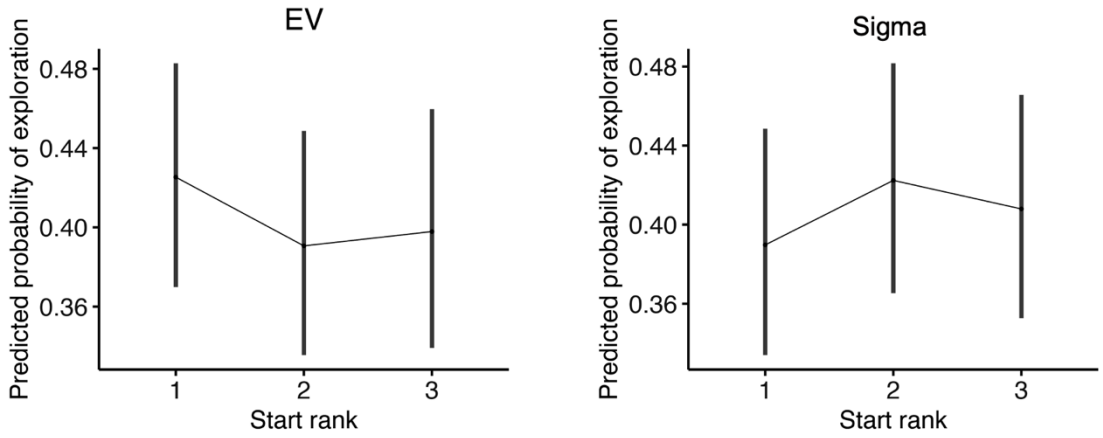
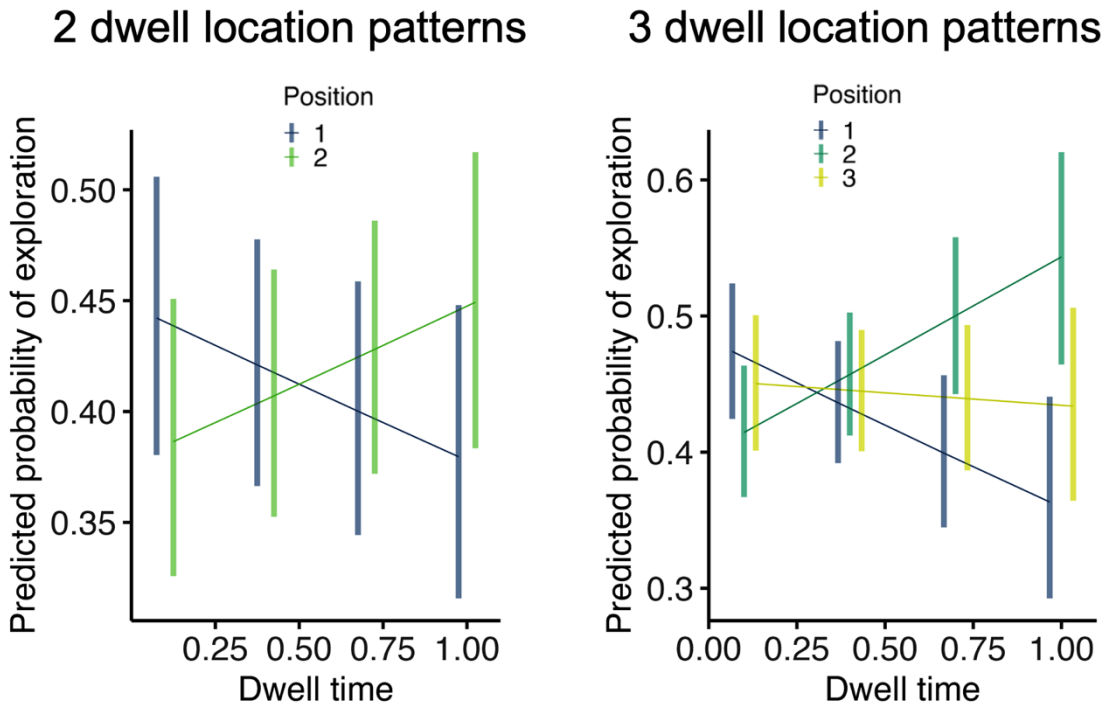


Figure 4-S4. Dwell time in each position in patterns with 2 and 3 dwell locations. Predicted probability: 0 on the y-axis corresponds to 100% probability of exploitation, 0% probability of exploration, 1 on the y-axis corresponds to 100% probability of exploration, 0% probability of exploitation.



