Core Reinforcement Learning Computations Underlying Distinct Behavioral Strategies and their Implications in Psychiatry

Thesis by

Weilun Ding

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2025
Defended November 25, 2024

ACKNOWLEDGEMENTS

My graduate studies and thesis completion would not have been possible without my advisor, professors, senior colleagues, lab mates, friends, and family. I could not write down all the names of those from whom I received help, but I thank them and appreciate that I have had them along the journey.

I would like to thank Prof. John O'Doherty, my advisor, who gave me academic advice and support in completing all the work in this thesis. Completing the PhD degree is never easy, and I feel blessed for Prof. O'Doherty's kindness and patience that helped my academic and personal growth over the years.

I would like to thank Prof. Colin Camerer, Prof. Dean Mobbs, and Prof. Kirby Nielsen, my thesis committee, who gave me support from the candidacy to the defense. I also would like to thank Prof. Antonio Rangel and Prof. Ralph Adolphs, the SDN faculty, for their academic input and support.

I would like to thank Prof. Ming Hsu at UC Berkeley and Prof. Michael Dorris at the Chinese Academy of Sciences, who gave me the opportunity to learn and gain when I was fresh to the field.

I would like to thank Dr. Jeffrey Cockburn, from whom I learned a lot and with whom I collaborated throughout my entire graduate study. I would like to thank him for his significant contribution and guidance on this thesis.

I would like to thank Dr. Mike Tyszka at Caltech Brain Imaging Center, who provided steady support and guidance during the fMRI data collection process over the years.

I would like to thank Sarah Oh, Scarlet Cho, Julia Simon, Emmily Hovhannisyan, and Amogh Johri for their help in participant recruitment and data collection for the large datasets presented here, which was a significant step in completing this thesis.

I would like to thank Sanghyun Yi, my study peer, who gave me various kinds of help as a friend in my graduate life. I would also like to thank Dr. Caroline Charpentier, Dr. Vincent Man, Logan Cross, Tomas Aquino, Sneha Aenugu, and Dr. Rani Gera in my lab for their help and kindness in my graduate studies.

Lastly, I would like to thank my family for their support and care throughout my life abroad.

I cannot thank everyone here, but I thank them in my heart for helping me get to where I am.

ABSTRACT

Reinforcement learning (RL) models have shown great capabilities in characterizing learning and decision-making in the real world. The dual systems of the model-free (MF) and model-based (MB) algorithms have been proposed to describe the computational mechanism underlying a reflexive habitual control and a cognitive goal-directed control, respectively. Given the dual systems under control, it is worth asking how the choice of which system to use is made with the changing environmental statistics of rewards and states. In Chapter 2, three types of prediction error signals from the dual systems are found to guide the arbitration process in a reliability-based RL framework. Moreover, an exploratory analysis was conducted to test for alternative arbitration theories that utilize the cost-benefit analysis on the goal-directed (or MB) system. Understanding learning and decision-making would not be complete without knowing how our neural machinery implements these RL computations when a given system is engaged. The robustness and replicability of neural encoding of learning and decision signals from the MF and MB systems are essential to set a reassuring path for future neurocomputational work on dual systems. In Chapter 3, we address recent concerns over the existence of the MF system and its neural computations in a widely-used Markov decision task (two-step task). By applying a model-based functional magnetic resonance imaging (fMRI) approach to a large number of participants, we found both MF and MB learning signals in the human striatum and that neural patterns of decision utility across different RL-strategy groups further add to the evidence of ubiquitous MF computations in Markov decisions. It turns out the framework of dual systems could not only account for normal learning behaviors but also inform us of what actually goes wrong in mental disorders. In Chapter 4, we show that, via the reliability-based arbitration framework, the MF behavioral bias observed in participants with high obsessive-compulsive tendency could be attributed to an enhanced encoding of MB reward prediction error in the anterior cingulate cortex, a region previously implicated in the error-monitoring process. Chapter 1 introduces basic concepts and example algorithms in RL; we also review relevant theoretical and neuroscientific works to build the knowledge base for subsequent chapters. Chapter 5 discusses the significance of empirical findings in this thesis, the values of adopting some of the methodologies herein, potential future research directions on the dual systems, and implications in computational psychiatry.

CONTENTS

Acknowledgements iii
Abstract v
Contents vi
List of Figures vii
Chapter I: General Introduction 1
1.1 Prediction and Prediction Error 1
1.2 The Model-Free and Model-Based Control and their Neural Substrates 4
1.3 The Dual System and the Arbitration 12
1.4 Obsessive-Compulsive Disorder and the Error Signal 25
Chapter II: Behavioral Investigations on Various Types of Prediction Errors Contributing to the Arbitration Between the Model-Free and Model-Based Reinforcement Learning 38
Abstract 39
2.1 Introduction 40
2.2 Results 42
2.3 Discussion 54
2.4 Methods 57
Chapter III: Investigations of Reward Prediction and System Engagement with Group-Level and Individual-Level fMRI in the Two-Step Markov Decision-Making Task 74
Abstract 75
3.1 Introduction 76
3.2 Results 79
3.3 Discussion 89
3.4 Methods 94
Chapter IV: The Enhanced Encoding of Model-Based Reward Prediction Error in Anterior Cingulate Cortex and the Model-Free Behavioral Bias among Individuals with High Obsessive-Compulsive Tendency 118
Abstract 119
4.1 Introduction 120
4.2 Results 123
4.3 Discussion 128
4.4 Methods 132
Chapter V: General Discussion 136

LIST OF FIGURES

Number *Page*

2.1 The two-step task structure and the group-level behaviors 67
2.2 Manipulations of three reliability signals 68
2.3 Transition Sensitivity Measure 69
2.4 Condition-based arbitration model 70
2.5 Transition Sensitivity Measure in the Replication Sample 71
2.6 Evidence of Inter-Reliability Arbitration 71
2.7 Testing of alternative arbitration theories 72
2.8 Transition Sensitivity Measure on Alternative Arbitration Theories . 73
2.9 Computational Modeling on Alternative Arbitration Theories 73
3.1 FMRI two-step task, clustering of the group behaviors and the overall and sub-group behaviors 84
3.2 Group-level fMRI results of Decision Utility and Reward Prediction Error 86
3.3 Across-group difference and individual differences in decision utility encoding 89
4.1 Correlations between Behavioral Measures and Obsessive-Compulsive Tendency 125
4.2 Neural Correlations between the encoding strength of MB RPE and the Obsessive-Compulsive Tendency 128

Chapter 1

GENERAL INTRODUCTION

1.1 Prediction and Prediction Error

A dog will salivate to a piece of nice food, a natural biological response beneficial to food digestion, which essentially improves the dog's fitness. Pigeons will peck in the presence of food also for biological fitness. Such innate and involuntary responses are not restricted to an appetitive domain but could also be observed in some aversive situations. For example, a person would startle when a loud noise is heard. Such a startling response is triggered through the acoustic startle reflex in the central nervous system, which helps the organism to prepare protective actions as the loud noise could indicate the presence of potential threats. Hand-withdrawing would be the reaction when the hand touches a hot surface to avoid skin tissue damage. In psychology, such a response is also called an unconditioned response. Such responses are enabled through the biological "hardware" built into the process of evolution rather than being learned within the life span of an organism through experiences. Given that these types of innate responses are essential to the organism's fitness and survival, the responsive circuit is naturally selected and acts as the biological "prior" so that, for the organism, the maintenance of fitness could be efficient and life-threatening risks could be minimized without too much cost (sometimes could be life itself).

As an organism, to navigate the environment not only successfully but also efficiently, only having spontaneous innate reactions to biologically significant events is not sufficient. An organism capable of predicting such biologically significant events could prospectively manage physical and cognitive resources to react better. As our environment is not solely composed of biologically significant stimuli but also neutral events that often co-occur with the stimuli statistically, such neutral events could provide great predictive power over the biologically significant events for the organisms to learn. A shaking bush by itself might not elicit any innate biological response (unconditioned response), yet while on a hiking trail, it could predict some unknown threats (e.g., predators) being present. Being able to predict the potential present predator from the shaking bush is essential for the organism to activate the "fight-or-flight" biological circuit in time to increase the chance of survival compared to only starting to react when the predator itself is encountered. Fortunately, animals are gifted with the capability to make such predictions. Pavlov (2010) found that a dog, usually salivating at the food itself, would start to salivate to a sound (conditioned stimulus) if that sound is frequently paired with food (unconditioned stimulus), even when the food is not present. Such a phenomenon is called Pavlovian conditioning (or classical conditioning). In essence, although probably unconsciously, the dog makes the prediction of food through a neutral stimulus with experiences and starts to salivate at the neutral stimulus (conditioned response) for food digestion. The Rescorla-Wagner model (1972) provides a theoretical account of Pavlovian conditioning. The process that leads the dog to express the conditioned response involves repetitions of pairing the food with the sound. If the sound starts with a zero value to the animal ($V(sound) = 0$), the repetitious pairing process will enable the sound to eventually acquire a value that is equal to the value of the food itself $R(food)$. Such a process is implemented through a repetitious comparison between the learned value of sound $V(sound)$ and the actual food value $R(food)$, and whenever a mismatch is detected, there would be a certain portion of updates made to the quantity $V(sound)$ to reduce the degree of mismatch. Mathematically, the comparison and the degree of mismatch on trial t can be written as:

$$\delta_t = R_t(food) - V_t(sound); \tag{1.1}$$

The reward value of food $R_t(food)$ is given, and the difference term δ is called prediction error. The prediction error would be a non-zero term as long as the value prediction from the sound about the food $V_t(sound)$ is imperfect to match the food's actual value. One note is that if the reward value of food is not fixed (e.g., the change of satiety state), there is a possibility that the prediction error becomes negative. The Rescorla-Wagner model utilizes this prediction error to make updates to the prediction so that the $V(sound)$ eventually converges to $R_t(food)$:

$$V_{t+1}(sound) = V_t(sound) + \alpha \times \delta_t; \tag{1.2}$$

The α term is the learning rate, which decides how fast the conversion happens. Even with a relatively small learning rate, with enough trials of pairing between the sound and the food, the predictive value of the sound will converge to the reward value of the food. This model could explain the Pavlovian conditioning that the dog would eventually exhibit a conditioned response to an originally neutral stimulus (i.e., sound) due to its good prediction of the reward value.

However, not all the learning behaviors can be explained by the Rescorla-Wagner model. Second-order conditioning describes the phenomenon that once the classical conditioning (first-order) is established, if now repetitively pairing conditioned stimulus (e.g., sound) with a new neutral stimulus (e.g., light) independent of the reward, the conditioned response would eventually be expressed towards the new neutral stimulus as well. In the example of Pavlov's dog, the dog would salivate at the light if the light was later paired with the sound after the dog started to salivate at the sound, although the light has never been directly paired with the food. According to the Rescorla-Wagner model, a new neutral stimulus can only attain value when the reward is experienced on the same trial as the neutral stimulus is experienced. Then, according to this model, in the case of second-order conditioning, the value of light $V(light)$ would always stay at zero as no rewards were experienced along with the light, and thus no conditioned response to the light should be observed. The assumption that all relevant stimuli and biologically significant events take place on the same trial gives the Rescorla-Wagner model some limitation in accounting for learning behaviors that are temporally distant or independent of the trial structure. Hence, the temporal difference (TD) model in the field of reinforcement learning was proposed to explain learning behavior that takes place beyond a constrained trial structure (Sutton, 2018).

In the TD model, the events are characterized in the unit of timestamp rather than trial so that there can be multiple events occurring at different timestamps within a single trial. Formally, the TD model learns the predictive value of each state at a time point t as a summation of the actual reward in the next state and the discounted expected predictive value across all possible next states at $t + 1$:

$$V_t(s_t) = r_t(s_{t+1}) + \gamma \times E(V_t(s_{t+1})|s_t), 0 \leq \gamma \leq 1; \tag{1.3}$$

This is a recursive form of current-state value computation as a function of the current-state reward and discounted value in the next state (discount factor γ), also called the Bellman Equation (Bellman, 1957). Learning the true expected value of the next state $E(V_t(s_{t+1})|s_t)$ is not trivial as it is a probability-weighted average of state value $V_t(s_{t+1})$ across all possible identities s_{t+1}. If the learning is not perfect, the TD model also uses a prediction error signal, the difference between the predictive value of the current state and the actual value (i.e., encountered reward added with the discounted predictive value of the next state), to update the predictive value of the current state:

$$\delta_t = r_t(s_{t+1}) + \gamma \times V_t(s_{t+1}) - V_t(s_t); \tag{1.4}$$

$$V_{t+1}(s_t) = V_t(s_t) + \alpha \times \delta_t; \tag{1.5}$$

Critically, in the TD learning model, the prediction error not only factors in the potential reward encountered at the current time t (i.e., $r_t(s_t)$) but also takes into account the predictive value of future rewards $V_t(s_{t+1})$. This gives the TD model the flexibility to learn the predictive value of the current state based upon the previously learned predictive value of possible future states, even in the absence of immediate rewards (i.e., when $r_t(s_t) = 0$). Such a feature essentially gives the model the capacity to learn predictions not only within a trial but also beyond the trial, without any premise on the presence of rewards.

Going back to the example of second-order conditioning, after the dog fully learns the predictive value of sound $V(sound)$ and exhibits a conditioned response to the sound, according to the TD learning model, the pairing of light and sound would eventually drive the predictive value of light towards that of sound thanks to the temporal difference prediction error even without any rewards in the trial:

$$\delta_t = V_t(light) - V_t(sound); \tag{1.6}$$

$$V_{t+1}(light) = V_t(light) + \alpha \times \delta_t. \tag{1.7}$$

The TD learning model is a classic model-free reinforcement learning model that opens up the possibility of modeling the real-world problems of achieving predictions over events that are temporarily extended and unconstrained in terms of their relevance to the rewards. As we will see in the next section, the TD learning model can go beyond the classical conditioning paradigm and be expanded to describe instrumental conditioning in cases with or without a "world model" of the environment. These variations of the reinforcement learning model gained further support from the neuroscience literature in animal models and humans.

1.2 The Model-Free and Model-Based Control and their Neural Substrates

Going back to Pavlov's dog, after repetitive sound-food association, the dog would start to salivate at the sound through gradual mitigation of prediction errors. Now, if, along with the sound, two touch pads were simultaneously offered and only touching the left pad could lead to the delivery of food, the dog could gradually learn which pad to touch; more interestingly, if the right pad becomes the rewarding pad, the dog could change its preference of the pad to adapt. Such learning behavior contrasts classical conditioning as the behavior exhibits flexible patterns to maximize the potential rewards. The dog could not only predict the presentation of reward through the onset of a certain stimulus but also learn the stimulus-response association as the action itself has predictive power of the outcome.

In classical conditioning, it is essentially the stimulus-outcome contingency that is learned, and the TD learning model explains how the stimulus gradually acquires the prediction of the outcome along the process. Another important aspect of prediction in the real world is the ability to predict the consequences (e.g., how much reward would be obtained) upon executing certain actions (e.g., a dog touching the pad). It is thus helpful to know the proper action upon encountering a given stimulus to maximize biological fitness. The dog learning which pad to touch is one type of instrumental conditioning, where a stimulus-action association is learned through experiencing the relationship between an action and the outcome. Thorndike (2017) described such phenomenon as the "Law of effect" - the stimulus-response association would be strengthened if the response leads to rewarding events and weakened if non-rewarding events were experienced. Such a reinforcing process on the stimulus-response association is later characterized as habitual learning (Dickinson, 1985). Importantly, the TD learning model could be easily adapted to describe such a habitual learning process by adding the "action" component alongside the "state" component as the learning unit of the prediction of rewards. Formally, the learning rule for the predictive value of a specific action in a given state can be described as:

$$\delta_t = [r_{t+1} + \gamma \times \max_a V_{t+1}(s,a)] - V_t(s,a); \tag{1.8}$$

$$V_{t+1}(s,a) = V_t(s,a) + \alpha \times \delta_t. \tag{1.9}$$

This is also known as the Q-learning algorithm, an off-policy reinforcement learning algorithm. The only difference in the prediction error of instrumental conditioning here compared to classical conditioning is that for the predictive value of the future state, only the action maximizing the predictive value would be considered. The update through the prediction error is then made only to the actual stimulus-action pair.

Now, consider another example where a rat is placed in a newly exposed maze, and a reward is placed at one of many exits. The rat would eventually learn the efficient route leading to the reward with some experiences, as the "Law of Effect" would implicate that the route that led to the reward in the past would be strengthened to be retaken in the future. For the rat to run through the route that has been reinforced in the past, a set of state-action predictive values would be needed in the maze to guide the rat. As covered in the previous section, such state-action pair predictive values could be learned through TD errors where the states are the locations in the maze, and the actions are the possible turn-makings. Hence, acquiring the stimulus-response association in the maze is done through a model-free (MF) reinforcement learning algorithm, as only the reward history is relied on to learn the predictive value. Still in the context of maze navigation, if the rat is firstly exposed to the maze to explore but without any exposure to any reward in any exits, then if later a reward is introduced to the environment, it was found that the rat with an exposure experience of the maze was faster to learn the optimal route to the reward. This is what Tolman (1948) found as another possible instrumental learning behavior different from the "Law of Effect," and the concept of "cognitive map" was introduced accordingly. Specifically, in the exposure phase, although without any reward experience, the rat could latently learn the maze's structure and how a specific location (e.g., a potential reward location) could be reached from some proceeding locations a few steps earlier. Such a "cognitive map" could later be retrieved and utilized to efficiently reach the rewarding location — whenever a reward location is reached, the knowledge of how to reach the current location from a previous state with the appropriate action could be used recursively back to the starting location in the maze. The knowledge of maze structure could be mathematically described with the state transition matrix $T(s, a, s')$, which encodes the probability of transitioning from state s to state s' via action a. The learning of the state transition matrix could be achieved via the state prediction errors:

$$\delta_t^{SPE} = 1 - T_t(s, a, s'); \tag{1.10}$$

$$T_{t+1}(s, a, s') = T_t(s, a, s') + \eta \times \delta_t^{SPE}; \tag{1.11}$$

$$T_{t+1}(s, a, s'') = (1 - \eta) \times T_t(s, a, s''), \forall s'' \neq s'; \tag{1.12}$$

The state prediction error (SPE) is computed at each step for experienced state transitions $s \longrightarrow s'$; for all other non-visited states s'', a normalization procedure is used so that all possible transition probabilities starting from the state s would add up to 1.

With the knowledge of state transition structure in the maze, any information of the reward location in the maze would lead to efficient computation of up-stream state values thanks to the state-transition matrix:

$$V_t(s,a) = \sum_{s'} T_t(s,a,s') \times [r_t(s') + \max_{a'} V_t(s',a')]; \tag{1.13}$$

The upstream state-action value is essentially a transition-probability-weighted downstream value of rewards (if any) and the maximal state-action value upon all available actions. At the reward location s_r, $V(s_r, a) = 0$, for any action a. Such a learning algorithm is called a model-based (MB) forward planner, as it builds a "model" of the task structure — the state-transition matrix in the case of maze navigation, and conducts action planning in a prospective way through dynamic programming. Compared to the MF algorithm, the MB algorithm offers the agent learning flexibility upon environmental changes such as the availability of previously viable routes or the change of the reward location. In any such events, the MF algorithm can adapt to a new viable route through a relatively long history of experiences. In contrast, the MB algorithm can take advantage of the knowledge of the state structure to quickly re-figure a new route plan.

Besides the difference in knowledge of the environment, another contrasting feature between the MF and MB algorithms is the online access to the outcome value itself. Consider an example that two rats in the same maze have already learned the best route towards the cheese location in the maze, one via the MF algorithm and the other via the MB algorithm. Now, they are suddenly fed with the cheese to satiety; the "MB" rat would immediately be unmotivated to move along the best route towards the cheese in the maze. In contrast, the "MF" rat would still move along that route towards the cheese until it realized the cheese at the end was no longer valuable. Such behavioral difference is because the "MB" rat would always have access to the online value of the end goal while performing forward planning for every intermediate step along the best route. The change of outcome value for the rat would then provide an immediate update on the predictive value of every intermediate step back to the starting location in the maze. On the contrary, the "MF" rat only has access to the outcome value at the last step before reaching the cheese and consequently would not update all the upstream state-action values before eventually reaching the reward that has a changed value. A note is that the dynamic information of outcome value is not always exogenously sent to the acting agent in the MB learning model (as described as a sudden devaluation in the above example), and therefore a typical reward learning mechanism using MB reward prediction error, just as in the case of MF reward prediction error, could well be used to learn the actual value of the possible outcomes.

Now that we have covered the MF and the MB reinforcement learning algorithm accounting for a variety of learning behaviors (in both classical conditioning and instrumental learning), it is worth asking how the brain implements the key computational elements of the two models — specifically the predictive values of the reward-related events, and the prediction errors that are used to reach the accurate predictions of the reward and of the environment. We will first start with the neuroscience findings on the computational elements of the MF algorithm.

Dopamine (DA) neurons in the mid-brain regions, specifically in the ventral tegmental area (VTA) and substantia nigra, have been found to be involved in reward processing and motivation (Wise, 1982). In the example of classical conditioning, when the reward (i.e., unconditioned stimulus) was associated with the neutral stimulus (conditioned stimulus) at the very beginning, using electrophysiological recording in monkeys, it has been found that dopamine neurons increased their firing response when the reward was presented (Schultz, Dayan, and Montague, 1997). This observation suggests that the dopamine response encodes some information related to the reward. It is unclear whether the DA response signals the reward itself or the reward prediction error. Yet, it was also observed that as more association between the reward and the neutral stimulus was experienced and the conditioned response was observed to the conditioned stimulus, the dopamine neurons started to respond to the conditioned stimulus as if the previously neutral stimulus was rewarding. Such observation would suggest the DA response aligns more with a reward prediction error account, which could be illustrated through a TD learning model (Sutton, 2018). Through the scope of the TD learning model, when the reward was first presented, a positive reward prediction error would be associated with the outcome onset time as the learned value for the time point is zero. As time goes on, the neutral stimulus acquires predictive value. A positive prediction error would arise when the neutral stimulus is presented as there is no expectancy of a valuable stimulus in predicting reward in any proceeding events. Moreover, as the TD learning model would predict, no further dopamine responses were observed after the conditioning was acquired when an expected reward occurred, and inhibited responses were observed when the expected reward was omitted, as it indicated a negative reward prediction error.

Beyond the simple conditioning paradigm, when the neutral stimuli were repetitively paired with reward delivery with varying probabilities, the dopamine neurons responded with the same strength as the TD errors (Fiorillo, Tobler, and Schultz, 2003). In a "blocking" paradigm, where a neutral stimulus that robustly predicts the reward is presented along with a new stimulus, followed by the predicted reward, the new stimulus does not elicit any conditioned response and becomes the "blocked" stimulus. When the blocked stimulus was followed by an absence of rewards, no change in dopamine responses was detected along the process, as no reward prediction errors would be elicited, consistent with the TD-error account (Waelti, Dickinson, and Schultz, 2001). The neural activity mimicking the TD-error has also been observed in human brains. Using functional magnetic c resonance imaging (fMRI), blood-oxygenation-level-dependent (BOLD) signals were measured as a proxy of neural activity in the corresponding regions across the whole brain, which offers more complete coverage of brain regions than electrophysiological recordings can do. There has been human neuroimaging work showing that the predictability of rewarding stimulus activated ventral striatum and ventromedial prefrontal cortex (Berns et al., 2001). Subsequent experiments further pin down the source of the observed neural activation in the human striatum as a temporal prediction error when the reward is unexpectedly delivered or omitted (McClure, Berns, and Montague, 2003). To strictly confirm the relationship between striatal neural activity and TD error in the reinforcement learning algorithm, TD reward prediction error signals in appetitive classical conditioning were computed from a TD model (Schultz, Dayan, and Montague, 1997) and found to be correlated with the BOLD signal in the human ventral striatum and in the orbitofrontal cortex (J. P. O’Doherty et al., 2003). A further study on prediction error signals in the human brain found the prediction error signals have overlapping representations in the dorsal striatum across different types of reinforcers (Valentin and J. P. O’Doherty, 2009), suggesting the role of striatum encoding a common error signal for reward learning in general.

Given dopamine neurons having projections to targeted regions such as the ventral striatum, nucleus accumbens, and prefrontal cortex, which have been associated with reward processing and instrumental learning (e.g., goal-directed control), the DA neurons sit well in the brain to help learn through the past reward history and reinforce the stimulus-action pair that leads more to the reward (e.g., rat navigating in the maze for rewards). Hence, the TD-error signal presumably leads to a better estimation of predictive values in the environment and could be well hosted in the activity of DA neurons.

Remember, when the rat navigates in the maze, it can use the history of trial and error to learn the best route to the reward location; presumably, the striatum facilitates such learning by communicating reward prediction errors to achieve good reward prediction from the environment. But if the rat is exposed to the maze beforehand, it could comprehend the environment by building a cognitive map of the maze (or state-transition knowledge) through the quantity of state prediction errors (SPE), a key element of the model-based algorithm. Initial studies on state-based learning focused on a qualitative error signal during the occurrence of novel stimuli or expectation violation of neutral stimuli (e.g., auditory stimulus) using EEG and fMRI (Opitz et al., 1999; Strobel et al., 2008). To investigate how a quantitative state-based error signal could be implemented in latent learning, Gläscher and colleagues (2010) used a probabilistic Markov decision task, in which the abstract state-action-state transition knowledge could help the human participants enhance the performance. A forward planner model-based algorithm, as described before, was fit to the behavior to compute the SPE signal for a model-based fMRI analysis, and the SPE signals were found to be correlated with the BOLD signal measured in intraparietal sulcus and lateral prefrontal cortex.

In the example of maze navigation, either learning the best route through the reward history with sequences of RPE using an MF algorithm or prospectively implementing the best route with the help of a cognitive map composed of state-transition probabilities in an MB manner, predictive values at the end of the day need to be computed in both types of strategies for all the possible intermediate states between the reward location and the starting state. If indeed the MF and MB type of strategies could well characterize the key computations underlying maze navigation and, more generally, the reward learning process, the brain should also represent the predictive value signal, learned either directly through the reward prediction errors or through weighing the state-action-state probability, to guide an optimal action selection process where the action leading to a higher predictive value should be favored.

As the ventromedial prefrontal cortex (vmPFC) had been implicated in reward-based decision-making and value encoding (Thorpe, Rolls, and Maddison, 1983; Bechara, Tranel, and Damasio, 2000; J. O'Doherty et al., 2003; Knutson et al., 2003; Padoa-Schioppa and Assad, 2006; Plassmann, J. O'doherty, and Rangel, 2007; Tobler et al., 2007) as well as in encoding abstract rules (Wallis, Anderson, and Miller, 2001; Genovesio et al., 2005), the region was studied specifically in terms of its role of value encoding in state-based learning. A probabilistic reversal learning task was used where the choice signals indexing the value from a standard reinforcement algorithm (MF) and that from a Bayesian Hidden State Markov Model (MB) could be dissociated. It was found that the BOLD activity in the medial prefrontal cortex, which was previously mainly implicated in MF value computation, explained the value signal better from the abstract state-based model than that from an MF algorithm (Hampton, Bossaerts, and O'doherty, 2006). To further pin down and dissociate the neural sources of predictive value signals of the MF and MB algorithm, Beierholm et al. (2011) have used a bandit-like neuroeconomic task where different options are associated with different reward probabilities, and the participants were incentivized to learn the reward probabilities and prioritized choosing options with higher reward probabilities. As the task successfully drives participants' behavior either to an experience-based MF strategy or a task-information-based Bayesian MB strategy, the computed value signal from both strategies has been found to correlate in non-overlapping regions in the medial prefrontal cortex (mPFC). Interestingly, the psychophysiological interaction analysis found the correlation between the ventral striatum and the ventromedial prefrontal cortex was increased when behaviors were more under the MF control, suggesting a neural process of MF learning integration between the error signal encoding region and the value encoding region. For the parallel computation of values pertinent to the MF and MB system, researchers have also created separate contexts where tree-search planning (i.e., MB) was demanded and reward-related choices were overtrained (i.e., MF) for human participants to maximize the rewards (Wunderlich, Dayan, and Dolan, 2012). Intriguingly, it was found that the value difference required in a tree-search strategy correlated significantly with the BOLD activity in caudate, whereas the values learned in the extensive training correlated significantly with the BOLD activity in putamen.

Given the past work on the neural representation of predictive value signals from the model-free and model-based systems, both systems' value signals seem to be represented independently and in parallel mainly in the medial prefrontal cortex and striatum, receiving learning signals from the ventral striatum to form an accurate representation of the reward environment. Given the availability of the two systems and their distinct learning modules, the acting agent has the flexibility to engage one system or the other for adaptive behavioral control on decision-making based on empirical needs. In the next section, we will cover multiple theoretical frameworks on the arbitration between the dual system and evidence from some computational and neural studies.

1.3 The Dual System and the Arbitration

The mental processes overall can roughly fall into two categories: a "fast" mode and a "slow" mode, which have been historically called "System 1" and "System 2". Such characterization originated from behavioral phenomena in experimental economics, social judgment, and animal learning (Kahneman, Frederick, et al., 2002; Loewenstein and O'Donoghue, 2004; Liberman, 2003; Killcross and Blundell, 2002; Dickinson and B. Balleine, 2002). Kahneman (2011) described "System 1" as fast, instinctive, and emotional and "System 2" as slow, deliberative, and logical. Hence, "System 1" involves mental processes that are relatively simple and effortless (e.g., driving on a usual commute route), whereas the processes in "System 2" entail mental expenses or cognitive efforts (e.g., navigating in a new city).

Such dual systems have realizations in the animal learning behaviors we have described in previous sections. In the example of a rat navigating in the maze, the rat could simply rely on the experiences of obtaining a reward or not to trigger necessary responses at each location of the maze, thus forming a stabilized route for which the learning process involves a sufficient number of trials and errors. Such a process falls into the category of "System 1", as a simple stimulus-response association is learned. On the other hand, if the rat has been exposed to the maze to explore beforehand, the rat could utilize the state-action-state association (or cognitive map) learned during the exposure phase to more efficiently figure out the route to the reward. As the rat has to "mentally simulate" the possible route by computing values at each node of branches in a tree-like decision space, and as the "tree" becomes larger, the mental simulation would become more laborious and intractable, which could well be characterized as a "System 2" process. Now, if the rat that learned a stabilized route through "System 1" faces the situation that the original route is blocked, then it has to relearn a new route from scratch through trials and errors. If the rat using the cognitive map faces the same situation, an alternative route could be figured out quickly through prospective planning over the maze. Also, as alluded to in the previous section, if the rat is fed to satiety, "System 1" would still implement the previously learned route as the stimulus-response association is intact so far, while "System 2" would stop pursuing the reward as the online update of the new outcome value (much lower) devalues all the state value in the maze. In cases of sudden reward environment change, such as a route block or a reward devaluation, "System 1" seems to persist in behavioral expression that was learned and effective before the change but could not effectively adjust towards the updated goal, and thus psychologists described such behavioral patterns as driven by habitual control (Dickinson, 1985). Conversely, "System 2" could adaptively adjust the behavioral policy online upon these sudden changes to adapt to continually pursue the current goal, and such behavioral patterns are characterized by psychologists as goal-directed control (Dickinson, 1985). Linking back to the reinforcement learning algorithms we introduced before, the habitual control can be mimicked with an MF agent learning the reward environment, and the goal-directed behavior can be achieved with an MB agent learning both the state and reward information.

As both habitual and goal-directed control could potentially be engaged in animal learning contexts, the question is under what circumstances each control would be engaged. We will describe the experimental findings first and then introduce the relevant arbitration theories.

Relevant to the outcome devaluation procedure described above, researchers have trained rodents to perform a stimulus-action pair (S1-A1) to obtain an outcome of O1. Then, the outcome (O1) was devalued, and the response rate of the rodents was compared to the control group, which did not go through the outcome devaluation procedure in terms of the response rate of action A1. If it is habitual control driving the behavior, then the response rate of A1 should be comparable between the devaluation group and the controls, displaying insensitivity to outcome devaluation. If goal-directed control drives behavior, the response rate of A1 should be reduced in rodents that experienced outcome devaluation, as the updated outcome values of O1 and O2 are considered. It turned out the rodents would display behaviors under habitual control if they were moderately trained and would display goal-directed behaviors if they were over-trained (Adams, 1982; B. W. Balleine and Dickinson, 1998). In human participants, Tricomi and colleagues (2009) used a free-operant task for training and found as it went from moderate training to extensive training, human participants shifted their behavior from being goal-directed to being habitual, echoing earlier findings in animal literature.

Given the good characterization of the MF and MB systems on the habitual and goal-directed behavior, respectively, a theoretical framework based upon the dual system uncertainty (or prediction accuracy) has been proposed to explain the arbitration between the habitual and goal-directed control from a computational perspective (Daw, Niv, and Dayan, 2005). Daw and colleagues (2005) framed the habitual and goal-directed control as the two ends of a spectrum going from an algorithm that is computationally simple but inflexible (model-free) to an algorithm that is computationally complex but could afford dynamic planning online (model-based).

Past research has found that lesions of the dopamine system (Faure et al., 2005) or the dorsolateral striatum (Yin, Knowlton, and B. W. Balleine, 2004) could disrupt the control shift to habit formation through extensive training, indicating persistent use of the MB system. Also, lesions of prelimbic cortex (B. W. Balleine and Dickinson, 1998; Coutureau and Killcross, 2003; Killcross and Coutureau, 2003) and prefrontal-associated regions of dorsomedial prefrontal cortex (Yin, Ostlund, et al., 2005) in rodents, and lesions of orbitofrontal cortex in monkeys (Izquierdo, Suda, and Murray, 2004) were found to lead to impaired MB system, that is, outcome-insensitive habits were displayed even with moderate training. Such neural findings suggest a potential parallel MF and MB system in the brain that could compete for actual behavioral control.

Consider an example where the rat was trained to perform a chain of two state-action pairs to successfully retrieve a reward. There are two actions that are always available for the rat to choose: 1) lever-pressing and 2) magazine-entering. At the initial state, the rat needs to press the lever (action 1) to have the food delivered and then enter the magazine (action 2) to obtain the food. Entering the magazine at the initial state or pressing the lever at the subsequent state (i.e., when the food is delivered) would lead to the no-reward states. Given all potential states, including the reward/no-reward state, the MB system would need to construct a correct tree structure (i.e., state-action-state transitions) to compute the state-action values. The MF system relies on experiences of reward or no reward, given an action at each state, to estimate the values. When each system estimates the state-action value in its own manner, there would be estimation uncertainty about the true value, and such uncertainty indicates how accurate each system is. The core of the uncertainty-based arbitration theory (Daw, Niv, and Dayan, 2005) is that the system with higher value estimation accuracy, that is, low uncertainty, should be favored in use. Then for the system selected to use, the possible actions were implemented with the probabilities that are proportional to their corresponding values. A Bayesian version of the two RL systems was implemented (Dearden, Friedman, and Russell, 1998; Mannor et al., 2004), and the posterior variance of the action value estimates serves as the uncertainty of the system in use for that action (here we focused only on the distal action, lever press, for simplicity). When the rat just starts with the task, the MB system has relatively lower value uncertainty as the initial stream of reward information can inform the action value more efficiently through propagations in the learned decision tree, whereas the MF system needs a sufficient number of bootstrapping to reach a stable estimation of the true value. Hence the distal action (pressing lever) would exhibit outcome devaluation sensitivity when the training is moderate. However, as the training goes on, the MF system could take advantage of the large number of reward samples to have the variance of value estimates gradually asymptote to a low level, while the MB system would asymptote to a relatively higher level of uncertainty, suffering from the "computational noise" in the additional step of tree search due to practical value approximations when using the MB system. As a consequence, after extensive training, the MF system would have a lower uncertainty for the distal action than the MB system would have. Hence, an insensitivity of the action to the outcome devaluation would be expressed.

In the situation where two concurrent actions are available to choose to obtain two different outcomes respectively (Kosaki and Dickinson, 2010), the prediction from the theory about the arbitration process would change. Since more data samples are needed to reduce the variance of value estimates from the MF system for the two concurrent actions, the resulting uncertainty of the MF system would asymptote to a higher level than that of the MB system according to the simulation, leading to the sustained use of the MB system, hence displaying sustained outcome devaluation sensitivity throughout the entire training process.

As the uncertainty-based arbitration theory (Daw, Niv, and Dayan, 2005) explained well the classic habit formation observed in rodents, a direct RL implementation of the theory, the reliability-based arbitration, was then proposed and tested in human decision making (Lee, Shimojo, and O'doherty, 2014; O'Doherty et al., 2021). The reliability-based arbitration theory approximates the system uncertainty through experienced prediction errors from each system. When a system's reliability is relatively high, then such a system should be assigned a large weight when controlling behavior. When deciding on the system weight, the reliabilities of both systems are considered.

For the reliability of the MF system, unsigned reward prediction errors serve as the proxy estimate, and the quantity of the unsigned reward prediction error Ω is learned through:

$$\Delta\Omega_t = \eta \times (|RPE|_t - \Omega_t); \tag{1.14}$$

$$\Omega_{t+1} = \Omega_t + \Delta\Omega_t, \tag{1.15}$$

where η is the learning rate for the unsigned reward prediction error. The reliability of the MF system is then defined as:

$$\chi_{MF} = \frac{(RPE_{max} - \Omega)}{RPE_{max}}, \tag{1.16}$$

where RPE_{max} is the experienced maximal reward prediction error. Hence the larger the learned unsigned reward prediction error, the lower the MF reliability is.

For the reliability of the MB system, Bayesian inference is used to estimate the conditional probability of zero and positive state prediction error (as SPE could not go negative). In detail:

$$P(SPE|\theta) = \begin{cases} \theta_1 & \text{if } SPE > \omega \\ \theta_0 & \text{if } 0 \leq SPE \leq \omega \end{cases}, \tag{1.17}$$

where ω denotes the tolerance level for a positive SPE (graininess of determination); θ_0 denotes the probability of making zero SPEs and θ_1 denotes the probability of making positive SPEs, and $\theta_0 + \theta_1 = 1$. The conditional probability of making zero or positive SPE is assumed to be from a Dirichlet prior distribution:

$$(\theta_0, \theta_1) \sim Dirichlet(\lambda_0, \lambda_1), \tag{1.18}$$

where $\lambda_0 + \lambda_1 = 1$. Assuming there is a set D of T events that consist of a subset D_{SPE_0} of zero SPE events and a subset $D_{SPE_1}$of positive SPE events:

$$D = \{SPE^1, SPE^2, ..., SPE^T\} \tag{1.19}$$

$$D_{SPE_0} = \{SPE^i\}, \text{where } 0 \leq SPE^i \leq \omega \text{ and } i \in \{1, 2, ..., T\} \tag{1.20}$$

$$D_{SPE_1} = \{SPE^i\}, \text{where } SPE^i > \omega \text{ and } i \in \{1, 2, ..., T\} \tag{1.21}$$

The posterior probability distribution of making zero or positive SPE would then be:

$$P(\theta|_D) \sim Dirichlet(\lambda_0 + |D_{SPE_0}|, \lambda_1 + |D_{SPE_1}|) \tag{1.22}$$

which is because the Dirichlet distribution is the conjugate prior distribution of a categorical distribution. Here $|D_{SPE_0}|$ and $|D_{SPE_1}|$ denote the cardinality of the corresponding subset.

Then the expected value and variance of the posterior probability distribution can be calculated as:

$$E(\theta_j|_D) = \frac{(1 + |D_{SPE_j}|)}{2 + |D|}; \quad j \in 0, 1 \tag{1.23}$$

$$Var(\theta_j|_D) = \frac{(1+|D_{SPE_j}|)(1+|D_{SPE_k}|)}{(2+|D|)^2(3+|D|)}; \quad j,k \in \{0,1\}; \quad k \neq j \tag{1.24}$$

The MB reliability informed by the SPE signal is then defined as:

$$\chi_{MB} = \frac{\chi_0}{\chi_0 + \chi_1}, \tag{1.25}$$

where $\chi_j = \frac{E(\theta_j|_D)}{Var(\theta_j|_D)}$. This mean-variance ratio is the inverse of a dispersion index that has been used to characterize the uncertainty in communication channels (Janesick et al., 1987) and efficiency of information transfer in neurons (Ma et al., 2006). Hence, the MB system reliability would decrease with the number of positive SPE observations and with inconsistency within all the observations.

Once the reliability metrics of the two systems have been established as above, to control the extent of each system influencing the behaviors, the weight of a system (P_{MB}, probability of MB) is decided through a dynamical two-state transition rule governed by the two systems' reliabilities χ_{MF} and χ_{MB}. The weight for a given system is probabilistic, and the weights of the two systems sum up to 1. Also, considering the fact that engaging the MB system could be more demanding in terms of cognitive efforts than engaging the MF system due to sophisticated tree-search computation, the transition between the high MF weight to high MB weight is biased in a way that the MF system is favored, when everything else equal. Also, through simulations, this specific RL implementation of the reliability-based arbitration was found to capture well the classic finding in instrumental learning that behavioral control would first get dominated by goal-directed control and shift to habitual control as the amount of training increases (Adams and Dickinson, 1981; Adams, 1982; B. W. Balleine and Dickinson, 1998; Tricomi, B. W. Balleine, and O'Doherty, 2009).

To test this implementation of the reliability-based arbitration model, Lee and colleagues (2014) used a Markov decision task for human participants to make a decision of a two-action sequence to transition through two layers of states for reward collections. There are fixed types of tokens with various reward amounts associated with each end-state in the task space. Importantly, the task is designed to elicit different levels of MF reliability in reward prediction and MB reliability in state prediction throughout the experiment. The types of tokens for rewards have periods where all types of tokens are valued ("flexible") vs. only a specific type of token is valuable ("specific," the specific type of token can change) in order to create periods when the unsigned reward prediction errors encountered in the MF system are larger ("specific" period) or smaller ("flexible" period). Additionally, the state-action-state transition probabilities were also manipulated with two conditions: low vs. high state-action-state transition uncertainty (90%/10% vs. 50%/50%), with which the level of state prediction errors could be set to low vs. high levels, respectively. Through examination of participants behaviors as well as computational model fitting, supportive evidence for the reliability-based arbitration hypothesis was found that during "specific" token periods (i.e., low MF reward prediction reliability) and low station-transition uncertainty periods (i.e., high MB state prediction reliability), more MB-consistent behaviors were expressed. The reliability-based arbitration model also predicted actual choices well and captured the shift between MF-consistent and MB-consistent behaviors throughout the experiment.

Besides the arbitration framework, which leverages the uncertainty (or prediction accuracy) of each system when making reward and state predictions, there have been other theories that trade off the benefits and costs of the dual systems for reward maximization (Keramati, Dezfouli, and Piray, 2011; Pezzulo, Rigoli, and Chersi, 2013). When a rat was put in a T-maze for choosing one of two concurrent options for a reward, it was observed that the rat would pause and exhibit head movements to both options before finally making a choice (Edward Chace Tolman, 1938; Edward Chace Tolman, 1939; Muenzinger, 1938). Such behavior is called "vicarious trial-and-error"(VTE) and indicates the deliberation process during planning. As training continues, the frequency of such VTE behaviors declines (Hu, Xu, and Gonzalez-Lima, 2006; Munn, 1950), and the decision time shortens, suggesting the control shifts from a goal-directed system to a habitual system (Redish, Jensen, and Johnson, 2008). Such observation highlights an important feature in the goal-directed system: although the deliberation could lead to potentially good information about the entire environment and sudden change in the environment could be properly incorporated, it comes with the time cost associated with deliberation. In contrast, the habitual system only learns the stimulus-action association, which is a simple computational process that elicits little time consumption, although it comes with the inflexibility to adapt to sudden environmental changes. As a result, using a goal-directed system would entail opportunity costs due to the fact that the deliberation time could otherwise be used for potential reward collection through a fast habitual system, with a caveat that the fast habitual system might be inaccurate in terms of evaluating the reward information in the environment, thus not necessarily harvesting more rewards in an absolute sense. Hence the selection between a goal-directed system and a habitual system becomes a problem of speed-accuracy trade-off, and Keramati and colleagues (2011) proposed a normative arbitration framework that compares the benefits and costs of using the goal-directed system to decide the actual system in use for instrumental behavior.

The speed-accuracy trade-off framework uses the MF and MB algorithms to computationally approximate the habitual and goal-directed control, respectively, and it relies on a key assumption that the goal-directed system would have an almost perfect value estimation of the rewards in the environment with more deliberation time. Thus, the benefits of the goal-directed system could be quantified as potential value gain when having the perfect information (value of perfect information, VPI; Howard, 1966), and the costs are then the potential rewards that could be accrued through a habitual system during the deliberation time. If the benefit (i.e., VPI) is larger than the cost, then the goal-directed system should be engaged; otherwise, the habitual system should be used. Specifically, the VPI can be considered as the value gain for a given state-action pair through the following metric (Dearden, Friedman, and Russell, 1998):

$$Gains_{s,a}(Q^*(s,a)) = \begin{cases} \hat{Q}^H(s,a_2) - Q^*(s,a) & \text{if } a = a_1 \text{ and } Q^*(s,a) < \hat{Q}^H(s,a_2), \\ Q^*(s,a) - \hat{Q}^H(s,a_1) & \text{if } a \neq a_1 \text{ and } Q^*(s,a) > \hat{Q}^H(s,a_1), \\ 0 & \text{otherwise.} \end{cases} \tag{1.26}$$

Here a_1 and a_2 are the best and second-best actions, which $\hat{Q}^H$ denote the value learned through the habitual system through the temporal difference learning algorithm, while the Q^* denotes the true value that could be learned through the goal-directed system. Hence, learning the true value of a state-action pair (s,a) is beneficial via a goal-directed system when 1) the current learned best action is found to be worse than the second-best action, or 2) some non-best action is actually better than the current best action. Given the definition of $Gains_{s,a}$, where the true values (i.e., Q^*) are not accessible but could be approximated by integrating over the probability distribution $\hat{Q}^H$ learned through the habitual system, the value of perfect information for a given state-action pair could be defined as:

$$VPI(s,a) = E(Gains_{s,a}(Q^*(s,a)) = \int_{-\infty}^{\infty} Gains_{s,a}(x)Pr(Q^H(s,a) = x)\,dx. \tag{1.27}$$

Thus, the VPI for each state-action pair, the potential gain of engaging the goal-directed system, could be computed from the value distribution learned in the habitual system. Intuitively, for a given non-best action in a state and the best action in that state, the VPI is proportional to the overlapping area between the value distributions of the two actions in consideration.

If assuming that a fixed amount of deliberation time for evaluating VPI is needed, denoted by τ, then the cost of engaging the goal-directed control could be defined as the amount of reward that could be obtained with a habitual system during the deliberation time period. If the average rate of reward obtained per unit of time is denoted as $\overline{R}$, mathematically, the cost of engaging the goal-directed system is $\overline{R} \times \tau$. Here the average rate of reward obtained per unit of time could be learned as:

$$\overline{R}_{t+1} = \overline{R}_t + \sigma \times (r_t - \overline{R}_t), \tag{1.28}$$

where $(r_t - \overline{R}_t)$ denotes the prediction error and σ is the learning rate. Hence, $VPI(s, a)$ and $\overline{R} \times \tau$ are compared against each other to determine whether to engage the goal-directed system.

In the example where the rat has to perform a sequence of two actions (i.e., lever pressing and magazine entering) to obtain a reward, the shift of dominant control from the goal-directed system to the habitual system when the training amount increases could be explained well by the speed-accuracy trade-off arbitration theory. When the reward sampling process just starts, the VPI of the two actions is high as there is a large uncertainty around which action has a truly higher value, manifested by the large overlapping area between the two value distributions. When the outcome is devalued after moderate training, the probability of the VPI being higher than the opportunity cost $\overline{R} \times \tau$ is relatively high, and therefore, in the test phase, the behavior would be driven by the goal-directed system. As training continues, the estimation of the true values of the two actions becomes more and more accurate (with less overlapping area between the two value distributions), and eventually, the VPI of the two actions would be smaller than the opportunity cost. Thus, when devaluation happens after extensive training, it is more likely that the habitual system would be the dominant behavioral control. Also, it accounts for the observation that as the training goes on, the reaction time would drop as the habitual system takes control, and the deliberation process would be saved in such cases (Edward Chace Tolman, 1938).

In the experiment where two concurrent actions are available for obtaining two different outcomes (Kosaki and Dickinson, 2010), because the value estimation from the habitual system through reinforcement learning could not entirely eliminate the variance of value estimation due to forgetting, and especially due to the fact that the two estimated values have the same reinforcing strength, the estimated value distribution of the two concurrent actions could not separate to enough of a degree so that the overlapping area would be high, leading to a higher level of VPI than the opportunity cost in a sustained manner. Consequently, no matter how long the training duration is, the goal-directed system would be the dominant behavioral control, explaining well the empirical finding in such a task setting (Kosaki and Dickinson, 2010). An interesting prediction from the speed-accuracy trade-off on the experiment with two concurrent action-outcome pairs is that if the reinforcing strengths (or intrinsic outcome values) are different enough, this will lead to a decreased overlapping area between the value distribution of the two actions, and thus a decrease of VPI. This suggests that as the training goes on, eventually the behavior would become habitual. This prediction would contrast with what the uncertainty-based arbitration framework would predict: as the relative value between the two concurrent actions is not considered, but only the estimation uncertainty of the dual systems is evaluated, the uncertainty-based arbitration would suggest a sustained use of the goal-directed system regardless of the relative value between the two actions.

Another arbitration model that leverages the benefit and cost of engaging the goal-directed/MB system is the "Mixed Instrumental Controller" (MIC; Pezzulo, Rigoli, and Chersi, 2013). The model is "mixed" in the sense that when an agent navigates in the reward environment and learns the state-action value, the goal-directed/MB system could conditionally join and contribute to the value estimation process along with the habitual/MF system to help with action selection. The habitual/MF system in MIC uses the typical Q-learning process (Watkins and Dayan, 1992) that learns the Q-values of a state-action pair based upon experienced rewards, and importantly it is the default system used for value estimations in MIC. The goal-directed/MB system in MIC is featured as using an internal model of the environment to perform mental simulations that sample potential actions and the associated outcomes. Performing the mental simulations brings the benefits of having better action value estimates, while at the same time, performing such mental simulations comes with a cost of cognitive efforts (Gershman and Daw, 2012) and the reward being delayed due to mental simulation time (Shadmehr, 2010). The benefit and cost of engaging mental simulations are compared to determine whether the goal-directed/MB system would be used to help estimate the values. Specifically, for the benefit of mental simulations, Pezzulo and colleagues (2013) used a simple method to approximate the metric of "Value of Information" (VoI; Howard, 1966), which quantifies the value of gaining more reward information via the mental simulations. For action 1 ($Act1$) of the two possible actions($Act1$ and $Act2$, its VoI is defined as:

$$VoI_{Act1} = \frac{C_{Act1}}{|Q_{Act1} - Q_{Act2}| + \epsilon}. \tag{1.29}$$

Here C_{Act1} denotes the uncertainty of the value estimate of action 1, Q_{Act1}. The denominator is the absolute difference between the learned Q-values of the two possible actions through past experiential learning or mental simulations and ϵ is added to ensure a non-zero denominator. This metric of VoI is then compared against the cost of mental simulations, set as a fixed threshold γ. If $VoI_{Act1} > \gamma$, mental simulations for that action and possible subsequent actions will be performed to have more pseudo-observations, which will be used to have a better posterior estimate of the Q-values for action selection; if $VoI_{Act1} < \gamma$, then the MIC would rely on the cached Q-values learned through past experiences for action selection. Based upon this metric, it can be seen that the benefit of mental simulation is high 1) when there is a lot of uncertainty around the value of $Act1$, or 2) when there is little difference between the estimated Q-values of the possible actions, according to this method. Additionally, the uncertainty of the executed action (assuming $Act1$) C_{Act1} and the Q-value of the action Q_{Act1} are learned online based on real observations when navigating through the task.

Pezzulo and colleagues (2013) used MIC to simulate agent behaviors in a double T-maze environment where a sequence of left-or-right decisions need to be made to obtain potential rewards placed in the maze. To test how changes of VoI could lead to changes in the behavioral expression of mental simulations, the authors manipulate the variance of available reward amount at specific maze locations as well as the relative value difference of concurrent rewards across multiple maze locations to achieve various levels of VoI in different experimental settings. In a simple and stable setting where the reward is placed in one location and the variance of the reward is small, the typical transfer from the goal-directed system to the habitual system as a function of training amount is observed, featuring a decrease in the number of mental simulations and the length of mental simulations, which is consistent with the classic behavioral findings (Adams and Dickinson, 1981; Adams, 1982; B. W. Balleine and Dickinson, 1998; Tricomi, B. W. Balleine, and O'Doherty, 2009) and the neural activity of reduced hippocampal forward sweeps and ventral striatal covert reward expectations (Van Der Meer and Redish, 2009). In addition, the MIC model also produces a testable prediction in this simple reward setting when varying the reward variance: the mental simulations can sustain for a longer period at the beginning of the setting when the reward variance is larger, but gradually the "cache" system would dominate due to sharp relative value difference of available actions, potentially explaining the patterns of forward sweeps in hippocampus as a function of environmental uncertainty (Gupta et al., 2010). Also, additional predictions of the mental simulation patterns are made with simulations of the MIC model: a) when there is a sudden change in reward location, the mental simulation is reintroduced to learn the new action-reward contingency due to elevated uncertainty of values; and b) when both the reward variance (i.e., high uncertainty of value estimates) and the reward availability at various locations (i.e., low relative value difference) are manipulated to be large, the mental simulation process sustains for the entire training period without complete habitualization, due to both high uncertainty of value estimates and small relative value difference of possible actions. These are all interesting theoretical predictions from the MIC model that need further empirical testing in a maze navigation setting.

From the perspective of cost-benefit analysis and with the MF/MB approximations of habitual/goal-directed system, empirical studies have been run to test relevant arbitration theories with the help of a two-step Markov decision task where MF and MB algorithm could be dissociated (Daw, Gershman, et al., 2011). By pairing the primary two-step task with a secondary numerical Stroop task that taxes working memory (Waldron and Ashby, 2001), it was firstly found that high cognitive load induced by the Stroop task drives behavior towards using the MF algorithm in the primary two-step task as the MB algorithm could be hard to implement due to insufficient cognitive resources (Otto et al., 2013), highlighting the impact of cognitive cost on the arbitration process. An arbitration framework that leverages the practical reward advantage against the cognitive costs was then proposed for empirical testing (Kool, Gershman, and Cushman, 2017): if the high reward stakes and the accuracy advantage carried by the MB system, which leads to an estimation of its overall reward advantage, could offset its cognitive costs, then the MB system should be favored over the MF system. Kool and colleagues (2017) used a variant of the two-step Markov task (Kool, Cushman, and Gershman, 2016; Doll et al., 2015) with manipulation of high vs. low reward stakes incorporated, to test how the usage of MB system changes when incentivized with different levels of rewards, assuming the rewards could manage to compensate the cognitive effort or not. Critically, it has been established that the degree of using the MB algorithm in this specific two-step task variant is positively associated with the reward rate (i.e., average rewards collected per trial; Kool, Cushman, and Gershman, 2016), which suggests that high stake trials, compared to low stake trials, should promote the usage of the MB system due to its enlarged reward advantage. Through behavioral analysis and computational model fitting, such prediction was confirmed. To further resolve a confound that the MB system would be favored due to high reward stakes regardless of the actual reward accuracy benefit the MB system could bring, the authors ran a second experiment where participants completed the original two-step task (Daw, Niv, and Dayan, 2005) where there is no performance difference between using the MF system and the MB system. No effects of high vs. low reward-stakes manipulation on the degree of expressing MB-consistent behavior were found in the original version of the two-step task. The two experiments together offered the empirical evidence that a cost-benefit analysis evaluating the computationally expensive MB system's reward advantage against its disadvantage in occupying more cognitive resources could potentially underlie the arbitration process between the MB and MF systems.

Summarizing the efforts so far in characterizing the arbitration process, past research has focused on different learning targets and the accuracy of predicting such targets, as well as the benefits and costs due to the fundamental difference in the computations of the dual systems. The early theory has focused on the uncertainty of how each system could represent and predict the environmental reward or state predictions (Daw, Niv, and Dayan, 2005), which characterizes well the classic findings in studies on instrumental control. Importantly, the uncertainty-based arbitration theory also gains supportive evidence through specific approximations of system reliability through prediction errors in reinforcement learning by conducting computational modeling and model-based fMRI analysis on key variables in the arbitration process (Lee, Shimojo, and O'doherty, 2014). Later, the cost difference of engaging the dual systems is taken into account to evaluate the value of engaging the computationally costly goal-directed/MB system (Keramati, Dezfouli, and Piray, 2011; Pezzulo, Rigoli, and Chersi, 2013), which could explain well previous findings in terms of the "vicarious trial-or-error" (VTE) behavior and the neural patterns of hippocampal forward sweeps. The utility of reward maximization (i.e., accuracy) to offset computational costs has also been considered, and there has been experimental support that the computation-demanding MB system should be favored, especially when it gains reward advantage over the MF system (Kool, Gershman, and Cushman, 2017). To further pin down the key variables in the arbitration of the dual systems, future work is needed to devise an experiment where multiple model candidates have different predictions in the dynamics of arbitration, so that model performance in capturing the behaviors and the alignment between neural data and the hypothesized arbitration process could be assessed.