Supplementary Tables

Table 4-S1. Fixations and dwell time reflect choice. Left - results of a logistic regression model predicting whether the bandit was fixated based on whether it was chosen and the trial type. Right - results of a linear regression model predicting the dwell time on the bandit based on whether it was chosen and the trial type. CI – confidence interval, p – p-value, R^2 – marginal R^2, N – number of subjects.

	Fixated			**Dwell time**			
Predictors	*Odds Ratios*	*CI*	*p*	*Estimates*	*CI*	*p*	R^2
(Intercept)	0.45	0.40 – 0.50	**<0.001**	0.52	0.50 – 0.54	**<0.001**	
trial type [explore]	1.52	1.45 – 1.60	**<0.001**	-0.02	-0.04 – -0.01	**<0.001**	0.0006
chosen [1]	14.53	13.63 – 15.50	**<0.001**	0.28	0.27 – 0.29	**<0.001**	0.1187
trial type [explore] × chosen [1]	0.54	0.49 – 0.60	**<0.001**	-0.04	-0.06 – -0.03	**<0.001**	0.0014
N	38 id			38 id			
Observations	47940			24979			
Marginal R^2 / Conditional R^2	0.288 / 0.312			0.180 / 0.230			

Table 4-S2. Gaze reflects computational model: EV- and sigma-based models. Results of logistic regression models predicting the trial type from the EV rank (left) / sigma rank (right) of the bandit and whether it was fixated. CI – confidence interval, p – p-value, N – number of subjects.

	Trial type [EV model]			**Trial type [sigma model]**		
Predictors	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*
(Intercept)	0.81	0.69 – 0.96	**0.015**	0.41	0.35 – 0.48	**<0.001**
rank [2]	0.48	0.44 – 0.52	**<0.001**	1.08	1.00 – 1.15	**0.036**
rank [3]	0.52	0.48 – 0.56	**<0.001**	1.42	1.32 – 1.52	**<0.001**
fixated [1]	0.55	0.51 – 0.60	**<0.001**	1.64	1.54 – 1.76	**<0.001**
rank [2] × fixated [1]	3.25	2.93 – 3.61	**<0.001**	0.87	0.79 – 0.95	**0.003**
rank [3] × fixated [1]	3.29	2.96 – 3.66	**<0.001**	0.51	0.47 – 0.57	**<0.001**
N	38 id			38 id		
Observations	47940			47940		

Marginal R^2 / Conditional R^2	0.021 / 0.084	0.009 / 0.073

Table 4-S3. **Gaze reflects computational model: combined EV and sigma model.** Results of logistic regression model predicting the trial type based on whether the bandit with EV rank 1 and the bandit with sigma rank 1 were fixated. CI – confidence interval, p – p-value, N – number of subjects.

	Trial type		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.72	0.60 – 0.86	**<0.001**
ev rank1 [1]	0.47	0.42 – 0.52	**<0.001**
sigma rank1 [1]	1.28	1.12 – 1.46	**<0.001**
ev rank1 [1] × sigma rank1 [1]	1.39	1.19 – 1.63	**<0.001**
N_{id}	38		
Observations	15980		
Marginal R^2 / Conditional R^2	0.037 / 0.098		

Table 4-S4. Number of dwell locations predicts trial type. Results of logistic regression model predicting the trial type based on the number of dwell locations visited on the respective trial. CI – confidence interval, p – p-value, N – number of subjects.