1.4 Obsessive-Compulsive Disorder and the Error Signal

Obsessive-compulsive disorder (OCD) is characterized by experiencing uncontrollable recurring thoughts, corresponding to the "obsessive" component, or by engaging in pointless excessive behavior that makes up the "compulsive" component of the disorder, or both (Franklin and Foa, 2011). Excessive hand-washing is an example of the "compulsive" behaviors in OCD. The adaptive goal of hand-washing is hygiene maintenance, yet when hand-washing is expressed to an abnormal extent in terms of frequency and duration, the benefit could typically be offset by the adverse consequence it will bring — the skin abrasion and the time wasted with no hygiene gain. Hence, compulsiveness can be characterized as a type of relentless and repetitive behavior despite its potential adverse consequences (Robbins et al., 2012). As illustrated in the hand-washing example, the agent expressing the compulsive behavior seemingly miscalculates the potential benefits and costs of the hand-washing action and the associated goal, as though the potential hygiene benefit brought by the excessive hand-washing would surpass the costs it induces due to its time-consuming nature, and stopping the washing action might cause higher costs than when the action is executed endlessly. Indeed, it has been suggested that the repetitive behaviors are intentionally expressed to avoid unwanted or aversive consequences (Salkovskis, 1985; McFall and Wollersheim, 1979; Stanley Rachman, 1998). Thus, the expressed compulsiveness could be due to some "cognitive bias" allocated to the value of the action options in the environment (with stopping-the-action as an option as well) to resolve the aversive worrying thoughts (Salkovskis et al., 2000) due to potential adverse consequences. According to this "cognitive" account, through a potentially biased value attribution, compulsive behaviors are a type of goal-directed or purposeful behavior to relieve worry or anxiety and to avoid the imagined aversive consequences (Rachman, 1976).

Given what has been covered on the habitual and goal-directed system, another account for the "compulsiveness" observed in OCD proposes that such compulsive behaviors arise due to the imbalanced allocation between the dual systems. In other words, OCD patients have no problem in correctly representing the benefits and costs of performing potential actions in the environment but, at the action-execution stage, suffer a maladaptive system allocation problem such that the goal-directed behavior cannot be expressed to a sufficient extent to effectively balance out the influence from the habitual system (Gillan and Robbins, 2014). Based upon the common neural circuit, the ’frontal-striatum’ circuit (Alexander, DeLong, and Strick, 1986), which is involved in habits (B. W. Balleine and J. P. O’doherty, 2010; Yin and Knowlton, 2006) and OCD (Alexander, DeLong, and Strick, 1986; Milad and Rauch, 2012; Haber and Heilbronner, 2013), it was suggested that OCD could be due to the maladaptive functioning of the habitual learning (Graybiel and Rauch, 2000).

Experimentally, through training in an appetitive instrumental learning paradigm where a specific action needs to be performed upon a given stimulus to receive rewards, it has been tested that OCD patients displayed learning of the stimulus-response association (habitual learning) to a greater extent but had difficulty in adapting to the updated outcome value, and were thus impaired in learning the response-outcome association or goal-directed learning (Gillan, Papmeyer, et al., 2011). Using a task of economic choices, it was also found that compared to healthy controls, OCD patients display reduced "potential regret," which served as a goal-directed marker underlying counterfactual processing for better performance (Gillan, Morein-Zamir, Kaser, et al., 2014). As typical compulsive behaviors are expressed to avoid potential aversive consequences, habitual/goal-directed learning has also been examined in the avoidance behavior. Specifically through training to avoid electric shocks, OCD patients expressed persistent avoidance behavior towards the devalued stimulus, indicating overreliance on habitual learning in the aversive domain (Gillan, Morein-Zamir, Urcelay, et al., 2014).

As for the explanations for the observed imbalance in habitual and goal-directed control in OCD patients, it could be due to deficits in the learning of action-outcome association; hence, the goal-directed process is affected, and the intact stimulus-action association is instead relied on (Gillan, Papmeyer, et al., 2011). Conversely, it could also be the reason that the stimulus-action association is established to an abnormally high degree that the normal learning or execution of the action-outcome association gets shadowed. Considering the theoretical account and empirical evidence of arbitration between the model-free and model-based reinforcement learning (Daw, Niv, and Dayan, 2005; Lee, Shimojo, and O’doherty, 2014), it could be the case that, as a third possible underlying cause, among OCD patients, both the stimulus-action link (habitual) and the action-outcome link (goal-directed) is intact, but, critically, the arbitration mechanism between the two systems is actually impaired (Kim et al., 2024).

To characterize the psychological mechanism of obsessive-compulsive disorder, the "error" signal is a fundamental driving factor underlying the expressed behavior. The mismatch between the expected consequent state of certain action and the perceived consequence, which is essentially an "error" signal, could induce error-correction behavior (e.g., checking or action repetition) until the mismatch as perceived is resolved (Pitman, 1987; Kate D Fitzgerald and Taylor, 2015). Such behavior would become maladaptive when the error signal is perceived as exaggeratedly large or when the errors are persistently perceived even after the correcting behavior. Thus the obsessive thoughts and subjective feelings of incomplete performance would be elicited due to the unresolved error. As a consequence, error-correction behavior would be overly expressed to address the perceived "large" errors, leading to repetitive but pointless compulsive behaviors.

Indeed, the role of abnormal "error" signals in OCD is well reflected based upon evidence from neuroscientific research. In electroencephalography (EEG), a negative deflection of electrical potential was observed starting at the event of error commission and reaches its negative maximum around 100ms post the error (Falkenstein et al., 1991; Gehring, Coles, et al., 1995), and such negative electrophysiological signal is called event-related negativity (ERN). This ERN signal has been empirically studied and traced back to its neural origin — the anterior cingulate cortex (ACC), with the approach of functional magnetic resonance imaging (fMRI; Ito et al., 2003; Holroyd et al., 2004), brain lesion research (Stemmer et al., 2004), and dipole source modeling (Dehaene, Posner, and Tucker, 1994). In a study with event-related fMRI, Carter and colleagues (1998) used a variant of the Continuous Performance Test (AX-CPT; Barch et al., 1997) to elicit response error and to create trials with high response competition. When examining the BOLD signal within the human brain, it was found that ACC expressed greater activity on error trials, implying its role in error detection, which, in relation to OCD, reflects the evaluation of the mismatch between the planned action execution and the actual action. Moreover, ACC also expressed greater activity in trials with high response conflict, suggesting ACC could also serve as a conflict detector in the error-prevention system underlying the OC behavior, which was further established with additional studies (Botvinick et al., 1999; Van Veen et al., 2001). Accordingly, the neural hypothesis of ACC being the conflict detector was supported by the evidence that ACC activity peaked before response in correct high-conflict trials as it resolved potential conflict in time but peaked after response in error trials as the unfinished evaluation of the ongoing conflict lingered (Veen and Carter, 2002).

Considering the role of ACC as both the error detector and the conflict detector, there were empirical studies specifically testing how activity in ACC shows a difference in OCD compared to controls. With respect to the role of ACC as the error detector, OCD patients showed enhanced ERN signal, possibly from ACC, compared to controls during action monitoring, and the degree of this enhancement was correlated with the severity of the OC symptoms (Gehring, Himle, and Nisenson, 2000). For better localization of the error processing mechanism in the brain, fMRI studies were conducted, and indeed, the activation in ACC was higher in OCD patients than in controls during error trials in the task of AX-CPT (Ursu et al., 2003); and in a "flanker interference" task (B. A. Eriksen and C. W. Eriksen, 1974), a similar effect of hyperactivity to errors in OCD was found in rostral ACC (Kate Dimond Fitzgerald et al., 2005). Such findings aligned well with the psychological mechanism that obsessive-compulsive behaviors are driven by hypersensitivity to the experienced error and manifested through subsequent persistent correction for the unsatisfactory performance (but for a possible confound of negative affect, see Luu, Collins, and Tucker, 2000). In parallel, although the task performance was comparable between the OCD patients and controls in the task of AX-CPT, a hyperactivity effect of ACC in OCD was still found during high-conflict trials (Ursu et al., 2003). Hence, the potential role of ACC as the conflict detector explains well the fact that although OCD patients can have unimpaired performance in cognitive tasks (Galderisi et al., 1995), performing the tasks is constantly accompanied by the subjective feeling of error and doubt. Further validating ACC's role in detecting error and conflict underlying the obsessive-compulsive behavior, the severity of OC symptoms was also found to positively correlate with the error-related hyperactivity of BOLD activity in ACC in the flanker interference task (Kate Dimond Fitzgerald et al., 2005) and have a trend of positive correlations with both the error-related and conflict-related BOLD hyperactivity in ACC in the task of AX-CPT (Ursu et al., 2003). Also, a meta-analysis of 9 fMRI studies showed that when comparing the whole-brain BOLD activity difference during errors between OCD and healthy controls on inhibitory control tasks (e.g., stop, go/no-go, Stroop, Simon, flanker, anti-saccade, and multisource interference), greater BOLD activation in the dorsal part of ACC was found in OCD patients than in healthy controls (Norman et al., 2019). Together, from both the electrophysiological and the fMRI studies, ACC as the error/conflict detector serves as a necessary module in establishing the psychological model of OCD.

To summarize the psychological models of OCD we have introduced so far, there are two accounts from the perspective of the action selection process: 1) an account of biased value attribution and 2) an account of the imbalance between the habitual and goal-directed system. The biased value attribution account of OCD proposes that the benefits and costs of potential action and no-action are computed when considering potential consequences of action/no-action, and such cost-benefit analysis leads to a goal-oriented behavior where the repetitive action execution minimizes the cost due to potential adverse consequences and reduces the worrying thoughts of potential aversiveness. For the account considering the imbalance between the habitual and goal-directed system, the occurrence of OC behavior could be driven by a more weighted emphasis on the expression of learned stimulus-action association, whereas the expression of the action-outcome association is compromised, which could be due to 1) an over-representation of the stimulus-action association, 2) an impaired cognitive link between action and outcome, or 3) an abnormal arbitration mechanism that the brain favors the habitual system over the goal-directed system. At the same time, OCD could be approached through the scope of an error correction/prevention process. Neural evidence has been presented that OCD patients show ACC hyperactivity in error/conflict scenarios compared to healthy controls, highlighting the role of the internal "error" signal (conflict as an unrealized error) in OCD, given ACC's general role in error/conflict detection in action monitoring. What is unknown yet is whether there could be an integrative model to associate the neuropsychological model of OCD in terms of resolving "error" and any of the computational models on the action selection process. One potential node for such model integration lies in the role of the "error" signal as it, in theory, could serve as a proxy estimate of the habitual/goal-directed system's reliability which drives the arbitrator to favor the more reliable system (i.e., the system generating fewer prediction errors; Daw, Niv, and Dayan, 2005; Lee, Shimojo, and O'doherty, 2014; O'Doherty et al., 2021). Thus, it implies that a biased neural representation of the errors from habitual and goal-directed systems could cause the apparent overreliance on the habitual system in OCD. To empirically test for evidence of such an integrative model, the framework of model-free and model-based system in reinforcement learning (Daw, Gershman, et al., 2011) could provide the necessary computational components — MF and MB prediction errors, to approximate the "error" signal from the habitual and goal-directed system. Such empirical works could help unravel the model of OCD and find the neural targets for its treatment.

References

Adams, Christopher D (1982). “Variations in the sensitivity of instrumental responding to reinforcer devaluation”. In: *The Quarterly Journal of Experimental Psychology* 34.2, pp. 77–98.

Adams, Christopher D and Anthony Dickinson (1981). “Instrumental responding following reinforcer devaluation”. In: *The Quarterly Journal of Experimental Psychology Section B* 33.2b, pp. 109–121.

Alexander, Garrett E, Mahlon R DeLong, and Peter L Strick (1986). “Parallel organization of functionally segregated circuits linking basal ganglia and cortex”. In: *Annual review of neuroscience* 9.1, pp. 357–381.

Balleine, Bernard W and Anthony Dickinson (1998). “Goal-directed instrumental action: contingency and incentive learning and their cortical substrates”. In: *Neuropharmacology* 37.4-5, pp. 407–419.

Balleine, Bernard W and John P O’doherty (2010). “Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action”. In: *Neuropsychopharmacology* 35.1, pp. 48–69.

Barch, Deanna M et al. (1997). “Dissociating working memory from task difficulty in human prefrontal cortex”. In: *Neuropsychologia* 35.10, pp. 1373–1380.

Bechara, Antoine, Daniel Tranel, and Hanna Damasio (2000). “Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions”. In: *Brain* 123.11, pp. 2189–2202.

Beierholm, Ulrik R et al. (2011). “Separate encoding of model-based and model-free valuations in the human brain”. In: *Neuroimage* 58.3, pp. 955–962.

Berns, Gregory S et al. (2001). “Predictability modulates human brain response to reward”. In: *Journal of neuroscience* 21.8, pp. 2793–2798.

Botvinick, Matthew et al. (1999). “Conflict monitoring versus selection-for-action in anterior cingulate cortex”. In: *Nature* 402.6758, pp. 179–181.

Carter, Cameron S et al. (1998). “Anterior cingulate cortex, error detection, and the online monitoring of performance”. In: *Science* 280.5364, pp. 747–749.

Coutureau, Etienne and Simon Killcross (2003). “Inactivation of the infralimbic prefrontal cortex reinstates goal-directed responding in overtrained rats”. In: *Behavioural brain research* 146.1-2, pp. 167–174.

Daw, Nathaniel D, Samuel J Gershman, et al. (2011). “Model-based influences on humans’ choices and striatal prediction errors”. In: *Neuron* 69.6, pp. 1204–1215.

Daw, Nathaniel D, Yael Niv, and Peter Dayan (2005). “Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control”. In: *Nature neuroscience* 8.12, pp. 1704–1711.

Dearden, Richard, Nir Friedman, and Stuart Russell (1998). “Bayesian Q-learning”. In: *Aaai/iaai* 1998, pp. 761–768.

Dehaene, Stanislas, Michael I Posner, and Don M Tucker (1994). “Localization of a neural system for error detection and compensation”. In: *Psychological science* 5.5, pp. 303–305.

Dickinson, Anthony (1985). “Actions and habits: the development of behavioural autonomy”. In: *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 308.1135, pp. 67–78.

Dickinson, Anthony and Bernard Balleine (2002). “The role of learning in the operation of motivational systems”. In: *Stevens’ handbook of experimental psychology* 3, pp. 497–533.

Doll, Bradley B et al. (2015). “Model-based choices involve prospective neural activity”. In: *Nature neuroscience* 18.5, pp. 767–772.

Eriksen, Barbara A and Charles W Eriksen (1974). “Effects of noise letters upon the identification of a target letter in a nonsearch task”. In: *Perception & psychophysics* 16.1, pp. 143–149.

Falkenstein, Michael et al. (1991). “Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks”. In: *Electroencephalography and clinical neurophysiology* 78.6, pp. 447–455.

Faure, Alexis et al. (2005). “Lesion to the nigrostriatal dopamine system disrupts stimulus-response habit formation”. In: *Journal of Neuroscience* 25.11, pp. 2771–2780.

Fiorillo, Christopher D, Philippe N Tobler, and Wolfram Schultz (2003). “Discrete coding of reward probability and uncertainty by dopamine neurons”. In: *Science* 299.5614, pp. 1898–1902.

Fitzgerald, Kate D and Stephan F Taylor (2015). “Error-processing abnormalities in pediatric anxiety and obsessive compulsive disorders”. In: *CNS spectrums* 20.4, pp. 346–354.

Fitzgerald, Kate Dimond et al. (2005). “Error-related hyperactivity of the anterior cingulate cortex in obsessive-compulsive disorder”. In: *Biological psychiatry* 57.3, pp. 287–294.

Franklin, Martin E and Edna B Foa (2011). “Treatment of obsessive compulsive disorder”. In: *Annual review of clinical psychology* 7.1, pp. 229–243.

Galderisi, Silvana et al. (1995). “Neuropsychological slowness in obsessive–compulsive patients: is it confined to tests involving the fronto-subcortical systems?” In: *The British Journal of Psychiatry* 167.3, pp. 394–398.

Gehring, William J, Michael GH Coles, et al. (1995). “A brain potential manifestation of error-related processing”. In: *Electroencephalography and Clinical Neurophysiology-Supplements only* 44, pp. 261–272.

Gehring, William J, Joseph Himle, and Laura G Nisenson (2000). "Action-monitoring dysfunction in obsessive-compulsive disorder". In: *Psychological science* 11.1, pp. 1–6.

Genovesio, Aldo et al. (2005). "Prefrontal cortex activity related to abstract response strategies". In: *Neuron* 47.2, pp. 307–320.

Gershman, Samuel J and Nathaniel D Daw (2012). "Perception, action, and utility: the tangled skein". In.

Gillan, Claire M, Sharon Morein-Zamir, Muzaffer Kaser, et al. (2014). "Counterfactual processing of economic action-outcome alternatives in obsessive-compulsive disorder: further evidence of impaired goal-directed behavior". In: *Biological psychiatry* 75.8, pp. 639–646.

Gillan, Claire M, Sharon Morein-Zamir, Gonzalo P Urcelay, et al. (2014). "Enhanced avoidance habits in obsessive-compulsive disorder". In: *Biological psychiatry* 75.8, pp. 631–638.

Gillan, Claire M, Martina Papmeyer, et al. (2011). "Disruption in the balance between goal-directed behavior and habit learning in obsessive-compulsive disorder". In: *American Journal of Psychiatry* 168.7, pp. 718–726.

Gillan, Claire M and Trevor W Robbins (2014). "Goal-directed learning and obsessive–compulsive disorder". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1655, p. 20130475.

Gläscher, Jan et al. (2010). "States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning". In: *Neuron* 66.4, pp. 585–595.

Graybiel, Ann M and Scott L Rauch (2000). "Toward a neurobiology of obsessive-compulsive disorder". In: *Neuron* 28.2, pp. 343–347.

Gupta, Anoopum S et al. (2010). "Hippocampal replay is not a simple function of experience". In: *Neuron* 65.5, pp. 695–705.

Haber, Suzanne N and Sarah R Heilbronner (2013). "Translational research in OCD: circuitry and mechanisms". In: *Neuropsychopharmacology* 38.1, p. 252.

Hampton, Alan N, Peter Bossaerts, and John P O'doherty (2006). "The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans". In: *Journal of Neuroscience* 26.32, pp. 8360–8367.

Holroyd, Clay B et al. (2004). "Dorsal anterior cingulate cortex shows fMRI response to internal and external error signals". In: *Nature neuroscience* 7.5, pp. 497–498.

Howard, Ronald A (1966). "Information value theory". In: *IEEE Transactions on systems science and cybernetics* 2.1, pp. 22–26.

Hu, Dan, Xiaojuan Xu, and Francisco Gonzalez-Lima (2006). “Vicarious trial-and-error behavior and hippocampal cytochrome oxidase activity during Y-maze discrimination learning in the rat”. In: *International Journal of Neuroscience* 116.3, pp. 265–280.

Ito, Shigehiko et al. (2003). “Performance monitoring by the anterior cingulate cortex during saccade countermanding”. In: *Science* 302.5642, pp. 120–122.

Izquierdo, Alicia, Robin K Suda, and Elisabeth A Murray (2004). “Bilateral orbital prefrontal cortex lesions in rhesus monkeys disrupt choices guided by both reward value and reward contingency”. In: *Journal of Neuroscience* 24.34, pp. 7540–7548.

Janesick, James R et al. (1987). “Scientific charge-coupled devices”. In: *Optical Engineering* 26.8, pp. 692–714.

Kahneman, Daniel (2011). “Thinking, fast and slow”. In: *Farrar, Straus and Giroux*.

Kahneman, Daniel, Shane Frederick, et al. (2002). “Representativeness revisited: Attribute substitution in intuitive judgment”. In: *Heuristics and biases: The psychology of intuitive judgment* 49.49-81, p. 74.

Keramati, Mehdi, Amir Dezfouli, and Payam Piray (2011). “Speed/accuracy trade-off between the habitual and the goal-directed processes”. In: *PLoS computational biology* 7.5, e1002055.

Killcross, Simon and Pam Blundell (2002). “Associative representations of emotionally significant outcomes”. In: *Emotional cognition*, p. 13.

Killcross, Simon and Etienne Coutureau (2003). “Coordination of actions and habits in the medial prefrontal cortex of rats”. In: *Cerebral cortex* 13.4, pp. 400–408.

Kim, Taekwan et al. (2024). “Neurocomputational model of compulsivity: deviating from an uncertain goal-directed system”. In: *Brain* 147.6, pp. 2230–2244.

Knutson, Brian et al. (2003). “A region of mesial prefrontal cortex tracks monetarily rewarding outcomes: characterization with rapid event-related fMRI”. In: *Neuroimage* 18.2, pp. 263–272.

Kool, Wouter, Fiery A Cushman, and Samuel J Gershman (2016). “When does model-based control pay off?” In: *PLoS computational biology* 12.8, e1005090.

Kool, Wouter, Samuel J Gershman, and Fiery A Cushman (2017). “Cost-benefit arbitration between multiple reinforcement-learning systems”. In: *Psychological science* 28.9, pp. 1321–1333.

Kosaki, Yutaka and Anthony Dickinson (2010). “Choice and contingency in the development of behavioral autonomy during instrumental conditioning.” In: *Journal of Experimental Psychology: Animal Behavior Processes* 36.3, p. 334.

Lee, Sang Wan, Shinsuke Shimojo, and John P O’doherty (2014). “Neural computations underlying arbitration between model-based and model-free learning”. In: *Neuron* 81.3, pp. 687–699.

Liberman, MD (2003). “Reflexive and reflective judgment processes: a social cognitive neuroscience approach”. In: *Social judgments: Implicit and explicit processes*, pp. 44–67.

Loewenstein, George and Ted O’Donoghue (2004). “Animal spirits: Affective and deliberative processes in economic behavior”. In: *Available at SSRN 539843*.

Luu, Phan, Paul Collins, and Don M Tucker (2000). “Mood, personality, and self-monitoring: negative affect and emotionality in relation to frontal lobe mechanisms of error monitoring.” In: *Journal of experimental psychology: General* 129.1, p. 43.

Ma, Wei Ji et al. (2006). “Bayesian inference with probabilistic population codes”. In: *Nature neuroscience* 9.11, pp. 1432–1438.

Mannor, Shie et al. (2004). “Bias and variance in value function estimation”. In: *Proceedings of the twenty-first international conference on Machine learning*, p. 72.

McClure, Samuel M, Gregory S Berns, and P Read Montague (2003). “Temporal prediction errors in a passive learning task activate human striatum”. In: *Neuron* 38.2, pp. 339–346.

McFall, Miles E and Janet P Wollersheim (1979). “Obsessive-compulsive neurosis: A cognitive-behavioral formulation and approach to treatment”. In: *Cognitive Therapy and Research* 3, pp. 333–348.

Milad, Mohammed R and Scott L Rauch (2012). “Obsessive-compulsive disorder: beyond segregated cortico-striatal pathways”. In: *Trends in cognitive sciences* 16.1, pp. 43–51.

Muenzinger, Karl F (1938). “Vicarious trial and error at a point of choice: I. A general survey of its relation to learning efficiency”. In: *The Pedagogical Seminary and Journal of Genetic Psychology* 53.1, pp. 75–86.

Munn, Norman L (1950). “Handbook of psychological research on the rat; an introduction to animal psychology.” In.

Norman, Luke J et al. (2019). “Error processing and inhibitory control in obsessive-compulsive disorder: a meta-analysis using statistical parametric maps”. In: *Biological Psychiatry* 85.9, pp. 713–725.

O’Doherty, John et al. (2003). “Dissociating valence of outcome from behavioral control in human orbital and ventral prefrontal cortices”. In: *Journal of neuroscience* 23.21, pp. 7931–7939.

O’Doherty, John P et al. (2003). “Temporal difference models and reward-related learning in the human brain”. In: *Neuron* 38.2, pp. 329–337.

O’Doherty, John P et al. (2021). “Why and how the brain weights contributions from a mixture of experts”. In: *Neuroscience & Biobehavioral Reviews* 123, pp. 14–23.

Opitz, Bertram et al. (1999). “The functional neuroanatomy of novelty processing: integrating ERP and fMRI results”. In: *Cerebral cortex* 9.4, pp. 379–391.

Otto, A Ross et al. (2013). “The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive”. In: *Psychological science* 24.5, pp. 751–761.

Padoa-Schioppa, Camillo and John A Assad (2006). “Neurons in the orbitofrontal cortex encode economic value”. In: *Nature* 441.7090, pp. 223–226.

Pavlov, P Ivan (2010). “Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex”. In: *Annals of neurosciences* 17.3, p. 136.

Pezzulo, Giovanni, Francesco Rigoli, and Fabian Chersi (2013). “The mixed instrumental controller: using value of information to combine habitual choice and mental simulation”. In: *Frontiers in psychology* 4, p. 92.

Pitman, Roger K (1987). “A cybernetic model of obsessive-compulsive psychopathology”. In: *Comprehensive psychiatry* 28.4, pp. 334–343.

Plassmann, Hilke, John O’doherty, and Antonio Rangel (2007). “Orbitofrontal cortex encodes willingness to pay in everyday economic transactions”. In: *Journal of neuroscience* 27.37, pp. 9984–9988.

Rachman, S (1976). “The modification of obsessions: A new formulation”. In: *Behaviour Research and Therapy* 14.6, pp. 437–443.

Rachman, Stanley (1998). “A cognitive theory of obsessions”. In: *Behavior and cognitive therapy today*. Elsevier, pp. 209–222.

Redish, A David, Steve Jensen, and Adam Johnson (2008). “Addiction as vulnerabilities in the decision process”. In: *Behavioral and brain sciences* 31.4, pp. 461–487.

Rescorla, Robert A (1972). “A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement”. In: *Classical conditioning, Current research and theory* 2, pp. 64–69.

Robbins, Trevor W et al. (2012). “Neurocognitive endophenotypes of impulsivity and compulsivity: towards dimensional psychiatry”. In: *Trends in cognitive sciences* 16.1, pp. 81–91.

Salkovskis, Paul M (1985). “Obsessional-compulsive problems: A cognitive-behavioural analysis”. In: *Behaviour research and therapy* 23.5, pp. 571–583.

Salkovskis, Paul M et al. (2000). “Responsibility attitudes and interpretations are characteristic of obsessive compulsive disorder”. In: *Behaviour research and therapy* 38.4, pp. 347–372.

Schultz, Wolfram, Peter Dayan, and P Read Montague (1997). “A neural substrate of prediction and reward”. In: *Science* 275.5306, pp. 1593–1599.

Shadmehr, Reza (2010). “Control of movements and temporal discounting of reward”. In: *Current opinion in neurobiology* 20.6, pp. 726–730.

Stemmer, Brigitte et al. (2004). “Error detection in patients with lesions to the medial prefrontal cortex: an ERP study”. In: *Neuropsychologia* 42.1, pp. 118–130.

Strobel, Alexander et al. (2008). “Novelty and target processing during an auditory novelty oddball: a simultaneous event-related potential and functional magnetic resonance imaging study”. In: *Neuroimage* 40.2, pp. 869–883.

Sutton, Richard S (2018). “Reinforcement learning: An introduction”. In: *A Bradford Book*.

Thorndike, Edward (2017). *Animal intelligence: Experimental studies*. Routledge.

Thorpe, SJ, ET Rolls, and S Maddison (1983). “The orbitofrontal cortex: neuronal activity in the behaving monkey”. In: *Experimental brain research* 49, pp. 93–115.

Tobler, Philippe N et al. (2007). “Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems”. In: *Journal of neurophysiology* 97.2, pp. 1621–1632.

Tolman, Edward C (1948). “Cognitive maps in rats and men.” In: *Psychological review* 55.4, p. 189.

Tolman, Edward Chace (1938). “The determiners of behavior at a choice point.” In: *Psychological review* 45.1, p. 1.

– (1939). “Prediction of vicarious trial and error by means of the schematic sowbug.” In: *Psychological review* 46.4, p. 318.

Tricomi, Elizabeth, Bernard W Balleine, and John P O’Doherty (2009). “A specific role for posterior dorsolateral striatum in human habit learning”. In: *European Journal of Neuroscience* 29.11, pp. 2225–2232.

Ursu, Stefan et al. (2003). “Overactive action monitoring in obsessive-compulsive disorder: evidence from functional magnetic resonance imaging”. In: *Psychological science* 14.4, pp. 347–353.