	Trial type		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.52	0.44 – 0.60	**<0.001**
n locations	1.34	1.29 – 1.38	**<0.001**
N_{id}	38		
Observations	15980		
Marginal R^2 / Conditional R^2	0.023 / 0.086		

Table 4-S5. Dwell patterns with 1 dwell location predict trial type: EV- and sigma-based models. Results of logistic regression models predicting the trial type based on the prevalence of dwell patterns with 1 dwell location based on EV ranks (left) and sigma ranks (right). Start bandit is the only bandit in the pattern. CI – confidence interval, p – p-value, N – number of subjects.

	Trial type [EV model]			Trial type [sigma model]		
Predictors	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*
(Intercept)	0.28	0.23 – 0.33	**<0.001**	0.49	0.40 – 0.59	**<0.001**
start [2]	2.38	2.10 – 2.70	**<0.001**	0.93	0.82 – 1.06	0.300
start [3]	2.51	2.15 – 2.92	**<0.001**	0.60	0.53 – 0.67	**<0.001**
N	38 $_{id}$			38 $_{id}$		
Observations	7924			7924		
Marginal R^2 / Conditional R^2	0.046 / 0.118			0.016 / 0.092		

Table 4-S6. **Dwell patterns with 1 dwell location predict trial type: combined EV and sigma model.** Dwell Results of logistic regression model predicting the trial type from EV and sigma ranks in the dwell patterns with 1 dwell location. CI – confidence interval, p – p-value, N – number of subjects.

	Trial type		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.35	0.28 – 0.43	**<0.001**
ev start [2]	1.97	1.58 – 2.45	**<0.001**
ev start [3]	1.95	1.52 – 2.48	**<0.001**
sig start [2]	0.90	0.74 – 1.10	0.308
sig start [3]	0.67	0.57 – 0.80	**<0.001**
ev start [2] × sig start [2]	1.09	0.81 – 1.49	0.565
ev start [3] × sig start [2]	1.13	0.78 – 1.63	0.515
ev start [2] × sig start [3]	1.26	0.91 – 1.73	0.162
ev start [3] × sig start [3]	1.58	1.07 – 2.32	**0.021**
N $_{id}$	38		
Observations	7924		
Marginal R^2 / Conditional R^2	0.053 / 0.124		

Table 4-S7. Start bandit in patterns with 2 dwell locations does not predict trial type: EV- and sigma-based models. Results of logistic regression models predicting the trial type based on EV rank

(left) and sigma rank (right) of the start location in patterns with 2 dwell locations. CI – confidence interval, p – p-value, N – number of subjects.

	Trial type [EV model]			Trial type [sigma model]		
Predictors	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*
(Intercept)	0.74	0.59 – 0.93	**0.011**	0.64	0.50 – 0.81	**<0.001**
start [2]	0.87	0.73 – 1.02	0.088	1.15	0.96 – 1.37	0.142
start [3]	0.89	0.74 – 1.08	0.234	1.08	0.91 – 1.29	0.395
N	38 $_{id}$			38 $_{id}$		
Observations	3394			3394		
Marginal R^2 / Conditional R^2	0.001 / 0.111			0.001 / 0.110		

Table 4-S8. End bandit in patterns with 2 dwell locations predicts trial type: EV- and sigma-based models. Results of logistic regression models predicting the trial type based on EV rank (left) and sigma rank (right) of the end location in patterns with 2 dwell locations. CI – confidence interval, p – p-value, N – number of subjects.

	Trial type [EV model]			Trial type [sigma model]		
Predictors	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*
(Intercept)	0.49	0.39 – 0.62	**<0.001**	0.85	0.67 – 1.09	0.202
end [2]	1.68	1.42 – 1.99	**<0.001**	0.79	0.66 – 0.94	**0.009**
end [3]	1.80	1.49 – 2.16	**<0.001**	0.66	0.55 – 0.79	**<0.001**
N	38 $_{id}$			38 $_{id}$		
Observations	3394			3394		
Marginal R^2 / Conditional R^2	0.019 / 0.129			0.008 / 0.117		

Table 4-S9. End bandit in patterns with 2 dwell locations predicts trial type: combined EV and sigma model. Results of logistic regression model predicting the trial type based on EV and sigma ranks of the end location in patterns with 2 dwell locations. CI – confidence interval, p – p-value, N – number of subjects.

	Trial type		
Predictors	*Odds Ratios*	*CI*	*p*

(Intercept)	0.58	0.42 – 0.81	**0.001**
ev end [2]	1.55	1.12 – 2.15	**0.009**
ev end [3]	1.83	1.31 – 2.56	**<0.001**
sig end [2]	0.83	0.59 – 1.16	0.276
sig end [3]	0.79	0.59 – 1.08	0.139
ev end [2] × sig end [2]	1.05	0.68 – 1.63	0.811
ev end [3] × sig end [2]	0.89	0.56 – 1.43	0.632
ev end [2] × sig end [3]	1.08	0.69 – 1.70	0.726
ev end [3] × sig end [3]	0.85	0.52 – 1.38	0.504
N_{id}	38		
Observations	3394		
Marginal R^2 / Conditional R^2	0.022 / 0.133		

Table 4-S10. Patterns with 3 dwell locations predict trial type: EV- and sigma-based models. Results of logistic regression models predicting the trial type based on whether it was an xyx-type pattern and EV (left) / sigma (right) rank of the start location in patterns with 3 dwell locations. CI – confidence interval, p – p-value, N – number of subjects.

	Trial type [EV model]			**Trial type [sigma model]**		
Predictors	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*
(Intercept)	1.01	0.77 – 1.32	0.943	0.81	0.61 – 1.08	0.150
xyx [1]	0.48	0.38 – 0.61	**<0.001**	0.94	0.72 – 1.24	0.672
start [2]	1.01	0.73 – 1.39	0.955	1.07	0.77 – 1.49	0.686
start [3]	0.77	0.55 – 1.06	0.110	1.36	0.98 – 1.88	0.064
xyx [1] × start [2]	1.96	1.37 – 2.80	**<0.001**	0.85	0.58 – 1.24	0.396
xyx [1] × start [3]	3.04	2.08 – 4.44	**<0.001**	0.60	0.42 – 0.87	**0.007**
N	38_{id}			38_{id}		
Observations	4353			4354		
Marginal R^2 / Conditional R^2	0.036 / 0.099			0.007 / 0.069		

Table 4-S11. Patterns with 3 dwell locations predict trial type: combined EV and sigma model. Results of logistic regression model predicting the trial type based on most prevalent xyx-type patterns based on EV and sigma ranks in patterns with 3 dwell locations. CI – confidence interval, p – p-value, N – number of subjects.

	Trial type		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.96	0.80 – 1.15	0.653
ev1y1 [1]	0.50	0.42 – 0.60	**<0.001**
sig3y3 [1]	1.21	0.96 – 1.51	0.102
ev1y1 [1] × sig3y3 [1]	0.84	0.62 – 1.13	0.246
N_{id}	38		
Observations	4353		
Marginal R^2 / Conditional R^2	0.035 / 0.097		

Table 4-S12. Dwell time spent in each position in patterns with 2 dwell locations. Results of logistic regression model predicting the trial type based on the dwell time spent in each position in patterns with 2 dwell locations. CI – confidence interval, p – p-value, N – number of subjects.

	Trial type		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.56	0.42 – 0.74	**<0.001**
position [2]	0.75	0.59 – 0.96	**0.022**
dwelltime	0.75	0.55 – 1.03	0.072
end [2]	1.70	1.51 – 1.92	**<0.001**
end [3]	1.83	1.61 – 2.09	**<0.001**
position [2] × dwelltime	1.78	1.14 – 2.78	**0.012**
N_{id}	38		
Observations	6788		
Marginal R^2 / Conditional R^2	0.020 / 0.140		

Table 4-S13. Dwell time spent in each position in patterns with 3 dwell locations. Results of logistic regression model predicting the trial type based on the dwell time spent in each position in patterns with 3 dwell locations. CI – confidence interval, p – p-value, N – number of subjects.