Valentin, Vivian V and John P O’Doherty (2009). “Overlapping prediction errors in dorsal striatum during instrumental learning with juice and money reward in the human brain”. In: *Journal of neurophysiology* 102.6, pp. 3384–3391.

Van Der Meer, Matthijs AA and A David Redish (2009). “Covert expectation-of-reward in rat ventral striatum at decision points”. In: *Frontiers in integrative neuroscience* 3, p. 458.

Van Veen, Vincent et al. (2001). “Anterior cingulate cortex, conflict monitoring, and levels of processing”. In: *Neuroimage* 14.6, pp. 1302–1308.

Veen, Vincent van and Cameron S Carter (2002). “The timing of action-monitoring processes in the anterior cingulate cortex”. In: *Journal of cognitive neuroscience* 14.4, pp. 593–602.

Waelti, Pascale, Anthony Dickinson, and Wolfram Schultz (2001). “Dopamine responses comply with basic assumptions of formal learning theory”. In: *Nature* 412.6842, pp. 43–48.

Waldron, Elliott M and F Gregory Ashby (2001). “The effects of concurrent task interference on category learning: Evidence for multiple category learning systems”. In: *Psychonomic bulletin & review* 8.1, pp. 168–176.

Wallis, Jonathan D, Kathleen C Anderson, and Earl K Miller (2001). “Single neurons in prefrontal cortex encode abstract rules”. In: *Nature* 411.6840, pp. 953–956.

Watkins, Christopher JCH and Peter Dayan (1992). “Q-learning”. In: *Machine learning* 8, pp. 279–292.

Wise, Roy A (1982). “Neuroleptics and operant behavior: the anhedonia hypothesis”. In: *Behavioral and brain sciences* 5.1, pp. 39–53.

Wunderlich, Klaus, Peter Dayan, and Raymond J Dolan (2012). “Mapping value based planning and extensively trained choice in the human brain”. In: *Nature neuroscience* 15.5, pp. 786–791.

Yin, Henry H and Barbara J Knowlton (2006). “The role of the basal ganglia in habit formation”. In: *Nature Reviews Neuroscience* 7.6, pp. 464–476.

Yin, Henry H, Barbara J Knowlton, and Bernard W Balleine (2004). “Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning”. In: *European journal of neuroscience* 19.1, pp. 181–189.

Yin, Henry H, Sean B Ostlund, et al. (2005). “The role of the dorsomedial striatum in instrumental conditioning”. In: *European Journal of Neuroscience* 22.2, pp. 513–523.

Chapter 2

BEHAVIORAL INVESTIGATIONS ON VARIOUS TYPES OF PREDICTION ERRORS CONTRIBUTING TO THE ARBITRATION BETWEEN THE MODEL-FREE AND MODEL-BASED REINFORCEMENT LEARNING

ABSTRACT

It has been proposed that the reflexive model-free (MF) system and a reflective model-based (MB) system could both guide decision-making through reinforcement learning, and an arbitration theory posits that the acting system between the two should more reliably make predictions about the environment. However, direct behavioral investigation of how control is allocated between the two systems remains scarce. In this chapter, to investigate the arbitration process, we used a novel variant of the two-step task with manipulations of three reliability signals from the two systems — MF reward prediction error, MB state prediction error, and MB reward prediction error, and performed statistical analysis on the arbitration behavior from two online samples of participants. We not only found behavioral evidence of MF reward prediction error and MB state prediction error driving control allocation, further supporting previous computational and neural evidence of reliability-based arbitration framework, but also found that MB reward prediction reliability also contributes to control allocation significantly. In an exploratory analysis, various reliability signals were found to interact to influence how behavior is guided by a certain system. Moreover, in a post-hoc way, we also tested alternative arbitration theories focusing on the cost-benefit analysis of the two systems but did not find positive results. Overall, this chapter expands the dictionary of basic building blocks to further study sophisticated arbitration behavior in real-world settings.

2.1 Introduction

In the animal psychology literature, a contrasting pair of control systems have been used to characterize certain stereotyped animal behaviors in learning paradigms: habitual control and goal-directed control (Balleine and Dickinson, 1998). Habitual control involves a system learning the association between a stimulus and an action, in which the action is reinforced through past rewards. As such, upon a given stimulus, an action would be executed as if there has been value cached to the action. In contrast, upon a given stimulus, the goal-directed control learns the association of action and its subsequent state, which is usually, but not limited to, an outcome state. In other words, the knowledge of the state-action-state transitions in the environment is needed to execute good goal-directed control. Thus, a specific goal state can be achieved by implementing the linking actions between different upstream states. Thanks to the development in the field of computer science and specifically in reinforcement learning, the aforementioned dual control systems have their own computational correspondence with the habitual control described as a model-free (MF) system and the goal-directed control described as a model-based (MB) system. Intuitively, the MF system learns a cached stimulus-action value through trials and errors, whereas the MB system computes the action value online through a "World Model," which encodes the state-action-state information. Empirical experiments with the computational approaches and the measure of underlying neural signals (Daw, Gershman, et al., 2011; Gläscher et al., 2010; Wunderlich, Dayan, and Dolan, 2012) have shown such computational characterization of MF vs. MB system attains good explanatory power on human learning.

Humans show a mixed use of the habitual (or MF) and goal-directed (or MB) systems in empirical studies of learning (Daw, Gershman, et al., 2011; Kool, Cushman, and Gershman, 2016; Wunderlich, Smittenaar, and Dolan, 2012; Dezfouli and Balleine, 2013; Otto et al., 2013; Smittenaar et al., 2013; Dezfouli, Lingawi, and Balleine, 2014). Given the prevalence of dual control systems in learning behaviors, it is worth asking which controller under what conditions should be prioritized. One theoretical work (Daw, Niv, and Dayan, 2005) has proposed an uncertainty-based arbitration framework, emphasizing the importance of the variance of value estimation in deciding the controller usage. Specifically, when a controller has a higher variance in its value estimation (e.g., habitual control), suggesting higher uncertainty in such controller, the other controller (e.g., goal-directed control) should be prioritized. The uncertainty-based arbitration framework has been further tested with "uncertainty" approximated by the "reliability" within the learning algorithms: the unsigned reward prediction error in an MF reinforcement learner and state-prediction error in a model-based Bayesian estimator (Lee, Shimojo, and O'doherty, 2014), which supported this reliability-based theory computationally and neurally well.

Researchers have been using a two-step Markov decision-making paradigm (i.e., two-step task, Daw, Gershman, et al., 2011) to study the usage of MF and MB systems. The paradigm is capable of recovering MF and MB control through the manipulation of transition probability and reward conditions, which are the critical learning components within the MF and MB reinforcement algorithms. The control flexibility encapsulated in the two-step task provides a platform for a more direct behavioral test of the arbitration hypothesis through systematic manipulations of computational variables based upon the previous work (Lee, Shimojo, and O'doherty, 2014).

Within this chapter, we first sought to establish the validity of the roles of reliability signals in the arbitration process between MF and MB controls in the two-step task through behavioral and computational analysis. For this aim, we modified the original two-step task to incorporate the manipulations of three reliability signals: 1) model-free reward prediction error, 2) state-prediction error, and 3) model-based reward-prediction error. Specifically, we would further confirm the role of MF reward prediction error and state prediction error as previously identified reliability signals (Lee, Shimojo, and O'doherty, 2014) on the arbitration process in a unified reinforcement learning framework. Moreover, the role of model-based reward prediction error, as another potential source of reliability signal for the MB system would be tested as well for its effect on arbitration. The results could potentially add to the basis of reliability-based arbitration theory with robust behavioral evidence. Beyond the influence of the change of a single reliability signal, our novel variant of the two-step task enabled the exploratory testing of how different reliability signals interact to influence the control allocation process, which is another key building stone of understanding sophisticated arbitration behaviors in the real world.

As there has been consideration of the extra deliberation time (Keramati, Dezfouli, and Piray, 2011) or cognitive efforts (Pezzulo, Rigoli, and Chersi, 2013) when engaging a goal-directed (or MB) system, there are alternative cost-benefit arbitration theories that leverage the benefit of engaging a goal-directed (or MB) control against its intrinsic costs. As a post-hoc analysis, we also further tested the possible validity of such alternative arbitration theories based on cost-benefit analysis in the two-step task.

2.2 Results

General Behaviors

The task is a novel variant of the original two-step task with a space mining theme, in which, on each trial, participants must choose between two spaceships (colored yellow and blue) in order to reach the landing pads on one of two planets, after which a mining operation will produce different rewarding outcomes. In each trial, participants are instructed to choose the spaceship in order to win as much reward as possible. As shown in Figure 2.1a, after choosing the spaceship, the participant transitions from the spaceship choice state to one of the two planets. The probability of transitioning to a particular planet, given the choice of a particular spaceship, is either 70% or 30%, thereby creating common and rare transitions. Following arrival at a planet, a subsequent transition occurs (without requiring a button press) from the planet to a particular landing pad also probabilistically. The timing and sequence of the trial event are shown in Figure 2.1c. Given the structure of the reward probabilities associated with the two spaceships (Figure 2.1b), there are periods where one spaceship is more rewarding than the other (reversals of which spaceship is most rewarding occur throughout the experiment). To see whether participants performed the task better than chance, we first evaluated the task performance of the M-Turk group (N=452) in terms of each individual's tendency to choose the more rewarding option (Figure 2.1d). We used a random choice agent to generate choices across the task to estimate a distribution for chance level performance (mean: 43.01%, standard deviation: 0.0279), and we compared this distribution to the actual probability of reward from all the participants (mean: 48.16%, standard deviation: 0.0371). Participants' performance was found to be significantly higher than that of a random agent ($T(451) = 24.6982, p < 0.0001$, paired t-test). This suggests that overall the participants learned the reward probability of the two options and adapted accordingly to obtain rewards.

Next, we aimed to determine whether participants exhibited choice behavior consistent with model-based (MB) or model-free (MF) reinforcement learning (RL) overall. In this task, as in the original two-step task (Daw, Gershman, et al., 2011), a model-free (MF) agent would choose the same spaceship in the following trial as chosen in the preceding trial after receiving a rewarded outcome on the preceding trial regardless of whether a common or rare transition occurred in that trial. A model-based (MB) agent, on the other hand, would consider the nature of the transition that occurred prior to reaching the rewarded outcome on the preceding trial when making a spaceship choice in the subsequent trial: after a rare transition, an

MB agent would be less likely to choose the same spaceship as chosen last time, favoring instead the alternative spaceship associated with the common transition more likely to lead to the same landing pad on the next trial. Prior studies have found that, on average, human behavior on this two-step task is a mix of MB and MF strategies (Daw, Gershman, et al., 2011). In order to diagnose if the behaviors are consistent with MF or MB control, as in previous studies, we classified trials by the outcome of the previous trial (i.e., reward vs. no-reward) and the 1st-stage transition type (i.e., common or rare) on the previous trial, and we then examined the probability of choosing the same option in the current trial as in the previous trial (i.e., p(stay)) as a function of outcome and transition type on the previous trial. Qualitatively, as shown in Figure 2, participants expressed a higher probability of choosing the same option in the current trial if the chosen option was rewarded in the previous trial, suggesting an MF component in participants' reward-maximizing strategy. However, participants also showed choice sensitivity to the transition type in the previous trial. They were: 1) more likely to switch the choice after a rewarded trial with a rare transition than with a common transition, and 2) more likely to stay with the choice after a no-reward trial with a rare transition than with a common transition, suggesting an MB component in participants' reward-maximizing strategy. Collectively, these results support the typical observation that participants' behavior reflects a mixture of both MB and MF components on average.

To quantify the degree of reward sensitivity and transition sensitivity, a mixed-effect logistic regression was run to test how the probability of choosing the same option is influenced by the previous outcome, previous transition type, and their interactions (see Methods for details). Consistently, it was found participants were more likely to repeat their choice if that choice was rewarded in the previous trial, reflected by a significant, MF-consistent, previous outcome effect ($\beta = 2.4914, SE = 0.049953, T = 49.875, p < 0.0001$). An MB-consistent effect was also found in that participants were more likely to switch to the other spaceship when they experienced a rare transition towards a reward, reflected by a significant effect of the interaction of the previous outcome and previous transition type ($\beta = -2.9328, SE = 0.10025, T = -29.256, p < 0.0001$). In sum, online participants showed a mixture of MF-consistent and MB-consistent behaviors, replicating the typical behavioral patterns observed in this two-step task (Daw, Gershman, et al., 2011).

Behavioral Evidence for the Effects of Reliability on the Balance between Model-Based and Model-Free Control

In order to test for the role of uncertainty in the predictions of the MB and MF systems in determining the degree to which behavior is MB and MF, the space-miner task featured three different manipulations aimed at perturbing the magnitude of the prediction errors generated within each system. First, we manipulated MF RPEs by altering the magnitude of reward outcomes received across blocks of trials (Figure 2.2b), yielding multiple trial blocks of high and low RPEs. By testing for signatures of MB and MF control in these blocks and comparing them, we could, therefore, assess whether manipulating RPEs, which would vary the reliability of MF predictions, influences the degree to which behavior is MB or MF. In addition, to manipulate uncertainty in the MB system, we changed the predictability of transitions from the planets to the landing pads while keeping the expected value of these transitions constant (Figure 2.2c). Blocks of trials with highly predictable transitions were interleaved with blocks where transitions were maximally uncertain — thereby producing periods with low and high state prediction errors, respectively. Finally, we introduced an additional manipulation of the reliability of the MB system by switching around reward contingencies such that model-based reward prediction errors were either high or low (Figure 2.2a).

According to the reliability-based arbitration framework (Lee, Shimojo, and O'doherty, 2014), MB control should increase when either 1) MF-RPEs are high, 2) SPEs are low, or 3) MB-RPEs are low. To measure the degree to which behavior is MB vs MF, we used a metric called the Transition Sensitivity (TS, see Methods), which quantifies the degree of the subsequent choice of "stay/switch" being sensitive to the 1st-stage transition type experienced in the previous trial. Mathematically, $TS_{rewarded}$ (or $TS_{Unrewarded}$) is the absolute difference of the p(stay) after experiencing reward (or no reward) upon a common transition vs. upon a rare transition. The $TS_{overall}$ are calculated as the summation of $TS_{rewarded}$ and $TS_{Unrewarded}$. Conceptually, an MF agent would be ignorant of the previous transition type and express a zero TS. In contrast, an MB agent would fully consider the previous transition type and express a positive TS. The more MB-consistent behaviors are expressed than the MF-consistent behaviors, the higher $TS_{overall}$ would be.

To assess the hypothesized shift of control due to arbitration, we performed a within-subject comparison of TS expressed in the high-level vs. low-level blocks for each of the manipulated reliability signals separately: 1)MF-RPE, 2) SPE, and

3) MB-RPE, respectively (Figure 2.3). For each analysis of the three reliability signals, we found 1) the TS expressed in the high MF-RPE condition was higher than in the low MF-RPE condition ($Z = 6.5285, p < 0.0001$, one-tailed Wilcoxon signed-rank test); 2) the TS expressed in the low SPE condition was higher than in the high SPE condition($Z = 2.3502, p = 0.0094$, one-tailed Wilcoxon signed-rank test); and 3) the TS expressed in the low MB-RPE condition was higher than in the high MB-RPE condition($Z = 9.4025, p < 0.0001$, one-tailed Wilcoxon signed-rank test). The results of the TS analysis were consistent with the hypothesis that the shift towards MB control was driven by the decreased uncertainty of the MB system (or increased MB reliability) signaled by the computed prediction errors (i.e., MF-RPE, SPE, and MB-RPE).

To quantitatively assess the increased MB control in response to the change in model reliability, a mixed-effect logistic regression was run with the following fixed effect included: the interaction of the previous outcome, previous transition type, and the condition-type of the manipulated reliability signals (e.g., $outcome \times transition \times highSPE$). For this analysis, all three conditional blocks independently interacted with the previous outcome and previous transition type within the same regression (see Methods for details). Since the interactional effect of the previous outcome and previous transition type indicates the degree of an MB-consistent component in the behavioral control (MB control), its additional interaction with the condition type could tell how the MB-consistent component would shift as a function of the level of the reliability signal in consideration. Consistent with the Transition Sensitivity analysis, the logistic regression analysis shows the probability of choosing the same option as in the previous trial is significantly modulated by the previous outcome, previous transition type, and the level of all three reliability signals. As for the M-Turk group (N=452), it was found that the MB control, indicated by the interaction of previous outcome and previous transition type, was increased during conditions where MF-RPEs are high($\beta = -0.1097, F(1, 136686) = 155.89, p < 0.0001$, F-test); also, increased MB control was found during low-SPE conditions ($\beta = -0.018941, F(1, 136686) = 5.4139, p = 0.019979$, F-test); lastly, during conditions where MB-RPEs are low, MB control was found to increase ($\beta = -0.17836, F(1, 136686) = 140.11, p < 0.0001$, F-test). Similarly, as for the replication sample (Prolific + virtual in-lab, N=226), this is consistent with what a reliability-based arbitration theory would predict in that available systems are arbitrated according to the reliability of the systems in consideration. When the level of MF-RPEs was high, suggesting the low reliability of the MF system for reward prediction, MB control would be more engaged than its MF counterpart in the two-step task. Similarly, when the level of SPEs was low, suggesting the high reliability of the MB system given its nature of learning state-state transition, MB control would be favored over MF control. Along a similar line, as the level of MB-RPEs decreased, suggesting a more reliable MB system in terms of reward prediction, MB control would be more relied on for the general behavioral control.

Computational Evidence of Reliability-Based Arbitration

As the evidence of reliability-based arbitration was observed at the behavioral level, we next tested if the observed behavioral arbitration could be accounted for by a reinforcement learning (RL) model that leverages the MB and MF control according to the levels of the manipulated reliability signals (Cockburn et al., n.d., see Methods). The RL model has the modules of MF and MB systems, where both systems learn reward magnitudes, and the values of the two concurrent options are computed within each system. Critically, for both MB and MF systems, the model entails condition-specific weights that are associated with high vs. low levels of reliability signals. Consequently, we have three independent mixture-weight models that entail model weights associated with high and low levels of the reliability signal of interest. For example, for the mixture-weight model that considers the conditional shift between high and low levels of MF-RPE signal (mixture-weight-MF-RPE model, Figure 2.4a), we have the mixed value for a specific option during the conditional block of high and low MF-RPE respectively as follows:

$$Value_{mix}^{highMFRPE} = w_{MF}^{highMFRPE} \times Value_{MF}^{highMFRPE} + w_{MB}^{highMFRPE} \times Value_{MB}^{highMFRPE}$$

$$Value_{mix}^{lowMFRPE} = w_{MF}^{highMFRPE} \times Value_{MF}^{lowMFRPE} + w_{MB}^{highMFRPE} \times Value_{MB}^{highMFRPE}$$

where the weight parameter w is assigned to the MF and MB system separately for both high and low MF-RPE blocks. Hence, there are, in total, four weight parameters fitted in the given model. An option is chosen based on a softmax function that takes as the input the computed mixed values in a given conditional block.

As in the mixture-weight-MF-RPE model, the same model structure was applied to the mixture-weight models that incorporate conditional weights for the reliability signals of SPE (mixture-weight-SPE model) and MB-RPE(mixture-weight-MB-RPE model). It can adapt to changes in the reward environment. We then searched for computational evidence of reliability-based arbitration by examining the fitted conditional MF and MB weights in the models described above. As the model weights associated with the MF and MB systems were independently fitted for high and low levels of reliability signals, we used the difference between the fitted model weights as the metric to characterize the relative model-based control (RMBC) in a given reliability condition. Take the mixture-weight-MF-RPE model as an example. The RMBC for high and low levels of MF-RPE conditions can be calculated as:

$$RMBC^{highMFRPE} = w_{MB}^{highMFRPE} - w_{MF}^{highMFRPE}$$

$$RMBC^{lowMFRPE} = w_{MB}^{lowMFRPE} - w_{MF}^{lowMFRPE}$$

The same RMBC measures were applied to the mixture-weight model that incorporates conditional weight for SPE (mixture-weight-SPE model) and MB-RPE (mixture-weight-MB-RPE model).

If arbitration were driven by the reliability of MB and MF system, in the current two-step task, the relative MB control (RMBC) would increase during periods of high MF-RPEs (i.e., low MF reliability), low SPEs (i.e., high MB reliability), and low MB-RPEs (i.e., high MB reliability).

Indeed, in the M-Turk sample we tested and through the group-level RMBC measure derived from the fitted mixture-weight-MF-RPE model, we found $RMBC_{highMFRPE} > RMBC_{lowMFRPE}$ ($Z = 3.4281, p < 0.001$, one-tailed Wilcoxon signed-rank test, Figure 2.4b). Also, as for the weights fitted through the mixture-weight-SPE model, the MB system was weighted more during low SPE condition than during high SPE condition, reflected by a trending effect of $RMBC_{lowSPE} > RMBC_{highSPE}$ ($Z = 1.4995, p = 0.0669$, one-tailed Wilcoxon signed-rank test, Figure 2.4c). Lastly, from the mixture-weight-MB-RPE model, MB control was found to increase during low MB-RPE condition compared to high MB-RPE condition, indicated by $RMBC_{lowMBRPE} > RMBC_{highMBRPE}$ ($Z = 6.6739, p < 0.0001$, one-tailed Wilcoxon signed-rank test, Figure 2.4d). To summarize, the computational evidence of arbitration according to system reliability was found for all manipulated reliability signals (i.e., MF-RPE, SPE, and MB-RPE, although to a lesser extent for the SPE), solidifying the behavioral evidence of arbitration measured by the Transition Sensitivity.

The Replication of Behavioral and Computational Evidence for the Reliability-based Arbitration

As we collected the same two-step task data as in the M-Turk sample through different platforms (N=226, Prolific and Virtual In-Lab), we repeated the same behavioral and regression analysis as in the previous section in this independent sample to further test the hypothesized effects of prediction reliability on the arbitration between model-based and model-free control.

We first found in this independent sample that participants' performance (mean:48.26%; standard deviation: 0.0398) was also significantly better than that of a random agent (mean: 43.28%; standard deviation: 0.0271, $T(225) = 15.7175, p < 0.0001$, paired t-test). Participants also expressed significant reward sensitivity and transition sensitivity to outcomes and transition types in the previous trial. Specifically, through the same mixed-effect logistic regression of modeling choice as a function of the outcome, transition type, and their interaction, we found the significant previous outcome effect ($\beta = -0.46832, F(1, 68560) = 148.73, p < 0.0001$, F-test) and the significant outcome-transition interaction ($\beta = -0.94594, F(1, 68560) = 346.83, p < 0.0001$, F-test).

Regarding the within-subject analysis of transition sensitivity (TS) measure across high vs. low levels of system reliability, we also found the hypothesized effects supported by the reliability-based arbitration theory in the replication sample (Figure 2.5) — that is, a higher transition sensitivity score, suggesting more MB control, when MF RPE is high ($Z = 3.4186, p < 0.001$, one-tailed Wilcoxon sign rank test), when MB RPE is low ($Z = 2.3983, p = 0.0082$, one-tailed Wilcoxon sign rank test) and a trending effect of higher TS when SPE is low ($Z = 1.5894, p = 0.0560$, one-tailed Wilcoxon sign rank test).

As for the mixed-effect logistic regression analysis on the choice sensitivity to the manipulated reliability signals, we found the hypothesized arbitration effects of three forms of prediction reliability as in the M-Turk sample, indicated by three significant three-way interactions of previous outcome, transition type, and reliability condition. That is, MB control was found to increase when MF-RPEs are high ($\beta = -0.089304, F(1, 68560) = 46.47, p < 0.0001$, F-test) and when MB-RPEs are low ($\beta = -0.059513, F(1, 68560) = 6.9092, p = 0.0085775$, F-test). Also, there is a trend of increased MB control when SPEs are low ($\beta = -0.020792, F(1, 68560) = 2.8531, p = 0.091201$, F-test).

To search for the computational evidence of reliability-based arbitration in the replication sample, we fitted the same set of three condition-based mixture arbitration models and examined, via the relative mode-based control (RMBC) measure, the arbitration effect from the shifts of prediction reliability as in the reliability-based arbitration theory. The hypothesized effect directions from the three reliability signals were found again but with mixed results in terms of statistical significance: relative MB control increases during periods when MF RPEs are high ($Z = 1.0274, p = 0.1521$, one-tailed Wilcoxon signed-rank test), when SPEs are low ($Z = 0.2165, p = 0.4143$, one-tailed Wilcoxon signed-rank test), and when MB RPEs are low ($Z = 2.8881, p = 0.0019$, one-tailed Wilcoxon signed-rank test). The MB RPE effect on the arbitration process remained strong, but the effects of MF RPEs and SPEs are weaker in this replication sample (N=226), which might be due to a smaller sample size compared to the M-Turk sample (N=452).

Exploratory Analysis of Inter-Reliability Arbitration

So far, we have shown how multiple forms of reliability signals can independently guide the arbitration between MF and MB control. As all three reliability signals are manipulated to dynamically shift between high and low levels orthogonally in the task (e.g., there are periods of high and low SPEs within the high MB-RPE condition), this enables us to explore how multiple forms of reliability signals could exert interactional influence on the arbitration process besides their independent roles. Specifically, given the fact that the manipulation of MB-RPEs was through different reward-contingency sessions (i.e., state-contingent session vs. stimulus-contingent session) and the manipulations of MF-RPE and SPE signals were frequently shifted within each session, we could probe how the reliability signal of MF-RPEs could modulate the arbitration process as a function of different levels of MB-RPE signal, and similarly as for the reliability signal of SPEs.

As the degree of shift in the relative model-based control measure (RMBC) across high and low levels of reliability indicates the strength of the arbitration effect, we specifically characterized the change of the arbitration strength via the RMBC measure across high and low MF-RPE conditions within the state-contingent session (i.e., low MB-RPE) and the same was conducted within the stimulus-contingent session (i.e., high MB-RPE):

$$diffRMBC^{MFRPE}_{state-contingent} = RMBC^{highMFRPE}_{state-contingent} - RMBC^{lowMFRPE}_{state-contingent}$$

$$diffRMBC^{MFRPE}_{stimulus-contingent} = RMBC^{highMFRPE}_{stimulus-contingent} - RMBC^{lowMFRPE}_{stimulus-contingent}$$

Similarly, the shift of RMBC across high and low SPE conditions was examined within the state-contingent session and stimulus-contingent session, respectively. With this exploratory analysis, we found a stronger arbitration effect induced by the shifted levels of MF-RPEs when MB-RPEs are low compared to when MB-RPEs are high ($Z = 2.9055, p = 0.0018$, one-tailed Wilcoxon signed-rank test, Figure 2.6(left)):

$$diffRMBC^{MFRPE}_{state-contingent} > diffRMBC^{MFRPE}_{stimulus-contingent}$$

However, for the arbitration effect driven by the change of SPE levels, no significant strength difference was found across the state-contingent and stimulus-contingent session ($Z = 0.1682$, $p = 0.4332$, one-tailed Wilcoxon signed-rank test, Figure 2.6(right)).

This analysis on inter-reliability arbitration was also conducted on the replication sample (N=226). A quantitatively similar interaction effect between MF RPEs and MB RPEs was detected on directing the relative MB control: the arbitration effect of MF RPE is more evidently expressed during state-contingent reward session (i.e., low MB RPEs) than during stimulus-contingent reward session (i.e., high MB RPEs), reflected by the trending change in RMBC measure across conditions ($Z = 1.3099, p = 0.0951$, one-tailed Wilcoxon sign-rank test). Also, similarly, as in the M-Turk sample, the arbitration effect of SPE is not significant either in this replication sample, as no significant change of RMBC change in high vs. low SPE conditions was found across state-contingent vs. stimulus-contingent reward session ($Z = -0.3719$, $p = 0.6450$, one-tailed Wilcoxon signed-rank test).

In sum, this exploratory analysis shows that the main effect of MF reward prediction reliability (indicated by MF-RPEs) on the arbitration process was more evident during the state-contingent session where the level of MB-RPEs is low, and the behavior is overall more under MB control. This might suggest a potential asymmetric arbitration mechanism that the arbitrator evaluates more of the MF-reliability to consider shifting to MF control when the behavior is overall under the MB control (i.e., due to high MB prediction reliability during state-contingent sessions), but not much of such evaluation when the behavior is overall under MF control. This is consistent with the proposal that MF control (or the habitual system) might serve as a default system (Lee, Shimojo, and O'doherty, 2014; Keramati, Dezfouli, and Piray, 2011; Pezzulo, Rigoli, and Chersi, 2013), and the arbitrator constantly evaluates the reliability of the MF system to consider the default even when the MB system is more engaged for the moment (Lee, Shimojo, and O'doherty, 2014).

Testing Alternative Theories of Arbitration between Controls

Through the behavioral and computational analysis so far, we found evidence of how the arbitration between MF and MB control is governed by the corresponding system uncertainty varied through multiple forms of prediction error signals, which supports an uncertainty-based arbitration theory (Daw, Niv, and Dayan, 2005; Lee, Shimojo, and O'doherty, 2014). On the other hand, to explain the shift of habitual and goal-directed control in the psychology literature, alternative arbitration theories have been proposed, focusing on the cost-benefit analysis of engaging the goal-directed control (Keramati, Dezfouli, and Piray, 2011; Pezzulo, Rigoli, and Chersi, 2013). Here, we attempted to test predictions of these alternative arbitration theories by assuming the correspondence between habitual control and MF control and the correspondence between goal-directed control and MB control, respectively, although further empirical evidence is needed to support the assumptions.

The alternative arbitration theories proposed by Keramati et al. (2011) and Pezzulo et al. (2013) entail a cost-benefit analysis of the two controllers and rely on an assumption about the cost of engaging a slow and deliberative goal-directed (or MB) system, which is either through 1) the forgone reward that could have been collected during the relatively long deliberation time (Keramati, Dezfouli, and Piray, 2011), or 2) the cognitive efforts and time spent during mental simulations (Pezzulo, Rigoli, and Chersi, 2013). Consequently, the arbitration scheme in both alternatives uses the "cached" value habitual (or MF) control as default and leverages the benefit of engaging the goal-directed (or MB) control against its associated cost to consider its usage.

Specifically, the speed/accuracy trade-off theory proposed by Keramati et al. (2011) assumes the goal-directed (or MB) control could use the knowledge of the environmental structure to compute the "value of perfect information" (VPI) of the concurrent actions and the arbitrator compares it against the opportunity cost (i.e., due to the longer deliberation time) for control selection. Conceptually, the VPI of a given action is the policy improvement if the true value of the considered action is known, and the VPI measure is proportional to the overlapping area between the value distribution of the two concurrent actions (Figure 2.7). In the paradigm of testing sensitivity to outcome-devaluation after moderate vs. extensive training, the VPI of the actions is high after moderate training since there is a higher probability that the true value of action 1 is higher than that of action 2 (Figure 2.7(left)), hence devaluation-sensitive behaviors are expressed (goal-directed control is beneficial). On the contrary, as the VPI of the concurrent actions becomes small as training becomes extensive (Figure 2.7(right)), devaluation-insensitive behaviors are expressed (not beneficial to activate the goal-directed control). As the moderate training schedule, compared to the extensive training schedule, induces a relatively smaller value difference between the two concurrent actions, one prediction then would be the smaller the value difference between the two concurrent actions, the higher the VPI of the given actions, and thus it is more beneficial for the arbitrator to activate the goal-directed (or MB) control as the benefit would be larger than the opportunity cost caused by the loss of time.

In parallel, Pezzulo et al.(2013) proposed the arbitrator evaluates a given action's "value of information" (VoI) against the cost of mental simulation on the decision of activating the goal-directed (or MB) control for value estimation of the given action. The VoI measure is defined as, in the case of two concurrent actions ($Act1$ and $Act2$), the ratio of value uncertainty over the difference between the considered action $Act1$ and the alternative action $Act2$:

$$VoI_{Act1} = \frac{C_{Act1}}{|Q_{Act1} - Q_{Act2}| + \epsilon}. \tag{2.1}$$

Here C_{Act1} denotes the uncertainty of the value estimates of $Act1$. Thus, if the value uncertainty is controlled, then again, the smaller the value difference between the concurrent actions, the more likely the VoI for the concurrent actions would be larger than the mental simulation cost, and thus the goal-directed control (or MB) control would have a higher chance to be activated.

Given the task structure that the reward probability associated with the two stimuli would have periods of small and large differences, the concurrent action values would hence have periods of small and large differences, respectively (Figure 2.1b). It then provides an opportunity to test whether the goal-directed or MB control would be more manifested when the value difference between the concurrent actions is small, according to the alternative arbitration theories proposed by Keramati et al. (2011) and Pezzulo et al. (2013). Specifically, first in the main M-Turk sample (N=452), we compared the degree of MB control via the Transition Sensitivity measure during periods of large value difference against that during periods when value difference is small. Yet no significant difference of TS across periods of large and small value difference was found (Figure 2.8 (left), $Z = 0.1942, p = 0.4230$, one-tailed Wilcoxon signed-rank test). We also used a mixed-effect logistic regression to examine how the effect of previous outcome and transition type on stay choices was modulated by a binary regressor on small vs. large value difference. Interestingly, we found a significant interaction between the previous outcome and value difference type on the stay choices ($\beta = -0.43204, F(1, 136694) = 29.981, p < 0.0001$, F-test), suggesting when the value difference of the concurrent actions is small, the behavior was less driven by the previous outcome (for which one possibility is the behavior being less under MF control); however, we did not find the significant three-way interaction of the previous outcome, previous transition type, and value difference type, which potentially corresponds to the increased MB control during periods when the value difference is small ($\beta = 0.00078283, F(1, 235160) = 0.0099999, p = 0.92034$, F-test). By fitting a condition-based arbitration model with separate weights for periods of large and small differences in reward probability (same structure as in Figure 2.4a), we compared the metric of relative model-based control (RMBC) across periods of small and large value differences and no significant difference of RMBC measure was found (Figure 2.9 (left), $Z = -0.7948$, $p = 0.7866$, one-tailed Wilcoxon signed-rank test). If any patterns exist, the RMBC is larger when the value difference is large, which is opposite to the prediction of the alternative arbitration theories.

We also examined the same measures of Transition Sensitivity, mixed-effect logistic regression estimates, and computational system weights in the replication sample (N=226). Similar to the M-Turk sample, we found no significantly different Transition Sensitivity measures across large and small value difference periods (Figure 2.8 (right), $Z = -0.1636, p = 0.5650$, one-tailed Wilcoxon signed-rank test). Also, for the mixed-effect logistic regression, a significant interaction effect of the previous outcome and value-difference type was found ($\beta = -0.048092, F(1, 68568) = 17.799, p < 0.0001$, F-test), whereas the interaction of the previous outcome, previous transition type, and value-difference type was again found to be non-significant ($\beta = -0.0019441, F(1, 68568) = 0.029475, p = 0.86369$, F-test). Through fitting the same condition-based arbitration model on the replication sample, we found a non-significant trend of larger RMBC during periods of small value difference (Figure 2.9 (right), $Z = 1.4156, p = 0.0784$, one-tailed Wilcoxon signed-rank test).

As the cost-benefit arbitration theories assume that goal-directed (or MB) control is more likely to be beneficial when the value difference between the concurrent actions is large, one concern regarding the current approach is whether the periods of small vs. large reward probability indeed correspond to periods of small vs. large action value difference. Through post-hoc examination of action values derived from the condition-based arbitration model, it was confirmed that action values have a larger difference in periods with large differences in reward probability than in periods with small differences in reward probability ($Z = 13.1624, p < 0.001$, Wilcoxon signed-rank test). Thus, from a reinforcement learning perspective, it is valid to examine whether the degree of MB control shifts as a function of value difference in concurrent actions, which are approximated by the reward probability shifts throughout the task.

In summary, the mixed results could not ensure the applicability of these alternative arbitration theories to the current data of arbitration between MB and MF control in this two-step task, suggesting more targeted empirical tests are needed for further assessment.

2.3 Discussion

Using a novel variant of the two-step task where reward magnitude, state-transition uncertainty, and reward contingency are manipulated systematically (Cockburn et al., n.d.), three forms of reliability signals are shifted to high and low levels for a test of how reliability-based arbitration framework could explain the allocation between MF and MB control. Not only the role of MF reward prediction error and state prediction error to exert influence on arbitration is solidified as found by Lee and colleagues (2014), accumulating more evidence for a reinforcement learning framework of the uncertainty-based arbitration theory, but also a new learning signal from the MB system, MB reward prediction error, was found to guide the arbitrator to allocate behavioral control in a significant way. Through statistical tests on transition sensitivity and regression analysis of choice sensitivity to high vs. low levels of various reliability signals, we established that MF RPE, SPE, and MB RPE effectively shift the behavioral control as the reliability-based arbitration theory predicted. Further consistent evidence was also found by fitting a reinforcement learning arbitration model incorporating the three reliability signals in controller weight determination. These findings were further supported in an independent replication sample. Statistically, MF RPE and MB RPE were found to exert a relatively strong influence on control shifting, whereas SPE was found to show the impact to a lesser extent.

We speculated the weak arbitrating effect of SPE could be due to a feature built into the task — the manipulation of state transition, designed to manipulate the SPE level. As in every trial, only one action is made at the first stage, and at the second stage, where the state transition is manipulated, there is no effective state-transition knowledge that would inform the optimality of the first-stage choice as either of the two states following the first transition would emit the same reward status. From the participant's perspective, the second-stage state transitions could be treated as the intermediate connecting state to the significant outcome stage, which is necessary to inform an optimal first-stage choice. Given its lesser importance in driving participants towards good performance, it is plausible that the state-transition manipulation did not elicit an arbitration effect as strong as magnitude and contingency manipulations on rewards. Still, it is intriguing that, replicating the previous work (Lee, Shimojo, and O'doherty, 2014), although not a huge effect, the manipulation itself did influence the arbitration process even when it is performance-irrelevant.

The specific design of the task also enabled us to conduct an exploratory analysis of how various forms of reliability signals interact to influence the allocation of control. Interestingly, through computational modeling, we found that the increase of MF RPE, indicating low reliability of the MF system, more strongly shifted participants' behavior to be under MB control when the reward delivery is contingent upon the terminal states than when the reward delivery is contingent upon first-stage stimulus. In other words, the arbitration effect of MF RPE is stronger when MB RPE is low. Hence, it suggests that the MF reliability signal (MF RPE) exerts asymmetric influence across high vs. low MB RPE levels. Given this asymmetric effect, it suggests the arbitrator might not consider the two types of reward prediction errors in a linear and equally weighted way. Also, this post-hoc observation might first seem contradictory in that the MF RPE signal guides behavior when the behavior is more MB-consistent. Yet it could be explained by a hypothesis that the MF system is always functioning at the back-end and the brain tracks the dynamics of MF RPE even when the behavior is more under MB control. Previously, Lee and colleagues (2014) had found that BOLD activity in neural arbitrator regions shows negative coupling with neural regions encoding MF values but no positive coupling with regions encoding MB values when behaviors are more MB-consistent. Also, the arbitrator activity modulates the effective connectivity between the MF value region and the integrated value region (vmPFC) based on the degree of demand in MF control. Such findings together suggested a possibility that the MF system is the default operating system, with the arbitrator modulating the MF system and the system's integration process when the reliability signal indicates that the MB system should be favored. So, our behavioral finding that the interactional arbitration effect between MF RPE and MB RPE is consistent with this neural arbitration view of the MF system being the default and functioning regardless of the expressed behavior. With respect to the state-transition manipulation, however, the SPE signal did not exert any arbitration effect across high vs. low MB RPE levels. Thus, it might imply that the interactional arbitration effect driven by different reliability signals might be subject to the type of the reliability signal. Since it is only a post-hoc analysis, more formal and stringent experimental testing and modeling works are needed to investigate what various reliability signals are integrated and in what manner to guide the arbitration process.

With a post-hoc test of alternative arbitration theories that leverage the cost of the goal-directed (or MB) control, the empirical evidence pointing towards these theories turned out to be mixed. There is an important caveat in this post-hoc attempt. The original arbitration theories from Keramati et al. (2011) and Pezzulo et al. (2013) were proposed in the context of explaining the controller switch from goal-directed control to habitual control as the training goes. The evaluation of the control engaged was typically through the outcome-devaluation procedure. Here in the two-step task, the control in use was framed as MF and MB, which was identified either via 1) choice patterns as a function of the states of the previous trial (i.e., outcome and transition type) or 2) extracting the parameter of controller weight from a reinforcement learning model. Hence, the approach makes the assumption that the control identified via the aforementioned approaches is intrinsically comparable to the control diagnosis procedure in the outcome-devaluation paradigm. Although the habitual control identified as outcome-devaluation insensitive shares with MF control the key feature of inflexible action execution upon accumulated reward history, there might be an intrinsic difference between the habit and the MF control in terms of the underlying cognitive mechanism in that non-RL strategies could take the camouflage of MF control in the two-step task (Cockburn et al., n.d.). More generally, it is another question whether habit/goal-directed control corresponds perfectly to the MF/MB control as the two pairs are identified from different tasks or definitive approaches and could potentially have distinct neural substrates for habitual/goal-directed control (Balleine and O'doherty, 2010) and MF/MB control (Beierholm et al., 2011; Wunderlich, Dayan, and Dolan, 2012), respectively. Additionally, there has been a proposal that the reward advantage carried by the MB system over the MF system is to be incorporated into the arbitration process (Kool, Gershman, and Cushman, 2017). As the MF and MB strategies have equal performance in our task design, we could not exclude the possibility that the strategy's performance could play a role in arbitration.

Overall, the work in this chapter further adds to the empirical evidence of the reliability-based arbitration theory through a two-step task with rich design features, which facilitates behaviorally distinguishing MF vs. MB control with three manipulations of reliability signal. The behavioral arbitration effect driven by the MB RPEs enriches the scope of the MB system's knowledge, both state-transition knowledge and the reward information, that contribute to the dual system arbitration process. With exploratory analysis, there were tentative results suggesting the integration of multiple reliability signals via the arbitrator might not be linear. Moreover, evidence for alternative arbitration theories leveraging the cost-benefit analysis of the goal-directed system was not found in a post-hoc analysis. The speculations from these exploratory analyses need to be formally examined in future works.

2.4 Methods

Participants

For recruitment on online platforms (i.e., Amazon Mechanical Turk and Prolific) and the virtual in-lab group, we recruited online participants who currently live in the United States and who are also fluent English speakers and readers. The age range for the studies is from 18 to 65 years. Specifically on M-Turk, besides the recruitment criteria specified above, workers are only recruited if 95% of their past M-Turk jobs have been accepted by the M-Turk requesters. Before the experiment, all participants signed the online informed consent approved by the California Institute of Technology's Institutional Review Board under protocols 19-0914 and 19-0916. All participants were paid in monetary form (either through an M-Turk account, Prolific account, or a peer-to-peer payment app). For the experiments conducted on M-Turk, 1028 workers completed the first session of the space miner task, and 512 of 1028 completed the second session of the task, thus making up the total number of samples for Experiment 1. As for recruitment criteria for the virtual in-lab group in Experiment 2, additional recruitment criteria were imposed so that the control participants would not have any history of anxiety disorder (Obsessive-Compulsive Disorder, Body Dysmorphia Disorder, generalized anxiety, social anxiety/social phobia) and/or depressive disorders (dysthymia, major depression).

Experimental Procedure

All participants from M-Turk, Prolific, and virtual in-lab group completed the same version of the two-step task. Participants from online platforms (i.e., M-Turk and Prolific) were included based on instruction comprehension and task completeness (see Exclusion Criteria). No statistical methods were used to predetermine the sample sizes. All participants were given a consent form online before the entire experiment and could only proceed to complete the task if they consented. For participants in the virtual in-lab group, one-on-one communications were made via email or phone calls to ensure the quality of task completion with the goal of excluding potential issues associated with carelessness that commonly occur in online studies. Specifically, to maximize the similarity of an online experiment to the in-lab experience, at an arranged time, the experimenter would contact the scheduled participant via phone or text message to review the online consent form and talk through the task instructions. Upon completion of the task, the experimenter and participant would open a text chat, where the participant would report the completion of each task, and the experimenter could answer questions.

Firstly, participants read through the instructions of the spaceship task. After the instructions, a few questions were asked to check whether participants understood the task and remembered the key features of the task. If the participant answered the questions wrongly, then they were sent back to the beginning stage of the instructions to redo the instructions until they answered the questions correctly. After the instructions, 12 trials of practice were done before the main task.

For the online sample from M-Turk (N=452) and Prolific (N= 160), the main task consists of two sessions, with 154 trials for each session. Specifically for the M-Turk group, 1,028 participants completed the first session of the spaceship task, in which the reward delivery was contingent upon the chosen spaceship regardless of the planets or landing pads they traveled to (i.e., stimulus-contingent). All 1,028 participants were invited back and offered to complete the second session of the task, where the reward delivery was contingent upon the terminal landing pads that were arrived at (i.e., state-contingent). 502 participants on M-Turk were included in the analysis, given they completed both sessions of the task. For the virtual in-lab group (N=66), each participant was in one-on-one online communication beforehand about the general structure of the experiment, in which they performed two sub-sessions (with a couple of weeks in between) for each of the stimulus-contingent and state-contingent versions of the task, respectively. In all the behavioral and computational analyses, we treated two separate sub-sessions as a combined session for the two reward-contingency conditions so that the data structure in the virtual in-lab group in all the analyses is effectively the same as in the M-Turk and Prolific group.

Space Miner Task

For the main task, participants were instructed to collect rewards through space mining on different planets. The background is that mining has begun on two planets (i.e., one identified as red and the other green) in space, and the goal is to earn as many points as possible by mining gems from the two planets. However, the mines on the two planets have changing conditions in their production. Sometimes gems could be found, but other times, worthless rocks could also be mined out. Specifically, there are two landing pads for the corresponding mines on each planet, one to the North and the other to the South. The landing pads are identified through their unique scenic view and are located in the upper (North) and bottom (South) parts of the planet on the screen.

Participants can choose between two different spaceships (identified as yellow and blue, with the screen locations of the spaceships fixed across trials) using a button-press to travel to the two planets for a miner. They are instructed that the yellow spaceship usually lands on the red planet, and the blue spaceship usually lands on the green planet. However, space travel can sometimes be a bit unpredictable due to space debris, so in some rare situations, the yellow ship will be forced to land on the green planet, and the blue spaceship will be forced to land on the red planet. Participants were instructed to use two buttons to choose the corresponding spaceships. In a given trial, once participants chose one spaceship, they observed the spaceship being highlighted and taken off and the planet appearing after the spaceship landed. Afterwards, the landing pad for the mine would also appear, either on the upper or the bottom part of the planet, given if they landed at the North or the South mine on the planet. Once the spaceship landed, the mine production appeared as either a gem with its price or a worthless rock. Specifically, participants were instructed that the gem's price is unpredictable and will change daily. Also, participants had no control over which mine (North or South) the spaceship would eventually land on. In general, the conditions at all four mines (two mines on each of the two planets) would change throughout the game regarding gem vs. stone production. Participants were encouraged to learn which mine produces gems the most reliably.

After going through the instructions of the spaceship task, participants were asked a few questions:
1. How many planets are there?

2. How many mines are there on each planet?
3. Which planet does the yellow ship usually land on?
4. Which planet does the blue ship usually land on?
5. How many points is a mined rock worth? (0 points vs. 1-100 points)
6. How many points is a gem worth? (0 points vs. 1-100 points)

Participants proceeded to the practice trials and then the main experiment if they answered all the questions correctly.

Task Design

The current variant of the two-step task shared a similar task structure as the original two-step task (Daw, Gershman, et al., 2011) overall but with some differences in details. The general structure consists of the first-step transition and reward probability shifts, and also three condition manipulations (i.e., reward magnitude, state transition, and reward contingency) were built into the space miner task. One key difference is that in the current task, there was only one action to be made, which was at the initial state, and the rest of the trial were all state transitions with no further actions needed. All three cohorts of participants experienced the same general structure of the space miner task. On each trial, the yellow and blue spaceships would appear on the left and right sides of the screen. If chosen, the transition via the yellow spaceship towards the red planet occurred with a probability of 0.7 and with a probability of 0.3 towards the green planet; corresponding probabilities were flipped for the transition via the blue spaceship towards the planets. After the transition from the planet stage to the landing-pad stage, reward or non-reward outcomes would appear. The underlying reward probability was shared across the two landing pads on a given planet (reward probability would also be associated with the spaceship chosen, depending on the reward contingency condition described later in this section), and there were specific periods built-in such that landing on one planet was more rewarding than landing on the other planet, and also periods where landing on either of the two planets was relatively comparable in terms of the reward probability. To implement this, the reward probability associated with two planets started from 1 vs. 0, and then by using the Sigmoid function, the reward probability of the rewarding planet decayed from 1 towards 0.3 (asymptote) within the time span of from 20 to 25 trials (the exact number of trials depended on whether a rare spaceship-planet transition is made); for the reward probability associated with the currently non-rewarding planet, the same Sigmoid decay rate and the flipped sign of decay slope was used to have the reward probability drift from 0 to 0.5 (asymptote) again within the time span of from 20 to 25 trials. Afterwards, the reward probabilities associated with the two planets were reset to 1 vs. 0, but the rewarding planet was reversed compared to the previous block of trials. This drifting of reward probabilities and the reversal of rewarding planet/landing pads occurred throughout the entire session of the experiments. The order of which planet was firstly rewarding was counterbalanced across the participants. The structure of going from strong preference (large gap of reward probabilities: 1 vs. 0) towards almost indifference (reward probabilities: 0.3 vs. 0.5) was first to facilitate learning of the rewarding option and then setting the value of two options towards indifference to prepare for learning after the preference reversal. Also, a critical trial of rare transition was built into the very beginning after a preference reversal to facilitate detecting stay/switch behaviors as a signature of the MB vs. MF strategy. It is worth noting that another manipulation of reward contingency (described later in this section) would change the contingency of reward delivery upon the landing pads (or planets) vs. the spaceships, yet the general fluctuating and reversal dynamics of the reward probability shift would remain the same across the two reward contingency conditions.

Reward Magnitude Manipulation

On top of the reward probabilities shifts throughout the main task, the reward magnitudes (points associated with the gem), if there was a reward, were manipulated to shift between two conditions: low reward prediction error (low RPE) vs. high reward prediction error (high RPE). For the reward magnitude manipulation, the magnitudes were drawn from a uniform distribution of (0.1, 0.19) for the low RPE condition and (0.3, 1) for the high RPE condition, for which the actual points were scaled by 100. Low vs. high RPE conditions were shifted every 26-27 trials, and the order of the low vs. high RPE conditions was counterbalanced across participants. The reward magnitude manipulation was the same and shared across all three sources of the online sample.

State-Transition Uncertainty Manipulation

For the second-step transition (with no actions required) after landing on a given planet, the landing pad would be shown subsequently to illustrate whether the North or South mine was landed on. A state-transition manipulation was built-in at this phase of the trial to shift the uncertainty of landing on a specific mine (North or South) given a specific planet: high state prediction error (high SPE) vs. low state prediction error (low SPE). During high SPE conditions, the transition to one of the two landing pads on a given planet was at the equal probability of 0.5 vs. 0.5, which elicited large state-transition uncertainties. On the other hand, during low SPE conditions, the transition to one of the two landing pads on a given planet was at the biased probability of 0.9 vs. 0.1, and whether the transition to the North or South landing pad was biased would change along the course of the task. Low vs. high SPE conditions were shifted every 26-27 trials, and the order was counterbalanced across participants. As participants would mostly see one type of transition within a conditional block, the state prediction was relatively certain under low SPE conditions. The state-transition manipulation was the same and shared across all three sources of the online sample.

Reward Contingency Manipulation

Besides manipulating reward magnitude and state-transition uncertainty, reward delivery was also manipulated to be contingent upon 1) what stimulus (i.e., spaceship) participants choose or 2) what terminal states (i.e., landing pads) were arrived at, meaning rewards are 1) stimulus-contingent or 2) state-contingent, respectively. The stimulus-contingent and state-contingent conditions were run by two separate sessions of the experiment, and the two sessions had the same number of trials. During the stimulus-contingent reward condition, the reward probability structure (described earlier in this section) was associated with the stimulus (i.e., spaceship) chosen by the participant in each trial, regardless of what second-stage state (i.e., planet) or terminal state (i.e., landing pad) was reached. During the state-contingent reward condition, the reward probability structure (described earlier in this section) was associated with the pair of landing pads on a given planet, so both the choice at the first stage and the actual terminal states that were arrived at would potentially influence the final outcome. Intuitively, under the state-contingent reward condition, in response to a win or a loss after a rare spaceship-planet transition in the previous trial, the first-stage transition probability representation (as used in a model-based strategy) would direct the subsequent choice towards the stimulus more likely leading to the rewarding terminal states; in contrast, during the stimulus-contingent reward condition in the similar situation, such a model-based strategy would end up choosing the option that was less likely to deliver a reward, effectively eliciting more model-based RPEs. For the M-Turk and Prolific sample (N=612), the stimulus-contingent and state-contingent reward conditions corresponded to each of the two sessions (154 trials per session) that every participant completed, and each contingency condition consisted of 154 trials; for the virtual in-lab part of the sample (N=66), each participant completed two sub-sessions within state-contingent and stimulus-contingent reward condition respectively (with 154 trials per sub-session) and each contingency condition is made up of 308 trials.

Within each reward-contingency session (i.e., the state-contingent and stimulus-contingent sessions), both reward magnitude and state-transition uncertainty were manipulated simultaneously along with the reversals of reward probability (either associated with the state or the stimulus). We ensured all the manipulated computational variables were orthogonal to each other, so the averaged overall effect of one conditional variable was solely independent rather than confounded by other manipulated variables. To realize the orthogonality between conditions, the starting trial index of each condition was offset to one another (i.e., with a lag of around 13 trials) so that no two conditional blocks (e.g., High RPE and Low SPE) would completely overlap, with the aim of obtaining the sole conditional effect by averaging across one type of conditional blocks and canceling out the correlations between manipulations.

Statistical Analysis

Transition Sensitivity Measure

We quantify the overall effect of transition on behavior according to a Transition Sensitivity (TS) measure, defined as the change of the probability of repeating the previous choice after experiencing a common vs. rare transition in the previous trial, summed over cases of experiencing reward and those of experiencing non-reward in the previous trial:

$$\begin{aligned} TS = {} & [p(Stay|Rewarded, Common) - p(Stay|Rewarded, Rare)] + \\ & [p(Stay|Unrewarded, Common) - p(Stay|Unrewarded, Rare)] \ . \end{aligned}$$

Logistic Regressions

For all the mixed-effect logistic regression analyses, "fitglme" function in Matlab was used. To quantify how the previous reward, previous transition type, and their interactions affected the stay choices, we used a mixed-effect logistic regression to model the probability of staying with the option chosen in the previous trial ($isStay$) as a function of the previous outcome ($isWin$, 1: reward, 0: no-reward), the previous transition type ($isRare$, 1:rare, 0:common) and their interactions. The fixed effects include the previous outcome, the previous transition type, and their interactions, and the slope of the previous outcome, the previous transition type, and their interaction were modeled as a random effect that could vary at the single-subject level (indicated by $subID$). The regression model is:

$$isStay \sim isWin \times isRare + (1 + isWin \times isRare|subID),$$

where "×" denotes the main effects and the interaction between each independent variable, and "1" denotes the intercept, which is the average stay probability for each subject. Trials, where the choice was not made within 2 seconds were excluded, before estimating the regression model.

To examine the behavioral evidence of arbitrating between MB-consistent vs. MF-consistent behaviors, three binary variables indicating the three conditional manipulations (i.e., reward magnitude, state-transition, and contingency) are added to the first regression model as three independent interactions with $Reward_{t-1} \times isRare_{t-1}$. Specifically, $HighRPE^{mf}_{t-1}$ is a binary variable indicating the previous trial's reward magnitude condition (1: high reward magnitude/high MF-RPE, 0: low reward magnitude/low MF-RPE). Similarly, $LowSPE^{mb}_{t-1}$ indicates the previous trial's state-transition uncertainty condition (1: low state-transition uncertainty/low SPE, 0: high state-transition uncertainty/high SPE), and $LowRPE^{mb}_{t-1}$ indicates the previous trial's reward contingency condition (1: state-contingent reward/low MB-RPE, 0: stimulus-contingent reward/ high MB-RPE). The fixed effects are the three full interactions of the previous outcome, the previous transition type, and the conditional variable, and the random effects are the same as in the first regression model, varying at the single-subject level (for random effects, we did not model the same full interaction as in the fixed effects due to the convergence issue). The full model is:

$$isStay \sim isWin \times isRare \times \left(HighRPE^{mf}_{t-1} + LowSPE^{mb}_{t-1} + LowRPE^{mb}_{t-1} \right) + \\ (1 + isWin \times isRare|subID),$$

where "×" denotes the main effects as well as the interaction between the independent variables, and "1" denotes the intercept, that is, the average stay probability for each subject. Trials, where the choice was not made within 2 seconds, were excluded before estimating the regression model.