	Trial type		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	1.10	0.88 – 1.38	0.414
position [2]	0.70	0.58 – 0.86	**0.001**
position [3]	0.87	0.71 – 1.06	0.174
xyx [1]	0.74	0.68 – 0.81	**<0.001**
dwelltime	0.60	0.41 – 0.89	**0.012**
position [2] × dwelltime	2.96	1.70 – 5.13	**<0.001**
position [3] × dwelltime	1.54	0.91 – 2.62	0.110
N_{id}	38		
Observations	13059		
Marginal R^2 / Conditional R^2	0.006 / 0.081		

Table 4-S14. Correlation results between the frequency of dwell patterns with 1 dwell location in each condition and behavioral measures. r – Pearson's correlation coefficient, rho – Spearman's correlation coefficient, p – p-value. Bold – correlations with p-value < 0.05, uncorrected.

Exploration								
	% optimal choice				% switch			
	Pearson		Spearman		Pearson		Spearman	
pattern	r	p	rho	p	r	p	rho	p
1	0.09	0.59	0.11	0.50	-0.11	0.49	-0.10	0.57
2	-0.01	0.96	-0.09	0.61	0.14	0.41	0.09	0.58
3	-0.15	0.38	-0.06	0.71	0.05	0.76	0.06	0.72
Exploitation								
	% optimal choice				% switch			
	Pearson		Spearman		Pearson		Spearman	
pattern	r	p	r	p	r	p	r	p
1	0.07	0.70	0.24	0.15	-0.22	0.18	-0.16	0.35
2	-0.02	0.91	-0.11	0.50	0.19	0.27	0.23	0.16
3	-0.25	0.14	-0.24	0.14	0.10	0.54	0.01	0.96

Table 4-S15. Correlation results between the frequency of dwell patterns with 2 dwell locations in each condition and behavioral measures. r – Pearson's correlation coefficient, rho – Spearman's correlation coefficient, p – p-value. Bold – correlations with p-value < 0.05, uncorrected.

Exploration								
	% optimal choice				% switch			
	Pearson		Spearman		Pearson		Spearman	
pattern	r	p	rho	p	r	p	rho	p
12	0.26	0.11	0.11	0.52	0.29	0.08	0.24	0.14
13	0.05	0.78	0.05	0.75	0.21	0.22	0.13	0.45
21	-0.05	0.78	0.03	0.86	-0.20	0.24	-0.05	0.76
23	-0.11	0.54	-0.03	0.88	0.17	0.33	0.24	0.18
31	-0.10	0.55	-0.22	0.18	-0.35	0.03	-0.38	0.02
32	-0.12	0.48	0.09	0.61	-0.37	0.02	-0.31	0.06
Exploitation								
	% optimal choice				% switch			
	Pearson		Spearman		Pearson		Spearman	
pattern	r	p	r	p	r	p	r	p
12	0.02	0.90	0.05	0.77	-0.10	0.54	-0.08	0.63
13	-0.10	0.58	-0.03	0.85	0.09	0.61	0.12	0.46
21	0.01	0.96	-0.04	0.79	0.03	0.86	0.04	0.81
23	-0.19	0.29	-0.18	0.33	-0.10	0.58	-0.004	0.98
31	0.09	0.58	0.18	0.28	0.02	0.92	-0.08	0.63
32	0.09	0.61	0.03	0.88	-0.31	0.08	-0.24	0.17

Table 4-S16. Correlation results between the frequency of dwell patterns with 3 dwell locations (xyx patterns only) in each condition and behavioral measures. r – Pearson's correlation coefficient, rho – Spearman's correlation coefficient, p – p-value. Bold – correlations with p-value < 0.05, uncorrected.