Exclusion Criteria

We defined a set of exclusion criteria to identify careless and/or malicious behavior in the task. This includes:

- Not responding in over 10% of the total trials.
- Selecting the same option excessively, quantified as response variance less than 3 standard deviations below the sample mean.
- Switching between options excessively, quantified as a response auto-correlation less than 3 standard deviations below the sample mean.
- Poor task comprehension, quantified by repeating the task instructions more than 5 times.

References

Balleine, Bernard W and Anthony Dickinson (1998). “Goal-directed instrumental action: contingency and incentive learning and their cortical substrates”. In: *Neuropharmacology* 37.4-5, pp. 407–419.

Balleine, Bernard W and John P O’doherty (2010). “Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action”. In: *Neuropsychopharmacology* 35.1, pp. 48–69.

Beierholm, Ulrik R et al. (2011). “Separate encoding of model-based and model-free valuations in the human brain”. In: *Neuroimage* 58.3, pp. 955–962.

Cockburn, Jeffrey et al. (n.d.). “Characterizing heterogeneity in human reinforcement learning and the arbitration of behavioral control”. In: ().

Daw, Nathaniel D, Samuel J Gershman, et al. (2011). “Model-based influences on humans’ choices and striatal prediction errors”. In: *Neuron* 69.6, pp. 1204–1215.

Daw, Nathaniel D, Yael Niv, and Peter Dayan (2005). “Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control”. In: *Nature neuroscience* 8.12, pp. 1704–1711.

Dezfouli, Amir and Bernard W Balleine (2013). “Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized”. In: *PLoS computational biology* 9.12, e1003364.

Dezfouli, Amir, Nura W Lingawi, and Bernard W Balleine (2014). “Habits as action sequences: hierarchical action control and changes in outcome value”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1655, p. 20130482.

Gläscher, Jan et al. (2010). “States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning”. In: *Neuron* 66.4, pp. 585–595.

Keramati, Mehdi, Amir Dezfouli, and Payam Piray (2011). “Speed/accuracy trade-off between the habitual and the goal-directed processes”. In: *PLoS computational biology* 7.5, e1002055.

Kool, Wouter, Fiery A Cushman, and Samuel J Gershman (2016). “When does model-based control pay off?” In: *PLoS computational biology* 12.8, e1005090.

Kool, Wouter, Samuel J Gershman, and Fiery A Cushman (2017). “Cost-benefit arbitration between multiple reinforcement-learning systems”. In: *Psychological science* 28.9, pp. 1321–1333.

Lee, Sang Wan, Shinsuke Shimojo, and John P O’doherty (2014). “Neural computations underlying arbitration between model-based and model-free learning”. In: *Neuron* 81.3, pp. 687–699.

Otto, A Ross et al. (2013). “Working-memory capacity protects model-based learning from stress”. In: *Proceedings of the National Academy of Sciences* 110.52, pp. 20941–20946.

Pezzulo, Giovanni, Francesco Rigoli, and Fabian Chersi (2013). “The mixed instrumental controller: using value of information to combine habitual choice and mental simulation”. In: *Frontiers in psychology* 4, p. 92.

Smittenaar, Peter et al. (2013). “Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans”. In: *Neuron* 80.4, pp. 914–919.

Wunderlich, Klaus, Peter Dayan, and Raymond J Dolan (2012). “Mapping value based planning and extensively trained choice in the human brain”. In: *Nature neuroscience* 15.5, pp. 786–791.

Wunderlich, Klaus, Peter Smittenaar, and Raymond J Dolan (2012). “Dopamine enhances model-based over model-free choice behavior”. In: *Neuron* 75.3, pp. 418–424.

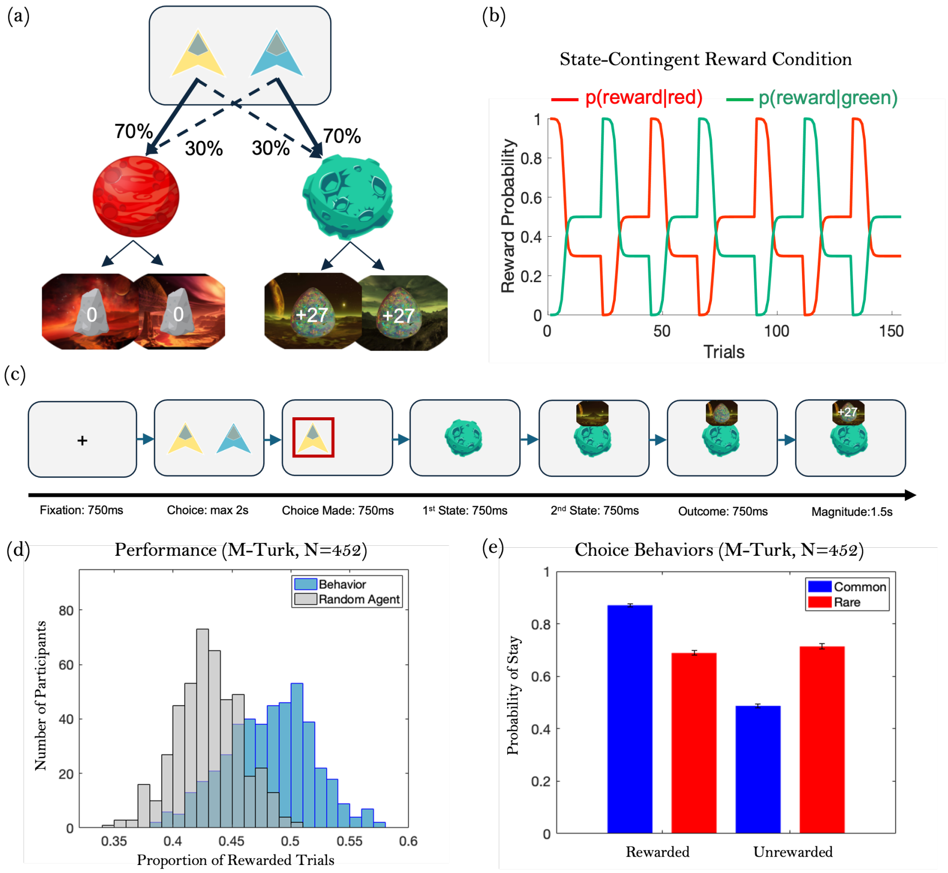


Figure 2.1: The two-step task structure and the group-level performance and choice behaviors. (a) The two-step task structure. In each play, participants started by choosing from two spaceships, and each spaceship had a common and rare transition to the two possible planets. Then, from the planet to the landing pads, the transition is manipulated to be probabilistic, and both landing pads associated with a given planet share the same reward probability function (see b). (b) The reward probability structure in the state-contingent reward condition. The red trajectory denotes the dynamics of reward probability of the landing pads on the red planet, whereas the green trajectory denotes that of the landing pads on the green planet. Same structure for stimulus-contingent reward condition (see Figure 2.2a). (c) The sequence of events of an example trial. After a fixation screen of 1s, the two stimuli showed up (yellow ship on the left and blue ship on the right as illustrated), and the participant had a maximum of 2s to make a choice; Choice-Planet interval: 1s; Planet-Pad interval: 1s; Pad-Outcome interval: 1s; Outcome-Fixation: 1s. (d) The distribution of the probability of obtaining rewards for a simulated random agent and the actual group performance. The light blue bars are the group-level probability of obtaining rewards (N=452). The grey bars are the counterpart distribution with a random agent running through the trial sequence experienced by each participant. (e) The stay probability as a function of the previous outcome (i.e., reward or no reward) and previous first-transition type (i.e., common in blue or rare in red). The overall group behaviors show a higher stay probability after a rewarded trial (MF-consistent) as well as an MB-consistent stay or switch choices (depending on the previous outcome) to a rare transition experienced in the previous trial. The black error bars reflect the within-subject SEMs.

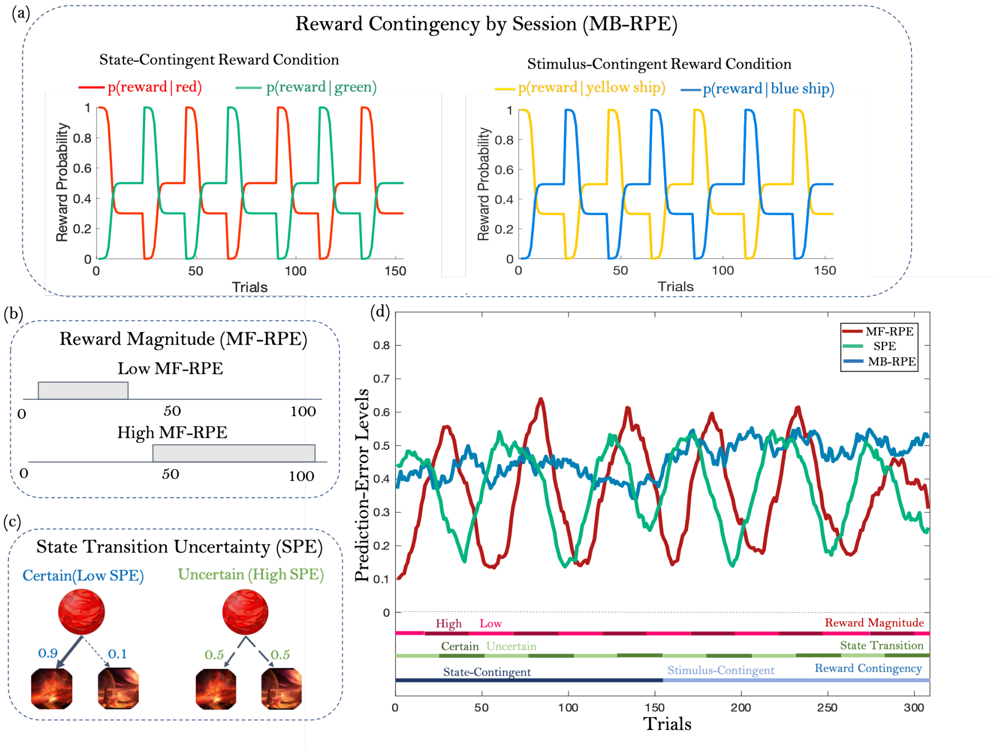


Figure 2.2: Manipulations of three reliability signals (MF-RPE, SPE, and MB-RPE). (a) Reward contingency manipulation. The high and low levels of MB-RPE were realized by applying the same reward probability structure to different reward contingency targets in separate sessions: 1) contingent on the arrived red vs. green outcome states (state-contingent) and 2) contingent on the chosen spaceship (stimulus-contingent). The reward probability trajectory is constituted of multiple sigmoid curves with reversals between 0% and 100% along each contingency session. (b) Reward magnitude manipulation. The varying levels of MF-RPEs were achieved by varying the reward magnitude range along the scale of 0-100. To achieve low MF-RPE, the magnitude was sampled from a uniform distribution of (30, 100); for high MF-RPE, the magnitude was sampled from a uniform distribution of (30, 100). (c) State-transition uncertainty manipulation. The high and low levels of SPE were achieved by manipulating the probabilistic second-stage transition (from the planet to the landing pad) to have high vs. low uncertainty. Under low uncertainty, a given planet will direct the spaceship to one landing pad with 90% probability and to the other with 10% probability, and the identity of the landing pad to which the spaceship would travel at 90% probability was randomized across periods of low state-transition uncertainty at the within-subject level. Under high uncertainty, the probability of transition to either of the two landing pads is 50%. (d) Example dynamics of three reliability signals under manipulations of reward magnitude, state-transition uncertainty and reward contingency. The plot shows an example of the averaged PE dynamics for a given design of task dynamics. The two separate reward-contingency sessions are concatenated here for illustration purposes. The PE signals were extracted by using group-level optimized parameters and the PE trajectory was moving-averaged with a 20-trial window.

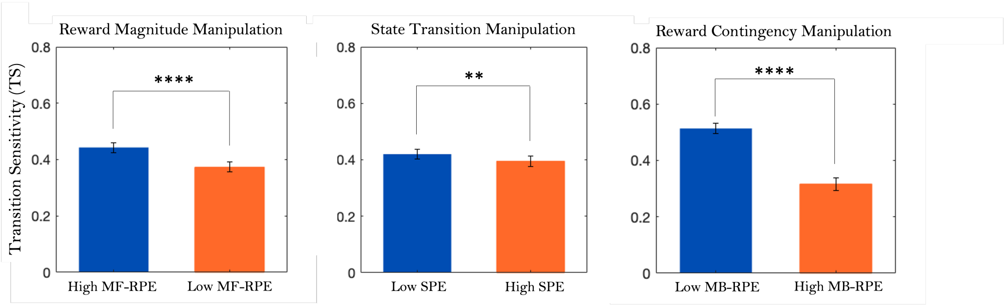


Figure 2.3: Transition Sensitivity across different levels of three reliability signals. Left: The TS measure was calculated across High vs. Low MF-RPE conditions for each subject. Middle: High vs. Low SPE conditions. Right: High vs. Low MB-RPE conditions. The black error bars reflect the within-subject SEMs, $****P{<}0.0001, ***P{<}0.001, **P{<}0.01, *P{<}0.05, n.s.P \geq 0.05$.

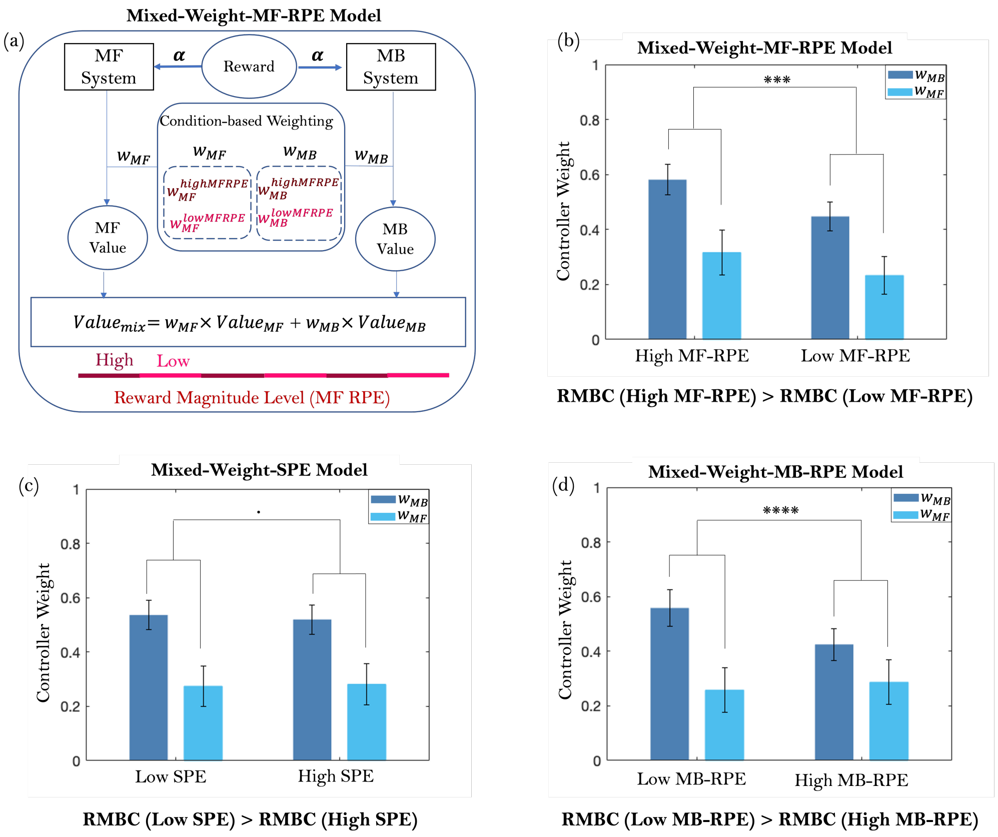


Figure 2.4: Illustration of an example condition-based arbitration model and the shift relative MB control under manipulations of three reliability signals. (a) The arbitration model based on the manipulated MF-RPE conditions, namely the Mixed-Weight-MF-RPE Model. Both MF and MB systems learn the stimulus value through binary reward outcomes with a common learning rate α. Exogenous independent conditional weights associated with high MF-RPE and low MF-RPE were fitted to the periods of interest to combine MF and MB values, respectively, for the computation of the mixed values. A softmax function was used on mixed-value for action selection. The arbitration model for SPE and MB-RPE conditions shared the same arbitration structure and the configuration of parameters, except the conditional weights were adjusted in the appropriate periods to its corresponding state-transition uncertainty condition and reward contingency condition (not shown). All three condition-based arbitration models were fitted independently for each subject. (b) Comparison of the relative model-based control (RMBC) across high vs. low MF-RPEs conditions by fitting the Mixture-Weight-MF-RPE Model. Dark blue bars show the fitted MB weight and the light blue bars show the fitted MF weight across High vs. Low MF-RPE conditions, the RMBC measure in a given condition is defined as the difference between the MB weight minus the MF weight in a given condition. (c) Comparison of the relative model-based control (RMBC) across high vs. low MF-RPEs conditions by fitting the Mixture-Weight-SPE Model. (d) Comparison of the relative model-based control (RMBC) across high vs. low MF-RPEs conditions by fitting the Mixture-Weight-MB-RPE Model. (b-d) The black error bars reflect the within-subject SEMs, $* * * * P<0.0001, * * * P<0.001, * * P<0.01, * P<0.05, n.s. P \geq 0.05$.

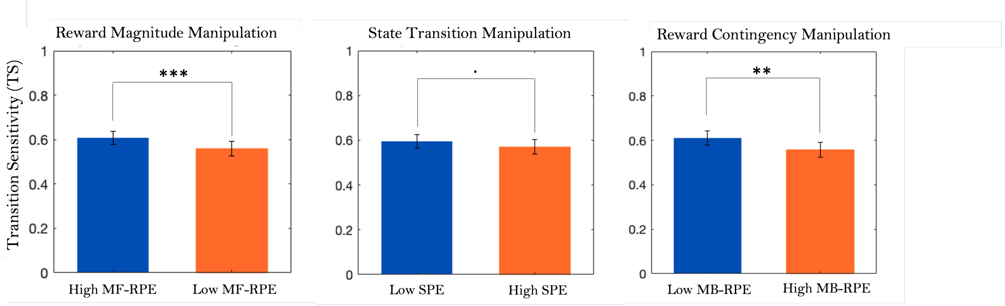


Figure 2.5: Transition Sensitivity across different levels of three reliability signals in the replication sample. Left: TS measure across High vs. Low MF-RPE conditions. Middle: TS measure across High vs. Low SPE conditions. Right: TS measure across High vs. Low MB-RPE conditions. The black error bars reflect the within-subject SEMs, $* * * * P<0.0001, * * *P<0.001, * * P<0.01, *P<0.05, n.s.P \geq 0.05$.

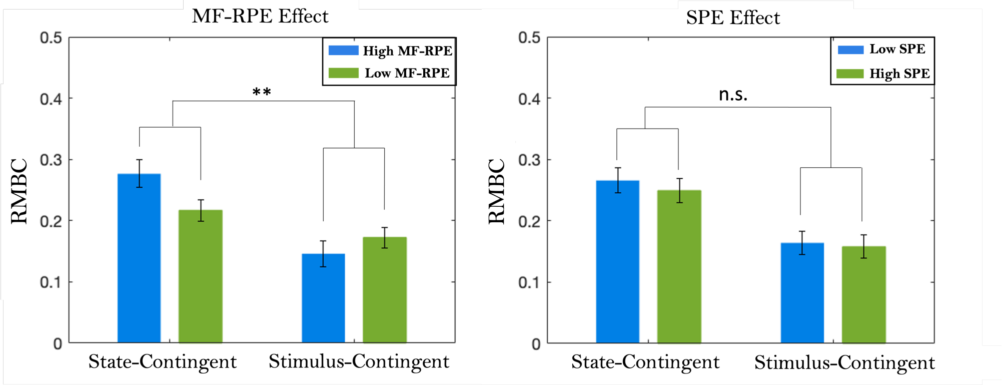
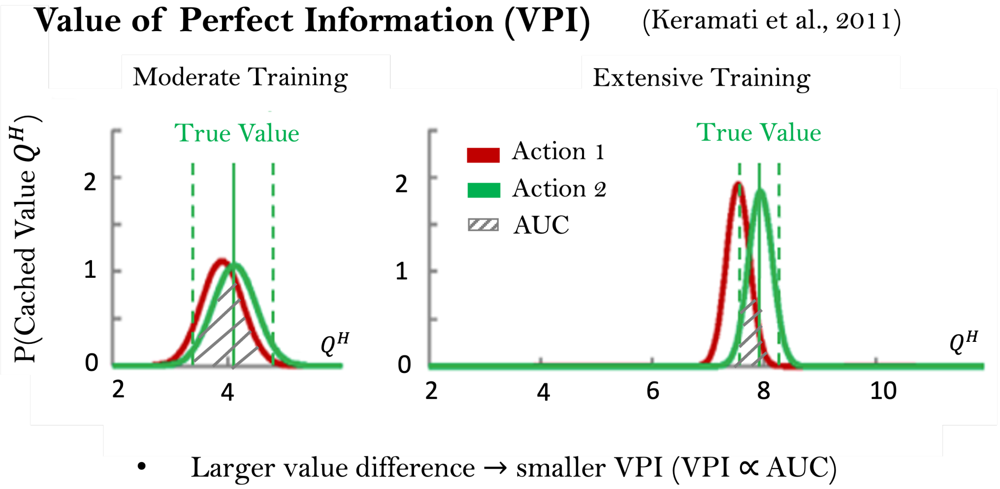


Figure 2.6: Inter-Reliability Arbitration between MF-RPE/SPE and MB-RPE. Left: RMBC shift high vs. low MF-RPEs as a function of the reward-contingency context; Right: RMBC shift across high vs. low SPEs as a function of the reward-contingency context. The black error bars reflect the within-subject SEMs, $* * * * P<0.0001, * * *P<0.001, * * P<0.01, *P<0.05, n.s.P \geq 0.05$.



Value of Perfect Information (VPI) (Keramati et al., 2011)

- Larger value difference → smaller VPI (VPI ∝ AUC)

Figure 2.7: Prediction testing of alternative arbitration theories. (a) The speed/accuracy trade-off arbitration theory proposed by Keramati et al. (2011). The value of perfect information (VPI) is computed and compared against the deliberation cost to decide on the activation of the goal-directed control (MB system in the current context). In testing sensitivity to outcome devaluation, compared to moderate training (left), extensive training (right) resolves the uncertainty of the two concurrent action values as the variance of the distribution mean decreases, and the VPI decreases to a smaller level, and thus it is less beneficial to activate the goal-directed control, and the behaviors are expressed as devaluation insensitive. The solid line and the two dashed lines illustrate the mean and the variance of the estimated true value (figure adapted from Keramati et al.(2011)). (b) The arbitration theory proposed by Pezzulo et al. (2013). The value of information for a given action is computed as the ratio of value uncertainty over the difference between the given action and the alternative action. Hence, the smaller the value difference between the two concurrent actions, the larger VoI would be, and the more beneficial it would be to activate the goal-directed control considering the cost of mental simulation. (c)Transition Sensitivity across small and large value differences. The black error bars reflect the within-subject SEMs, $****P<0.0001, ***P<0.001, **P<0.01, *P<0.05, n.s.P \geq 0.05$.

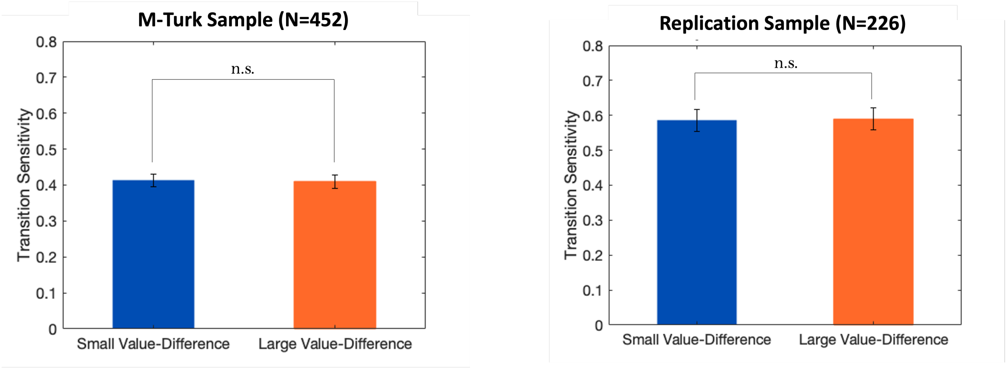


Figure 2.8: Transition Sensitivity Measure across small vs. large value difference. Left: M-Turk Sample (N=452); Right: Replication Sample (N=226). The black error bars reflect the within-subject SEMs, $* * * * \ P$<0.0001, $* * *P$<0.001, $* *$ P<0.01, $*P$<0.05, $n.s.P \geq 0.05$.

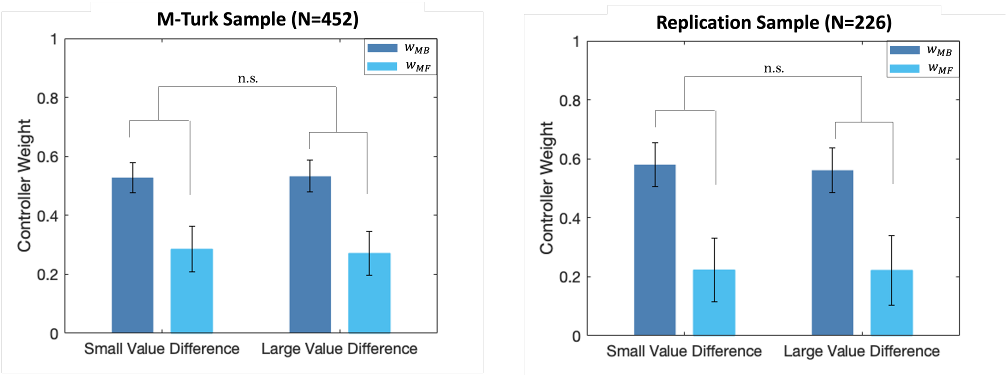


Figure 2.9: Comparison of the relative model-based control (RMBC) across small vs. large value difference conditions by fitting the Mixture-Weight-Value-Difference Model. Left: M-Turk Sample (N=452); Right: Replication Sample (N=226). The black error bars reflect the within-subject SEMs, $* * * * \ P$<0.0001, $* * *P$<0.001, $* *$ P<0.01, $*P$<0.05, $n.s.P \geq 0.05$.

Chapter 3

INVESTIGATIONS OF REWARD PREDICTION AND SYSTEM ENGAGEMENT WITH GROUP-LEVEL AND INDIVIDUAL-LEVEL FMRI IN THE TWO-STEP MARKOV DECISION-MAKING TASK

ABSTRACT

The model-free (MF) and model-based (MB) algorithms in the field of reinforcement learning (RL) field could decently characterize the reflexive habitual control and the reflective goal-directed control identified in animal and human instrumental learning behaviors. A Markov two-step decision task has been designed and used to study the use of MF and MB systems and their neural correlates. However, it has been recently argued that the two-step task was insufficient to solicit the engagement of MF control in human participants, and the previously identified RPE signal in the striatum might be due to the misattribution of variance from the reward signal. In this study, by having a large number of participants complete the two-step task with the measurement of functional magnetic resonance imaging (fMRI), we investigated the basic learning and decision-making computations entailed in the MF and MB system at the group level and specifically conducted statistical tests on the existence of the RPE signal from both systems in the striatum. Moreover, we deployed the approach of individual differences to classify participants into sub-groups of distinct RL and non-RL strategies, which provided a detailed profile of the behavioral strategy engaged in the two-step task and the neural substrates underlying these different behavioral strategies. In sum, we found evidence of both MF RPE and MB RPE encoding in the striatum, with behavioral strategies on the MF-MB spectrum present in the two-step task, demonstrating the validity of using the two-step task to study RL-related strategies with varying degrees of MF system. Our finer-grained neural analysis on decision utility within both RL and non-RL strategy groups also shows the ubiquitous involvement of the MF system in Markov decisions. Potential evidence of an MF-default control allocation process was also discussed.

3.1 Introduction

It has been proposed that multiple decision-making systems underlie choice behaviors(Kahneman, Frederick, et al., 2002; Loewenstein and O'Donoghue, 2004; Killcross and Blundell, 2002; Dickinson and B. Balleine, 2002; Rangel, Camerer, and Montague, 2008; Sloman, 1996), and a reflexive habitual system and a cognitive goal-directed system are two contrasting learning strategies that could equally well guide behaviors with their own signatures (B. W. Balleine and Dickinson, 1998). Researchers have used two types of reinforcement learning (RL) algorithms to characterize these two contrasting decision-making strategies: model-free (MF) control for habitual system (Houk, Adams, and Barto, 1994; Schultz, Dayan, and Montague, 1997 and model-based (MB) control for goal-directed system (Sutton, 2018). The MF control navigates in an environment with rewards through trial and error, accumulating values for potential actions by experiencing the associated outcomes. On the contrary, the MB control uses a "World Model" of the environment, which has access to the learned information of state transitions upon actions, facilitating flexible online planning. Although the MF and MB controls have qualitatively different learning mechanisms by definition, one common mechanism for both controllers to perform well is the process of reward learning. The temporal-difference error (TD error), of which the neural signal has been repeatedly found in the striatum (Schultz, Dayan, and Montague, 1997; O'Doherty et al., 2003; McClure, Berns, and Montague, 2003), serves as a critical building block of the system for both MF and MB controller to navigate in the environment to collect rewards effectively.

However, there has been some recent debate (Feher da Silva and Hare, 2020; Feher da Silva, Lombardi, et al., 2023) on the existence of such TD error signal in the human striatum, specifically when a Markov decision two-step task was used (Daw et al., 2011), arguing that some of the previously observed RPE signals in the striatum could simply reflect the information of reward itself. Given the prevalent usage of the two-step task in characterizing the MF and MB control use and in studying the corresponding neural processes, it is necessary to understand better the complete picture that the current debate is on and investigate further into the validity of previous TD error finding and the usage of the two-step task. Also, along with the doubt about the MF RPE signal's existence in the brain, the use of the MF system in the two-step task is also in question, for which a finer-grained behavioral characterization of the strategies within the RL and the non-RL categories, as well as neural investigations into the existence of key MF learning and decision-making signals, are needed.

To answer these questions, first, by using the model-based fMRI approach, we re-examined the existence of MF RPE and MB RPE neural signals, specifically in the striatum when human participants were engaged in a two-step task using reinforcement learning strategies. Such investigation can, through neural measures, speak to the engagement of the MF system in the two-step task. Secondly, by leveraging the individual differences in usage of MF vs. MB strategies in the two-step task and specifically clustering on participants' behavioral metrics, we also established individual groups of distinct strategies within the RL spectrum and beyond. The details of all possible RL strategies identified through the clustering process provided a better picture of the extent to which participants engaged the MF system in solving the task. Importantly, the usage of the MF and MB systems could also be reflected in the encoding profile of the corresponding decision utility signal from the two systems both at the overall group and at the sub-group level, where each sub-group used distinct behavioral strategies according to the cluster allocation. Examining the overall encoding of MF and MB decision signals at the level of the entire participant pool could first hint at the engagement of the dual systems in the decision-making process, where the MF decision signal is specifically in our research interest. Moreover, the finer-grained classification of behavioral strategies would also allow us to examine how dual systems' decision signals are neurally implemented across groups using distinct strategies (both RL and non-RL strategies). Hence, we can not only validate the behavioral cluster allocation of strategies such as MF and MB strategies by searching for evidence of the MF and MB decision signal from the MF and MB systems, respectively, but also investigate how the decision signal of one RL system (e.g., MF system) is engaged when the other RL system (e.g., MB system) also exerts behavioral influence, or when the behavior is more under the influence of some non-RL strategies. Specifically, the engagement of the MF system across distinct behavioral strategies beyond an MF strategy could be reflected in the encoding of the MF decision signal across all behavioral groups we classified. This would also directly speak to the involvement of the MF system in the Markov two-step task or beyond. In the literature studying the MF and MB system, the MF and MB values were found to be encoded in the medial prefrontal cortex (mPFC) (Beierholm et al., 2011; Hampton, Bossaerts, and O'doherty, 2006; S. W. Lee, Shimojo, and O'doherty, 2014); hence mPFC would be our region of interest for this purpose.

In addition to tracing down the actual RL computations and the neural substrates in the two-step task, it is natural to ask how the MF and MB systems cooperate or compete with each other to express adaptive behavioral outputs. A reliability-based arbitration framework, where prediction errors serve as the proxy of the system reliability, has been proposed accordingly and empirically tested (S. W. Lee, Shimojo, and O'doherty, 2014; Cockburn et al., n.d.). In this previous work, Lee and colleagues (2014) showed evidence that supports a potential control allocation hypothesis that between the MF and MB systems, the MF system is the default system in use and only gets inhibited when more MB control is needed according to the arbitrator via tracking the system reliability. One potential deduction of this hypothesis is that the computational variables, such as value or decision-related signals of the default MF system, should be present no matter what category the behaviors fall into MF, MB, Mixture, or even the non-RL group, whereas the computations associated with the MB system should only be present when the behaviors are under the guidance of MB control as the MB computations would only get manifested when the default MF control is inhibited. If this is true, we should then observe MF value/utility signals represented in the region of interest (i.e., medial prefrontal cortex) in the MF, MB, and Mixture group; furthermore, we should only observe the neural representation of MB value/utility signals in mPFC in the MB and Mixture groups but not in the MF group.

Overall, in this work, with a large sample of participants completing the two-step task while being scanned, we examine the group-level overall encoding of MF RPE and MB RPE signals with a model-based fMRI approach with a special interest in the striatum to address recent literature debate on the MF-existence in the two-step task. Additionally, through clustering participants' behavioral strategy on their choice and reaction time measure, we characterize behavioral strategies with different degrees of MF or MB control used at the individual level: 1) MF group, 2) MB group, 3) Mixture group, and 4) non-RL group. Thanks to the decent sample size of individual groups, we are able to present the profile of distinct strategy groups on the MF-MB spectrum and examine the neural signals for strategy-specific computations (e.g., MF decision utility and MB decision utility) within each behavioral group to exclusively pin down the neural substrates when different strategies on the MF-MB spectrum are engaged. The findings that the MF strategy is used by a decent number of participants and that the MF decision signals are neurally ubiquitous add more evidence of the MF system as a fundamental RL module contributing to behavioral controls in Markov decisions. Also, the neural findings of MF and MB decision signals we found are relevant to a control allocation hypothesis suggesting the MF system is the default controller and the MB system is engaged when needed.

3.2 Results

General Behaviors

Participants were instructed to perform a variant of the two-step task (Figure 3.1a). In each trial, participants started by choosing between two options (i.e., spaceships) at the first stage, which would lead to one of two planets at the second stage in a probabilistic manner — one spaceship reaches one planet more likely than another (70% vs. 30%). Then a transition from the planet to one of the two landing pads would happen before the participants finally observed the outcome, which was either a reward with some magnitudes or no reward. Participants only had to make one choice on each trial, and the reward probability associated with choosing either of the two options fluctuated throughout the experiments (i.e., the more rewarding option would switch between the two). From the perspective of a participant, achieving a good performance depends on constantly learning the action value of the two options throughout the experiment. Representing the internal task structure when necessary (see the next paragraph, e.g., the first stage probabilistic transition from the spaceship to the planet) could also be beneficial by design when the reward is contingent upon the planet they reached. To evaluate how well participants perform the task in general, we calculated the actual probability of obtaining a reward across all trials for each participant and obtained the population distribution of such probability of obtaining rewards. The distribution was compared to a null distribution of the probability of obtaining rewards by simulating the same number of random agents as the number of participants performing the same task. Participants' performance was found to be significantly higher than the chance level indicated by the null distribution ($p < 0.001$paired t-test), where the actual probability of reward from all the participants has a mean of 47.19% ($sd = 0.393$), and the null distribution has a mean of 43.08% ($sd = 0.030$). This suggests that overall the participants learned the reward probability of the two options well to obtain rewards.

Next, we aimed to determine whether participants exhibited choice behavior consistent with model-based (MB) or model-free (MF) reinforcement learning (RL) overall. In this task, as in the original two-step task (Daw et al., 2011), an MF-consistent behavior can be dissociated from an MB-consistent behavior by examining the choice repeating pattern when the transition type in the preceding trial is considered. A model-free (MF) agent would choose the same spaceship as chosen on the preceding trial after receiving a reward on the preceding trial, regardless of whether a common or rare transition occurred on that trial. A model-based (MB) agent, on the other hand, would consider the nature of the state transition that occurred prior to reaching the reward on the preceding trial when making a spaceship choice: when a rare transition happened in the preceding rewarded trial, an MB agent would be less likely to repeat the spaceship choice if the reward is contingent upon the state, favoring instead the alternative spaceship associated with the common transition more likely to lead to the same state (i.e., planet) on the next trial. Prior studies have found that, on average, human behavior on this two-step task is a mix of MB and MF strategies (Daw et al., 2011). In order to diagnose if the behaviors are consistent with MF vs. MB control, as in previous studies, we classified trials by the outcome of the previous trial (i.e., reward vs. no-reward) and the first-stage transition type (i.e., common vs. rare) in the previous trial, and we then examined the probability of choosing the same option in the current trial as in the previous trial (i.e., p(stay)) as a function of outcome and transition type on the previous trial. Qualitatively, as shown in Figure 3.1b, participants expressed a choice sensitivity to the reward, indicated by a higher probability of choosing the same option on the current trial if the chosen option was rewarded in the previous trial (averaging across preceding trials with both common and rare transitions), suggesting a model-free component in participants' reward-maximizing strategy. In addition, participants also showed choice sensitivity to the transition type in the previous trial. They were 1) more likely to switch the choice after a rewarded trial with a rare transition than with a common transition, and 2) more likely to stay with the choice after a no-reward trial with a rare transition than with a common transition, suggesting an MB component in participants' reward-maximizing strategy. Collectively, these results support the typical observation that participants' behavior reflects a mix of both MB and MF components on average.

To quantify the degree of reward sensitivity and transition sensitivity of the current choice, a mixed-effect logistic regression was run to test how the probability of choosing the same option is influenced by the previous outcome, previous transition type, and their interactions (see Methods for details). Consistent with the qualitative behavioral results, an MF-consistent behavioral pattern is observed that participants were more likely to repeat their choice if that choice was rewarded in the previous trial, reflected by an MF-consistent main effect of the previous outcome ($\beta = 0.65716, SE = 0.047577, T = 13.813, p < 0.001$). An MB-consistent effect was also found in that participants were more likely to switch to the other spaceship when they experienced a rare transition towards the reward, reflected by a significant effect of the interaction between the previous outcome and the previous transition type ($\beta = -0.67636, SE = 0.060504, T = -11.179, p < 0.001$). In sum, the participants showed a mixture of MF-consistent and MB-consistent behaviors, replicating the typical behavioral patterns observed in the two-step task (Daw et al., 2011).

Computational Variables and Behavioral Clustering

As participants express a mixture of MF and MB strategies at the population level, it is important to understand whether the manifested mixed behavioral strategies overall could be decomposed into different groups of individuals engaging different strategies in light of their behavioral features. To approach this problem, we relied on an external dataset with a much larger sample (N=678; Cockburn et al., n.d.) to classify the entire participant pool on a set of behavioral features derived from choice patterns and reaction time patterns conditioned on the preceding trial's outcome and transition type (see Methods). There are four cluster centroids from the external dataset, the label of which was created to make a sensible interpretation of behaviors in the two-step task. The breakdown of the four groups is 1) Mixture Group (N=40), 2) MF Group(N=33), 3) MB Group(N=44), and 4) Other Group (N=62), corresponding to the behavioral signature of each group. As shown in Figure 3.1c, by plotting cluster identity against each individual's regression estimates from the mixed-effect regression, the Mixture Group has an intermediate level of both outcome effect and outcome-transition interaction effect, indicating a mixture of MF and MB strategy usage within these individuals. The MF Group tends to show a larger main effect of outcome but a weaker interaction effect of outcome and transition, whereas the MB Group tends to show a smaller main effect of outcome but a stronger interaction effect of outcome and transition. Lastly, the Other Group has low measures on either regression estimate, indicating the potential use of non-RL strategies. When plotting out each group's choice probability as a function of the previous outcome and transition type, stereotyped behaviors of each identified behavioral group were observed (Figure 3.1d). The classification of each individual's behavioral strategy provides a clear footing for studying what neural computations in the RL process are conducted in support of different types of RL strategies. The shared and distinct neural computations could be unveiled by a model-based fMRI approach within each of these behavioral clusters.

After confirming the existence of Mixture, MB, and MF strategies when participants performed the two-step task, we sought to characterize the key MF and MB computations within the corresponding system to facilitate the expression of MF and MB signature. For this purpose, we fitted a hybrid reinforcement learning model with separate MF and MB modules with independent condition-specific weights assigned to the MF and MB system (Cockburn et al., n.d.; see Methods for details). The model's posterior predictions align with the participants' actual behavior quite well (Figure 3.1d), which indicates good model performance so that we can rely on the learning and decision variables used by the model to study their neural substrates. Within the model, the Q-learning algorithm for the MF system is used to learn the magnitude of the outcome and the expected state values of each intermediate stage. In each trial, as there are three transitions, there are, in total, three MF reward prediction errors (RPE) calculated as the difference between the expected state value and the encountered state value, with the state value at the outcome stage as the received reward magnitude. In detail, the first MF-RPE is the RPE encountered at the planet stage after choosing the spaceship, the second MF-RPE is the RPE encountered at the transition from the planet to the landing pad, and finally, the third MF-RPE is the difference between the received reward magnitude and the value of the landing pad. In parallel, the MB system represents the three MB reward prediction errors in a similar manner as the MF system — computing the difference between the expected state value and the actual encountered, although the MB agent learns value differently. Specifically, the MB system first learns the state value at each stage by learning the final state value through the final outcome and, subsequently, with the transition probability between states, computing the expected values of previous states. Similar to the MF system, the first MB-RPE, second MB-RPE, and third MB-RPE are calculated within each trial as the transition goes from the chosen spaceship to the final outcome observed. One important difference between the MB and MF systems is that the MB system learns binary reward outcome (reward vs. no reward) as opposed to magnitude reward in the MF system, and this is to account for the task design that choice optimality depends on reward probability rather than the magnitude itself, which participants are instructed on beforehand. Consequently, the MF RPEs and MB RPEs at the three stages are dissociable and could be leveraged to investigate their corresponding neural correlates through the model-based fMRI analysis, which was conducted at the population level rather than within each of the behavioral groups.

Besides the reward prediction error signal from the MF and MB systems, the decision utility signals from each system are also of interest to understand the RL computations engaged in the two-step task as the construct of decision utility is critical for each system to enable the downstream action selection process. The decision utility of each option could be acquired over a relatively longer time scale (e.g., multiple past trials) and simultaneously also sensitive to rewards obtained recently, for example, in the preceding trial (Iigaya et al., 2019). The learning mechanism just described in the previous paragraph is known as a slow learning mechanism that integrates past reward information over multiple past trials. In addition to such "slowly" learned option value, a fast learning mechanism is also incorporated into the model for MF and MB systems, respectively, to capture the "fast" value component updated on a trial-by-trial basis. For the MF system, a value component would be added or subtracted based on the outcome of the preceding trial, whereas for the MB system, the value component would be adjusted not only based on the outcome but also on the transition type in the preceding trial. Hence, the decision utility for MF and MB systems is computed as the chosen option value minus the rejected option value, where the option values are composed of "slow" and "fast" value components. The MF and MB decision utilities derived from the model were studied within each behavioral group identified by the clustering algorithm with a model-based fMRI approach. It is an intriguing question as to how MF and MB systems are engaged at the neural level across groups with distinct behavioral strategies, and we specifically approach the problem through the lens of decision utility, given it is a key computational variable in the reinforcement learning process.

Neural Correlates of Reward Prediction Error in the Model-Free and Model-Based System

As for the model-based fMRI analysis, an event-related design matrix was used for the general linear model fitted to the fMRI BOLD data (see Methods for details). The events of interest in the design matrix for each trial are stimulus onset, response onset, planet onset, pad onset, and outcome onset, which are set as stick functions at the corresponding event onset time. Since there are three RPE signals from the MF and MB systems arising in each trial, the derived three RPE variables are set as the parametric modulators at their corresponding time points: planet onset, pad onset, and outcome onset. Importantly, to increase the statistical power of capturing the RPE-specific variance in the BOLD signal, we combine the three event regressors into one chained regressor "planet-pad-outcome onset," which essentially means for each trial, there are three stick functions built into the regressor rather than one stick function per trial. The combined "planet-pad-outcome" regressor, which entails three stick functions, is parametrically modulated by both the MF RPE and MB RPE at each stage. Consequently, using the combined event regressor for the RPE signal would have three times more observation points than setting three separate RPE regressors at each stage, but at the same time, would not differentiate RPE encoding

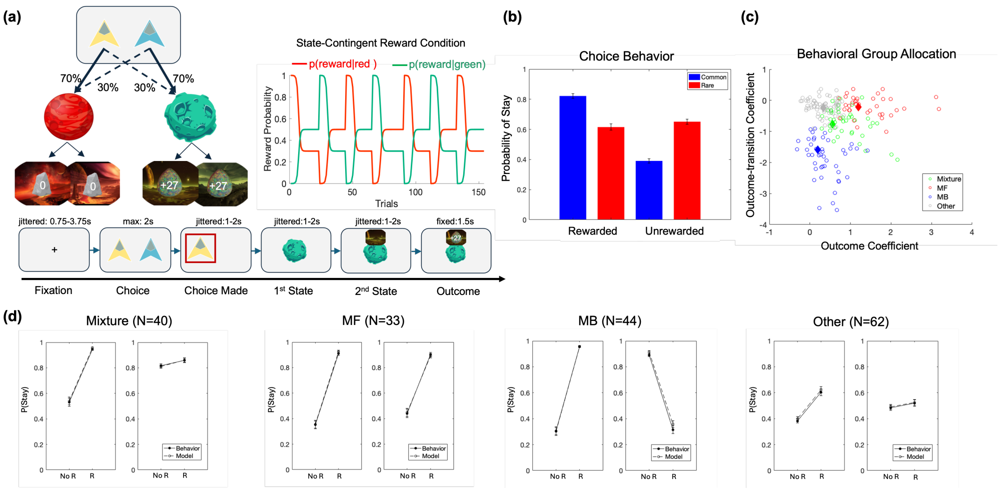


Figure 3.1: The two-step task structure and the group-level performance and choice behaviors. (a) Left: The two-step task structure. In each play, participants started by choosing from two spaceships, and each spaceship had a common and rare transition to the two possible planets. Then, from the planet to the landing pads, the transition is manipulated to be probabilistic, and both landing pads associated with a given planet share the same reward probability function. Right: The reward probability function throughout the task; Bottom: The timing of the trial sequence. (b) The probability of repeating the previous action as a function of the previous trial's outcome and transition type. (c) The clustering result after allocating the participant's behavioral features to the four cluster centroids identified from an external large dataset. Four diamonds denote the four centroids from the external dataset. One empty circle denotes one individual in our fMRI sample. (d) The line plot of the information shown in (b) for each behavioral group using group membership in (c). The solid lines are the participants' actual behavior, and the dashed lines are the predicted behavior from the computational model fitted to the participants. The black error bars reflect the across-subject SEMs.

at different trial stages. Also, both MF RPE and MB RPE are simultaneously put as parametric modulators of the combined event regressor without orthogonalization so that the identified signal ascribed to either MF or MB RPE regressor is beyond the shared variance and belongs to the regressor itself. The three-stage MF RPE and MB RPE signals are both z-scored across all trials, and we fitted the first-level GLM to each participant with the outcome magnitude as a parametric modulator at the outcome onset stage to control for signal-related outcome magnitude. For data inclusion in this analysis, as we are interested in the RPE computations in RL, we excluded the Other Group and reported the results from analyzing the data of the three RL-related behavioral groups (i.e., the Mixture, MF, and MB Group).

From the group-level analysis of the MF RPE and MB RPE, we found that neural activity in the right caudate significantly correlated with the combined MF RPE signal ($p = 0.025$, cluster-level FWE, small-volume corrected, Figure 3.2a left) and found that neural activity in the left ventral striatum significantly correlated with the combined MB RPE signal ($p = 0.026$, cluster-level FWE, small-volume corrected, Figure 3.2a middle). After establishing the existence of MF RPE and MB RPE computations in the brain, we sought to investigate whether the just-described neural results are exclusively attributed to the RPE signal in each system. In other words, it is a question of whether the neural activity in the right caudate significantly correlated with MF RPE would also significantly encode the RPE signal from the MB system and vice versa for the neural activity in the left ventral striatum. To answer this question, we ran a group-level paired t-test on beta maps of the MF RPE and MB RPE in the regions previously identified for the RPE signals. Specifically, we tested whether, in the right caudate, the beta maps for MF RPE are higher than that for MB RPE and tested whether the betas coefficients for MB RPE in the left ventral striatum are higher than that for MF RPE. We found that there was a cluster in the right caudate that has higher betas for MF RPE than for MB RPE ($p = 0.031$, cluster-level FWE, small-volume corrected, Figure 3.2a right). On the other hand, we did not find any clusters surviving correction in the ventral striatum that encode MB RPE more strongly than MF RPE. We also took an ROI-based analytic approach to investigate how the MF RPE and MB RPE signals are encoded in the caudate and the ventral striatum by extracting the mean beta coefficients of the RPE contrasts within the pre-defined ROIs (see Methods: Regions of Interest and Small Volume Correction), respectively. Specifically, we found that, in the Caudate ROI, MF RPE were encoded with a trending significance ($p = 0.0727$), whereas the MB RPE was not significantly encoded in the region (Figure 3.2b); in contrast, in the ventral striatum ROI, MB RPE, but not MF RPE, was significantly encoded in the region ($p = 0.0148$, Figure 3.2b). The ROI-based results further illustrated the potential specificity of the striatal regions in the encoding of RPE signals, as we found in the voxel-based analysis on the brain maps.

Neural Correlates of Decision Utility Underlying Model-Free and Model-based Strategy

To study what RL computations are carried out neurally to give rise to the different behavioral strategies that emerge in the two-step task, we built into the design matrix decision utility variables from both MF and MB systems to unveil their neural correlates. Specifically, the chosen utility and rejected utility of the MF and

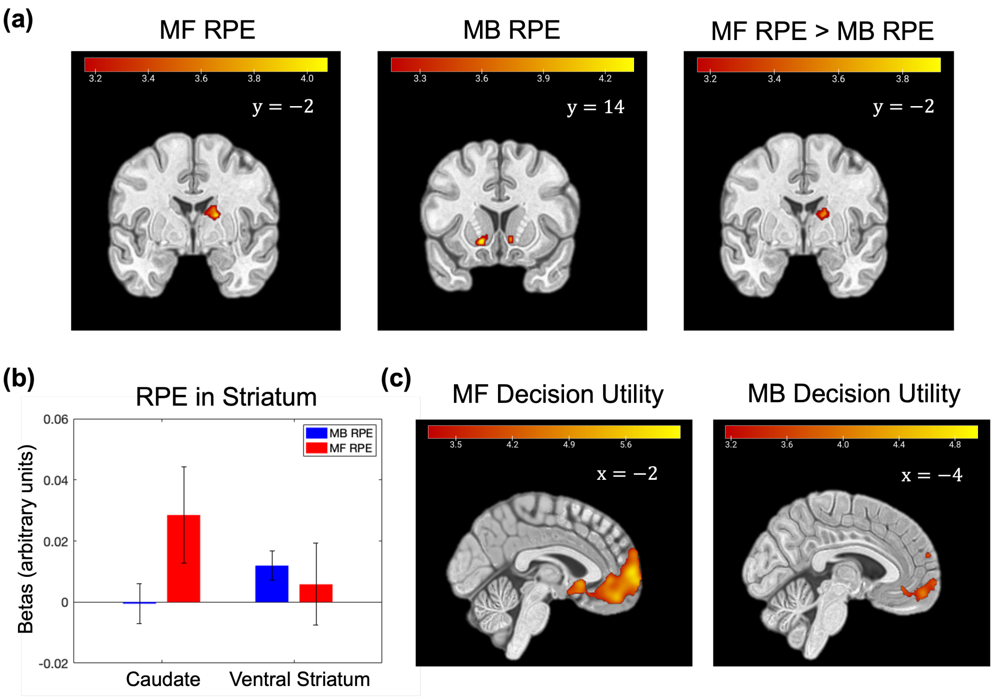


Figure 3.2: The encoding of decision utility and reward prediction error at the group level. All brain images show the T-maps with cluster-forming threshold of p<0.001, uncorrected, in the corresponding region of interest, k denotes the number of voxels in the cluster; SVC denotes the application of small volume correction. (a) Left: The cluster in the right caudate that correlates with MF RPE, k=20, peak voxel (14, 2, 16), T=3.58, SVC; Middle: The clusters in bilateral ventral striatum that correlates with MB RPE, only the cluster in the left ventral striatum survives the small-volume correction, k=51, peak voxel (-14, 14, -8). T =4.49; Right: The cluster in the right caudate has a stronger correlation with MF RPE than with MB RPE, k=16, peak voxel (14,-2,12), T =3.54. (b): The group-level beta coefficients of the MF RPE and MB RPE contrasts in the caudate and the ventral striatum ROIs (t-test, MB RPE in Caudate: $p > 0.05$; MF RPE in Caudate: $p = 0.0727$; MB RPE in ventral striatum: $p = 0.0148$; MF RPE in ventral striatum: $p > 0.05$). (c) Left: The cluster in vmPFC that correlates with MF decision utility (i.e., the contrast of chosen MF utility minus rejected MF utility), k=3486, peak voxel (-4,62,6), T= 6.34; right: The cluster in vmPFC that correlates with MB decision utility (i.e., the contrast of chosen MB utility minus rejected MB utility), k=1297, peak voxel (-12,58,30), T= 5.04.

MB systems were put simultaneously as parametric modulators for the stick function at the time of stimulus onset (i.e., spaceship onset). The same GLM was run for both the previous section on RPE and the current section on decision utility but with different focused contrasts (see Methods for details). We fit the first-level GLM to each individual and looked into the group-level results as a function of the entire group as well as different behavioral groups: 1) Mixture Group, 2) MF Group, 3) MB Group, and 4) Other Group. For the overall group-level analysis, similar to the group-level analysis on RPE, we excluded the Other Group and reported the results from analyzing the data of the three RL-related behavioral groups.

Firstly, when examining the contrast of decision utility (i.e., chosen utility minus rejected utility), we observed a significant neural cluster correlating with MF decision utility ($p < 0.001$, cluster-level FWE, Figure 3.2c left) and MB decision utility ($p < 0.001$, cluster-level FWE, Figure 3.2c right) within the ventromedial prefrontal cortex area (vmPFC), replicating the previous findings on neural correlates of MF and MB value signals (Hampton, Bossaerts, and O'doherty, 2006; Beierholm et al., 2011; S. W. Lee, Shimojo, and O'doherty, 2014). Interestingly, when looking into each individual behavioral group, the MF decision utility was represented significantly by all three RL-related groups and the non-RL group in vmPFC (cluster-level FWE, MF Group: $p < 0.001$; Mixture Group: $p = 0.023$; MB Group: $p < 0.001$; Other Group: $p < 0.001$; Figure 3.3a), meaning that MF decision utility was also computed in the groups whose behaviors aligned more with a pattern of MB control or in the group using a presumably non-RL strategy. In contrast, the MB decision utility in vmPFC was predominantly found in the MB group but with no significant clusters (cluster-forming threshold: p<0.001) surviving corrections for multiple comparisons (FWE, p<0.05) in vmPFC in the Mixture, MF, or Other Group group (cluster-level FWE, $p < 0.001$; Figure 3.3b)

Using an independent ROI-based analysis focusing on the vmPFC, a mask derived from a neural meta-analysis on more than 200 studies of the valuation process (Bartra, McGuire, and Kable, 2013), similar results were found as the whole-brain analysis that MF decision utility was represented significantly in all three groups using RL strategies — in the Mixture group ($p < 0.001$, one-sample t-test), MB group ($p = 0.0249$, one-sample t-test), and MF group ($p = 0.0155$, one-sample t-test). In contrast, the neural encoding of MB decision utility was mainly in the MB group ($p < 0.001$, one-sample t-test), only showing the encoding tentatively in the hypothesized direction in the Mixture group ($p = 0.3527$, one-sample t-test) but not showing positive correlations at all in the MF group ($p = 0.7393$, one-sample t-test). To estimate the overall signal correlations in the vmPFC ROI and to compare the representation of the decision utility of MF and MB systems from group to group, we ran two regression models on the 1st-level beta estimates in the vmPFC (ROI) with an intercept and an ordinal group variable, one for MF decision utility and the other for MB decision utility (see Methods). Consistent with the whole-brain fMRI analysis, we found, in the vmPFC ROI, there are significant overall positive correlations for MF decision utility ($\beta = 0.2499, p = 0.036039$) and for MB decision utility ($\beta = 0.66366, p < 0.001$), indicating the significant intercepts of the two regression models, respectively. As for the across-group comparison of the neural representation of decision utility, patterns from the MF and the MB system diverge. As shown by an insignificant ordinal group effect on the MF decision utility ($\beta = -0.011614, p = 0.8387$, Figure 3.3c), the representation of MF decision utility in vmPFC across the three groups was comparable regardless of the extent to which model-free behaviors are expressed. On the contrary, the MB decision utility was represented in vmPFC differently across the three groups, as indicated by a significant ordinal group effect ($\beta = -0.24765, p = 0.0033$, Figure 3.3c). More specifically, the representation of MB decision utility in vmPFC was most strongly represented in the MB group, but as the MB component in the behavior shrank, the neural representation of MB decision utility waned. In addition to the comparison in the behaviorally categorical group, we also examined how neural representation of decision utility varied across varying degrees of MF (or MB) expression through a more continuous measure: the MF weight parameter (wMF) from the computational model. Generalizing the previous findings across categorical behavioral groups, through Spearman Correlations, it was found that when the individual's behavior went from MB to MF (increasing wMF), the representation of MF decision utility did not shift significantly (Spearman $r = -0.0113, p = 0.9033$, Figure 3.3d left) while the representation strength of MB decisions utility significantly dropped (Spearman $r = -0.3370, p < 0.001$, Figure 3.3d right).