Exploration								
	% optimal choice				% switch			
	Pearson		Spearman		Pearson		Spearman	
pattern	r	p	rho	p	r	p	rho	p
121	0.02	0.90	-0.01	0.95	-0.12	0.46	-0.07	0.67
131	0.10	0.57	0.07	0.68	-0.17	0.30	-0.03	0.84
212	-0.03	0.88	-0.03	0.87	-0.01	0.95	-0.01	0.97
232	-0.32	0.07	-0.23	0.19	-0.06	0.72	-0.13	0.45
313	0.23	0.18	0.19	0.27	-0.05	0.76	-0.10	0.59
323	-0.31	0.07	-0.40	0.02	-0.38	0.03	-0.39	0.02
Exploitation								
	% optimal choice				% switch			

	Pearson		Spearman		Pearson		Spearman	
pattern	r	p	r	p	r	p	r	p
121	-0.001	0.99	0.12	0.46	-0.20	0.22	-0.14	0.41
131	0.01	0.93	0.07	0.67	-0.19	0.25	-0.16	0.34
212	0.22	0.20	0.16	0.36	0.07	0.69	0.04	0.81
232	0.13	0.49	0.02	0.90	-0.05	0.78	0.09	0.61
313	-0.18	0.29	-0.20	0.26	0.10	0.57	-0.04	0.80
323	-0.20	0.30	-0.29	0.12	-0.25	0.19	-0.23	0.22

E. Declaration of contributions

I. General Information
Last name, first name: Polanski, Liliana
Institute: Fachbereich Erziehungswissenschaft und Psychologie, Max Planck Institute for Human Development
Doctoral study subjects: Psychology, Computational Cognitive Neuroscience
Title: Master of Science (M. Sc.)

II. The work included in this dissertation has not yet been published.

III. Explanation of own share of the work: Liliana Polanski led the work at all stages of this dissertation (project administration, project conceptualization, task design, data collection, data analysis, computational modeling, coding, manuscript writing, review & editing).

Other contributions:

Douglas Garrett supervised the work and provided advice at all stages of this dissertation (project administration, project conceptualization, task design, data collection, data analysis, computational modeling, manuscript review & editing).

Tobias Hauser, Magda Dubois and Alexander Skowron made a substantial contribution to computational modeling (conceptualization of models, sharing code, providing feedback and advice to improve models, discussing results).

Alexander Skowron made a substantial contribution to data collection in the lab study (testing participants with fMRI and eye tracking).

Tobias Hauser made a contribution to task design (providing feedback on early design ideas).

IV. Name, address, and e-mail address of the relevant contributors (in alphabetical order):

Dubois, Magda (*magda.dubois.18@alumni.ucl.ac.uk*), Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Russell Square House, 10-12 Russell Square, London, WC1B 5EH, United Kingdom (last academic affiliation)

Garrett, Douglas (*garrett@mpib-berlin.mpg.de*), Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

Hauser, Tobias (*tobias.hauser@uni-tuebingen.de*), Eberhard Karls Universität Tübingen, Calwerstraße 14, 72076 Tübingen, Germany

Skowron, Alexander (*skowron@mpib-berlin.mpg.de*), Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

Liliana Polanski
23.05.2024

V. Written confirmation of III. by contributors:

I confirm the declaration made by Liliana Polanski under III.:

Name: Magda Dubois Signature: . . . . . .

Name: Douglas Garrett Signature: . . . . . .

Name: Tobias Hauser Signature: . . . . . .

Name: Alexander Skowron Signature: . . . . . .

Written confirmation was obtained from all contributors. For reasons of data protection, the signatures are separately submitted to the Doctoral Examination Office (promotion@ewi-psy.fu-berlin.de), and are not directly included here.

F. Declaration of independent work

I hereby declare that I completed this doctoral thesis independently. Except where otherwise stated, I confirm that the work presented in this thesis is my own. Where information has been derived from other sources (including internet- and AI-based sources), I confirm that this has been indicated in the thesis. I confirm that this dissertation has not been accepted or rejected in a procedure for obtaining a doctoral degree elsewhere. I have acknowledged the Doctoral Degree Regulations which underlie the procedure of the Department of Education and Psychology of Freie Universität Berlin, as amended on 22.05.2023. The principles of Freie Universität Berlin for ensuring good academic practice have been complied with.

Liliana Polanski

Berlin, 09.07.2024