Based upon the observation that both MF decision utility and MB decision utility were represented in the behaviorally MB group but only MF decision utility was represented in the behaviorally MF group, we conducted a two-sample t-test to formally compare MB group and MF group in terms of whether they represent the MF decision utility and MB decision utility differently. The statistical tests coherently indicated that MB decision utility is more strongly represented in vmPFC in the MB group than in the MF group ($p = 0.001$, cluster-level FWE), yet there is no difference in vmPFC activity correlating with MF decision utility between the MB group and MF group, underscoring the contrasting difference of the decision utility computation underlying an MF strategy vs. an MB strategy.

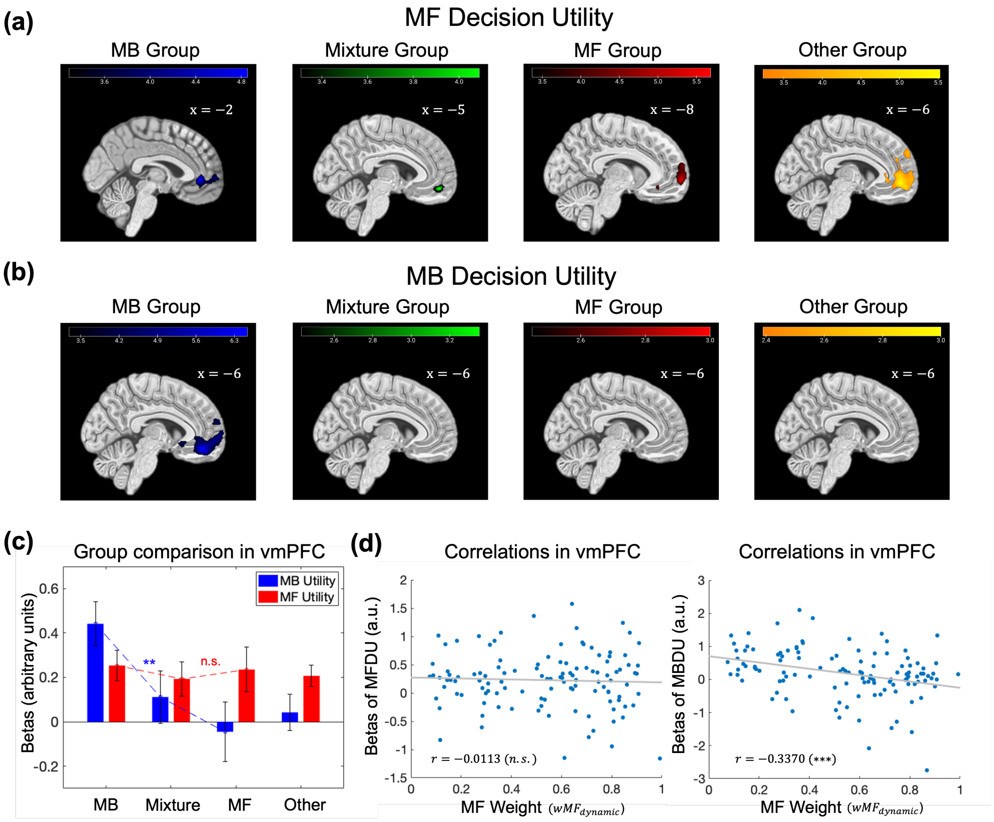


Figure 3.3: The encoding of decision utility across different behavioral groups and across individuals with various degrees of MF (or MB) behavior. (a): The clusters of MF and MB decision utility of each behavioral group using the RL strategy with a cluster-forming threshold of p<0.001, uncorrected, in the region of the prefrontal cortex. (b): The clusters of MB decision utility of each behavioral group using the RL strategy in the region of the prefrontal cortex, with a cluster-forming threshold of p<0.001 (uncorrected) for the MB group and a cluster-forming threshold of p<0.01 (uncorrected) for the Mixture, MF, and Other group. (c): The beta coefficients of the MF and MB decision utility to account for BOLD activity in the vmPFC ROI from all behavioral groups. (d) Left: The correlation between the beta coefficients of MF decision utility in the vmPFC ROI and the weight of MF system from the computational model; Right: The correlation between the beta coefficients of MB decision utility in the vmPFC ROI and the weight of MF system from the computational model (Spearman correlation, "$***$": $p < 0.001$).

3.3 Discussion

With a large group of participants completing the two-step task, an overall mixture of MF and MB strategies was evident in the sample we collected, replicating the previous findings about the behavior in the two-step task (Daw et al., 2011; Kool, Cushman, and Gershman, 2016; Wunderlich, Smittenaar, and Dolan, 2012; Dezfouli and B. W. Balleine, 2013; Otto et al., 2013; Smittenaar et al., 2013; Dezfouli, Lingawi, and B. W. Balleine, 2014). Specifically, finer-grained behavioral clustering analysis enabled us to identify the strategy subtypes that were actually used by the participants. We leveraged the cluster centroids obtained from a much larger external sample (Cockburn et al., n.d.) to assign group identity in our current dataset for classification robustness and reliability. Using clustering features of various behavioral aspects, such as how choice and reaction time are influenced by the preceding trial's information, four groups of participants were identified: 1) Mixture Group, 2) MF Group, 3) MB Group, and 4) Other Group. The number of the behavioral cluster groups was found through the Gap statistical method in the external dataset (Tibshirani, Walther, and Hastie, 2001), and the labels of four formed groups were allocated by evaluating how choices were influenced by the previous trial's outcome and transition type, and each group's reaction time measure was found to be consistent with their distinct choice processes (Cockburn et al., n.d.; Konovalov and Krajbich, 2016). Thus, we first illustrate that a variety of behavioral strategies on the entire spectrum from MF to MB behavior were actually deployed in the two-step task. There have been some recent studies on people's actual strategy engaged in the two-step task and found the MB strategy was primarily engaged if participants were given detailed and story-based instruction, and there was no involvement of the MF strategy (Feher da Silva and Hare, 2020; Feher da Silva, Lombardi, et al., 2023). With our current large sample of participants, who also received detailed and story-based instructions and whose behavioral strategies were identified based upon a reliable criterion set by an external sample, we showed that strategies with MF components (Mixture Group and MF Group) are also evident in a large number of participants' behaviors in the two-step task. This first helps validate that in the two-step task, the MF strategy is also expressed in a certain proportion of the entire participant population besides the MB strategy, and it shows the possibility of arbitration between the two strategies within the same task as well as the possibility of studying the neural substrates of the MF and MB strategies and the mix of both given a spectrum of behaviors are present in this task. We also note that there is a group of participants (Other Group) using apparently non-RL strategies in the task, who could use certain heuristics or simply be less engaged in the task. Although our main analysis of group-level fMRI or individual differences does not include these participants that used non-RL strategy, it suggests that, indeed, a variety of behavioral strategies could be engaged in the two-step task, and special care needs to be taken to tease apart different kinds of strategies before drawing conclusions to avoid misattributing certain behavioral results due to co-existence of multiple strategies and inattentive behavior.

By fitting a computational model of reinforcement learning algorithms to the behavior of all participants that used the reinforcement learning strategy (Mixture Group, MF Group, and MB group, but excluding the Other Group), we first validated that the behavioral signatures of the participants can be well captured by a mixture of model-free and model-based RL model through posterior predictive checks. Given the reliable model characterization, we then identified the key RL computations during the two-step task at the group level: 1) the decision utility, which guides the 1st-stage choice on each trial, and 2) the reward prediction errors, which are experienced at multiple stages of the trial to update the learned value of multiple stages. Specifically, the model enabled us to extract the decision utility and reward prediction error signal from the MF and MB systems, respectively. By fitting these extracted computational variable time series to the fMRI BOLD signal, we found that the decision utility and reward prediction errors from both the MF and MB systems were significantly encoded in the brain, which further confirmed the use of both MF and MB systems when learning and making decisions in the two-step task.

The MF and MB decision utility signals were encoded significantly in the ventromedial prefrontal cortex area (vmPFC) across all groups of participants using the RL strategy, indicating a mixture of MF and MB systems are engaged, aligned well with the group-level behavioral finding in this study. When looking into how these learning and value signals were represented in each individual group of participants using the RL strategy, we further found a ubiquitous MF decision utility encoding in vmPFC across all groups, from using a pure MF strategy to a mix of MF and MB strategy to a pure MB strategy. This is a specifically interesting finding as we found neural evidence of engaging the MF strategy by showing that the MF decision utility signal is indeed represented not only in the group of participants who behaved in an MF-consistent manner but also in participants who were mainly deploying an MB or a Mixture strategy or a non-RL strategy. Such a finding further supports the ubiquitous engagement of the MF system in the Markov decision-making process, as shown in the current two-step task. The ubiquity of the MF decision utility signal also supports a hypothesis of the arbitration theory of control allocation (S. W. Lee, Shimojo, and O'doherty, 2014): the MF system serves as the default control system and gets inhibited when MB control is needed, according to the arbitrator. Although it was not tested in the current study how the arbitrator inhibits MF control when system reliability indicates the need, the presence of MF decision utility in the Mixture Group and MB Group suggests that the MF system might always serve in the neural computation even when the behavioral outputs contain the MB elements according to the arbitrator. This is possible because even though some participants' mean behavioral patterns throughout the task might be categorized as MB, a participant's behavior who is using the MB strategy overall might not stick to using the same degree of MB control from the beginning to the end, and there could be unmeasured control shift between MF system and MB system throughout the task, similarly for participants showing an overall mixture of MF and MB behavior. The result of no correlation between the encoding strength of MF decision utility and the degree of the MF behavior is consistent with the view that the MF module always operates at the back-end to provide inputs for the arbitrator. With respect to the encoding of the MB decision utility signal, the gradual decrease of encoding strength of MB decision utility with an increasing degree of using MF strategy further supports the aforementioned hypothesis about the arbitration process between the MF and MB system. It is hence well expected, given such a hypothesis, that we observed the significant encoding of the MB decision utility in the behaviorally MB Group but no encoding trend at all in the MF Group. The fact that the encoding of MB decision utility is weak in the Mixture Group might be a natural product of the cluster group allocation process within our data sample, as each group's neural result is sensitive to some individual-level encoding strength.

We also found a significant encoding of reward prediction error signal from both MF and MB systems in the striatal area of the brain, with MF reward prediction error (MF RPE) in the right dorsal caudate and MB reward prediction error (MB RPE) in the left ventral striatum. Specifically, we found that the right dorsal caudate region exclusively encodes the RPE signal from the MF system. These neural results further consolidate earlier behavioral and neural evidence of the co-existence of MF and MB strategy when participants use an RL strategy in the two-step task, with the current RPE results suggesting that both MF and MB system are updating their own estimated value in the environment through a temporal-difference learning algorithm. A note is that the significant encoding of MF RPE in the current study was not found in the ventral striatum as typically found in early human studies (Berns et al., 2001; Pagnoni et al., 2002; O'Doherty et al., 2003; McClure, Berns, and Montague, 2003). This might be because these early studies used a simple conditioning paradigm where there is no action to make (e.g., in the Pavlovian conditioning paradigm) or the response is so simple that there are no concurrent options. Thus, the neural computation of the RPE in these no-choice situations could be very different from the computation of RPE and how it contributes to value updates in a task that involves flexible choices and more complex task structures, such as in the two-step task. In an earlier study, Jessup & O'Doherty (2011) found that dorsal caudate was engaged in RL-consistent choices, and its BOLD activity was correlated with the prediction error signal in the MF reinforcement learning algorithm. Our ROI analysis of MF RPE in the dorsal caudate region, of which the mask was created from the aforementioned study (Jessup and O'Doherty, 2011), reassures the proposed role of dorsal caudate underlying the MF learning of stimulus-response association, potentially adding further evidence of the dorsal striatum serving as the "actor" in the actor-critic framework implemented in the striatum (Montague, Dayan, and Sejnowski, 1996; B. W. Balleine and O'doherty, 2010). Intriguingly, we found MB RPE signals were significantly encoded within the ventral striatum, a region typically involved in the MF reward learning process. The MB module in our model requires not only the knowledge of state-transition structure but also the binary reward information at the terminal state so that the initial action can be adaptively guided through the state transitions to the more rewarding terminal state. The induced MB RPEs are dissociable from the MF RPE generated from the reward magnitude learning at the outcome stage, enabling us to investigate the corresponding neural signatures of both systems' RPE signals. The finding of the MB RPE signal in the ventral striatum opens up the possibility that a shared neural mechanism of reward learning might underlie both MF and MB strategy; when computing the choice values, the MF system could relay the learning-related signal to PFC directly while the MB system needs to first relate the state-transition information in the cortex (Gläscher et al., 2010) to the learned state value in the basal ganglia for decision value in PFC. Further connectivity work would be needed to confirm such a hypothesis of the MB system in the brain. There have been concerns in recent literature that the MF RPE signal observed in the two-step task could result from GLM design specification, hence misattributing the reward signal to the reward prediction error (Feher da Silva, Lombardi, et al., 2023). We want to note that our current results of MF RPE and MB RPE encoding in the striatum are obtained with a GLM design that addresses such concerns. Although we used a design where the multiple intermediate-stage MF RPE signals are combined into a single regressor (to increase statistical power), by incorporating reward magnitude as a co-regressor at the outcome stage into the GLM to control for BOLD variance related to the reward signal, we could ensure the MF RPE encoding result in the caudate we observed was not mixed with signals from the reward magnitude. Also, using a combined-regressor GLM approach is valid here as in our computational model specification, the MB RPE (binary reward learning) and MF RPE (magnitude reward learning) are distinguishable (although with high correlations) at all stages.

Drawing insights about the neural substrates of reinforcement learning strategy with a group-level analysis across a limited number of participants could sometimes be challenging, considering the diversity in behavioral strategies and the variability of meaningful neural signals across individuals. In the current study, we embraced this challenge by conducting group-level analysis on a large number of participants as well as leveraging the individual differences in the behavioral strategies and the relevant neural computations. Given the identified striatal MF RPE encoding and the stereotyped MF behavior in a decent number of individuals, a literature debate about the usage of the MF system while engaging the two-step task could now be reevaluated with the new evidence. Moreover, the individual difference analysis on decision utility signals across groups with different behavioral strategies also sheds light on the role of the MF system in the control allocation process when both RL and non-RL strategies are engaged. As real-world decisions usually involve choosing strategies in a flexible manner, our current finding could help provide insights into an accurate arbitration model of strategy shifting for future studies.

3.4 Methods

Participants

For participant recruitment, we recruited participants who reside in the United States and who are fluent English speakers and readers. The age range for the studies is from 18 to 65 years. Before the experiment, all participants signed the informed consent approved by the California Institute of Technology's Institutional Review Board under protocol 19-0914. All participants were reimbursed in monetary format for base pay and their performance bonus. There are a total of 179 participants (105 females) with usable data after exclusions, who have a mean age of 30.3743 years (sd = 10.1886). As for recruitment criteria, participants would not have any history of substance/alcohol use disorder, anxiety disorder (Obsessive-Compulsive Disorder, Body Dysmorphia Disorder, generalized anxiety, social anxiety/social phobia), and/or depressive disorders (dysthymia, major depression). Also, participants would not use any medications for any subclinical psychiatric disorder treatment.

Experimental Procedure

There were multiple stages to complete the entire experiment, and all participants were given a consent form before participating in the experiment and could only proceed to the next stage if they consented. After consenting to participate, the participants were first screened through an interview by a psychiatrist, in which participants' mental health conditions and medication usage were evaluated to decide whether they were eligible to participate as healthy controls or patients or ineligible to participate. We only analyzed the data of the healthy controls from this procedure. After the eligibility evaluation, participants were asked to complete a few questionnaires on psychiatric measures before coming to the Brain Imaging Center at the California Institute of Technology to conduct the functional magnetic resonance imaging study.

Firstly, upon arrival at the Brain Imaging Center, participants were asked to complete a short questionnaire asking about feelings and emotions. Afterwards, before the fMRI scan, participants read through the instructions for the spaceship task. After the instructions, a few questions were asked to ensure participants understood the task and remembered the key features of the task. After the instructions, 115 trials of practice rounds of the task were done before participants performed the main task in the scanner. The spaceship task was paired with another task, studying habits, and the participants completed both tasks during the scan. The MRI scans were split into two half-sessions where participants completed one task per session; completing the spaceship task in the first or the second half was counterbalanced across participants. During the structural scan in the scanner (either in the first or the second half session), participants also completed another session (105 trials) of the spaceship task right before the functional session (154 trials). The behaviors of the structural scan and the functional scan were combined for the behavioral analysis as well as computational modeling. Specifically, both sessions were used for behavioral clustering analysis and for finding the best parameters of the computational models fitted to each individual, which could offer us more data to generate a reliable model fit to participants' behaviors. After completing the two tasks in the scanner, participants were debriefed about the task and received their reimbursement.

Space Miner Task

For the main task, participants were instructed to collect rewards through space mining on different planets. The background is that mining has begun on two planets (i.e., one identified as red and the other green) in space, and the goal is to earn as many points as possible by mining gems from the two planets. However, the mines on the two planets have changing conditions in their production. Sometimes gems could be found, but other times, worthless rocks could also be mined out. Specifically, there are two landing pads for the corresponding mines on each planet, one to the North and the other to the South. The landing pads are identified through their unique scenic view and are located in the upper (North) and bottom (South) parts of the planet on the screen.

Participants can choose between two different spaceships (identified as yellow and blue, with the screen locations of the spaceships fixed across trials) using a button-press to travel to the two planets for a miner. They are instructed that the yellow spaceship usually lands on the red planet, and the blue spaceship usually lands on the green planet. However, space travel can sometimes be a bit unpredictable due to space debris, so in some rare situations, the yellow ship will be forced to land on the green planet, and the blue spaceship will be forced to land on the red planet. Participants were instructed to use two buttons to choose the corresponding spaceships. In a given trial, once participants chose one spaceship, they observed the spaceship being highlighted and taken off and the planet appearing after the spaceship landed. Afterwards, the landing pad for the mine would also appear, either on the upper or the bottom part of the planet, given if they landed at the North or the South mine on the planet. Once the spaceship landed, the mine production appeared as either a gem with its price or a worthless rock. Specifically, participants were instructed that the gem's price is unpredictable and will change daily. Also, participants had no control over which mine (North or South) the spaceship would eventually land on. In general, the conditions at all four mines (two mines on each of the two planets) would change throughout the game regarding gem vs. stone production. Participants were encouraged to learn which mine produces gems the most reliably.

After going through the instructions of the spaceship task, participants were asked a few questions:

1. How many planets are there?
2. How many mines are there on each planet?
3. Which planet does the yellow ship usually land on?
4. Which planet does the blue ship usually land on?
5. How many points is a mined rock worth? (0 points vs. 1-100 points)

6. How many points is a gem worth? (0 points vs. 1-100 points)

Participants proceeded to the practice trials and then the main experiment if they answered all the questions correctly.

Task Design

The current variant of the two-step task shared a similar task structure as the original two-step task (Daw et al., 2011) overall but with some differences in details. The general structure consists of the first-step transition and reward probability shifts, and also three condition manipulations (i.e., reward magnitude, state transition, and reward contingency) were built into the space miner task. One key difference is that in the current task, there was only one action to be made, which was at the initial state, and the rest of the trial were all state transitions with no further actions needed. All three cohorts of participants experienced the same general structure of the space miner task. On each trial, the yellow and blue spaceships would appear on the left and right sides of the screen. If chosen, the transition via the yellow spaceship towards the red planet occurred with a probability of 0.7 and with a probability of 0.3 towards the green planet; corresponding probabilities were flipped for the transition via the blue spaceship towards the planets. After the transition from the planet stage to the landing-pad stage, reward or non-reward outcomes would appear. The underlying reward probability was shared across the two landing pads on a given planet (reward probability would also be associated with the spaceship chosen, depending on the reward contingency condition described later in this section), and there were specific periods built-in such that landing on one planet was more rewarding than landing on the other planet, and also periods where landing on either of the two planets was relatively comparable in terms of the reward probability. To implement this, the reward probability associated with two planets started from 1 vs. 0, and then by using the Sigmoid function, the reward probability of the rewarding planet decayed from 1 towards 0.3 (asymptote) within the time span of from 20 to 25 trials (the exact number of trials depended on whether a rare spaceship-planet transition is made); for the reward probability associated with the currently non-rewarding planet, the same Sigmoid decay rate and the flipped sign of decay slope was used to have the reward probability drift from 0 to 0.5 (asymptote) again within the time span of from 20 to 25 trials. Afterwards, the reward probabilities associated with the two planets were reset to 1 vs. 0, but the rewarding planet was reversed compared to the previous block of trials. This drifting of reward probabilities and the reversal of rewarding planet/landing pads occurred throughout the entire session of the experiments. The order of which planet was firstly rewarding was counterbalanced across the participants. The structure of going from strong preference (large gap of reward probabilities: 1 vs. 0) towards almost indifference (reward probabilities: 0.3 vs. 0.5) was first to facilitate learning of the rewarding option and then setting the value of two options towards indifference to prepare for learning after the preference reversal. Also, a critical trial of rare transition was built into the very beginning after a preference reversal to facilitate detecting stay/switch behaviors as a signature of the MB vs. MF strategy. It is worth noting that another manipulation of reward contingency (described later in this section) would change the contingency of reward delivery upon the landing pads (or planets) vs. the spaceships, yet the general fluctuating and reversal dynamics of the reward probability shift would remain the same across the two reward contingency conditions.

Reward Magnitude Manipulation

On top of the reward probabilities shifts throughout the main task, the reward magnitudes (points associated with the gem), if there was a reward, were manipulated to shift between two conditions: low reward prediction error (low RPE) vs. high reward prediction error (high RPE). For the reward magnitude manipulation, the magnitudes were drawn from a uniform distribution of (0.1, 0.19) for the low RPE condition and (0.3, 1) for the high RPE condition, for which the actual points were scaled by 100. Low vs. high RPE conditions were shifted every 26-27 trials, and the order of the low vs. high RPE conditions was counterbalanced across participants.

State-Transition Uncertainty Manipulation

For the second-step transition (with no actions required) after landing on a given planet, the landing pad would be shown subsequently to illustrate whether the North or South mine was landed on. A state-transition manipulation was built-in at this phase of the trial to shift the uncertainty of landing on a specific mine (North or South) given a specific planet: high state prediction error (high SPE) vs. low state prediction error (low SPE). During high SPE conditions, the transition to one of the two landing pads on a given planet was at the equal probability of 0.5 vs. 0.5, which elicited large state-transition uncertainties. On the other hand, during low SPE conditions, the transition to one of the two landing pads on a given planet was at the biased probability of 0.9 vs. 0.1, and whether the transition to the North or South landing pad was biased would change along the course of the task. Low vs. high SPE conditions were shifted every 26-27 trials, and the order was counterbalanced across participants. As participants would mostly see one type of transition within a conditional block, the state prediction was relatively certain under low SPE conditions.

Reward Contingency Manipulation

Besides manipulating reward magnitude and state-transition uncertainty, reward delivery was also manipulated to be contingent upon 1) what stimulus (i.e., spaceship) participants choose or 2) what terminal states (i.e., landing pads) were arrived at, meaning rewards are 1) stimulus-contingent or 2) state-contingent, respectively. The stimulus-contingent and state-contingent conditions were run by two separate fMRI blocks of the experiment, and the two fMRI blocks had the same number of trials (77 trials each, 154 trials in total), and the order of the two contingency conditions was also counterbalanced across participants. During the stimulus-contingent reward condition, the reward probability structure (described earlier in this section) was associated with the stimulus (i.e., spaceship) chosen by the participant in each trial, regardless of what second-stage state (i.e., planet) or terminal state (i.e., landing pad) that was reached. During the state-contingent reward condition, the reward probability structure (described earlier in this section) was associated with the pair of landing pads on a given planet, so both the choice at the first stage and the actual terminal states that were arrived at would potentially influence the final outcome. Intuitively, under the state-contingent reward condition, in response to a win or a loss after a rare spaceship-planet transition in the previous trial, the first-stage transition probability representation (as used in a model-based strategy) would direct the subsequent choice towards the stimulus more likely leading to the rewarding terminal states; in contrast, during the stimulus-contingent reward condition in the similar situation, such a model-based strategy would end up choosing the option that was less likely to deliver a reward, effectively eliciting more model-based RPEs.

Within each reward-contingency block (i.e., the state-contingent and stimulus-contingent blocks), both reward magnitude and state-transition uncertainty were manipulated simultaneously along with the reversals of reward probability (either associated with the state or the stimulus). We ensured all the manipulated computational variables were orthogonal to each other, so the averaged overall effect of one conditional variable was solely independent rather than confounded by other manipulated variables. To realize the orthogonality between conditions, the starting trial index of each condition was offset to one another (i.e., with a lag of around 13 trials) so that no two conditional blocks (e.g., High RPE and Low SPE) would completely overlap, with the aim of obtaining the sole conditional effect by averaging across one type of conditional blocks and canceling out the correlations between manipulations.

Behavioral Clustering

There are, in general, two types of behavioral metrics we leveraged to perform cluster allocation on participants' behavior: 1) choice measure: choice pattern as a function of the outcome and transition type in the preceding trial, and 2) reaction time measure: reaction time pattern as a function of current choice, the outcome and transition type in the preceding trial. Specifically within the choice measure, summary statistics and regression estimates of each participant served as the features for the clustering. For the summary statistics, we calculate the following metrics for each participant:

1. $p(Stay_t|Common_{t-1}Rewarded_{t-1})$
2. $p(Stay_t|Common_{t-1}Unrewarded_{t-1})$
3. $p(Stay_t|Rare_{t-1}Rewarded_{t-1})$
4. $p(Stay_t|Rare_{t-1}Unrewarded_{t-1})$

5. Reward Sensitivity:
$$[p(Stay_t|Common_{t-1}Rewarded_{t-1}) + p(Stay_t|Rare_{t-1}Rewarded_{t-1})]$$
$$-[p(Stay_t|Common_{t-1}Unrewarded_{t-1}) + p(Stay_t|Rare_{t-1}Unrewarded_{t-1})]$$

6. Transition Sensitivity:
$$[p(Stay_t|Common_{t-1}Rewarded_{t-1}) - p(Stay_t|Rare_{t-1}Rewarded_{t-1})]$$
$$-[p(Stay_t|Common_{t-1}Unrewarded_{t-1}) - p(Stay_t|Rare_{t-1}Unrewarded_{t-1})]$$

7. Absolute Reward Sensitivity:
$$|p(Stay_t|Common_{t-1}Rewarded_{t-1}) - p(Stay_t|Rare_{t-1}Rewarded_{t-1})|$$
$$-|p(Stay_t|Common_{t-1}Unrewarded_{t-1}) - p(Stay_t|Rare_{t-1}Unrewarded_{t-1})|$$

For the regression estimates of the choice measure, we used a mixed-effect logistic regression to model the choice repeating probability as a function of the intercept, outcome, transition type, and their interaction in the preceding trial, which serve as the fixed effects. The dependent variable "isStay" denotes whether choosing the same option as in the previous trial, and all regressors are dummy variables identifying the status in the previous trial: "isWin" denotes the outcome in the previous trial; "isRare" denotes the spaceship-planet transition type in the previous trial. Three binary variables indicating the conditional manipulations are added as three independent interactions with $isWin \times isRare$. Specifically, $HighRPE^{mf}_{t-1}$ is a binary variable indicating the previous trial's reward magnitude condition (1: high reward magnitude/high MF-RPE, 0: low reward magnitude/low MF-RPE). Similarly, $LowSPE^{mb}_{t-1}$ indicates the previous trial's state-transition uncertainty condition (1: low state-transition uncertainty/low SPE, 0: high state-transition uncertainty/high SPE), and $LowRPE^{mb}_{t-1}$ indicates the previous trial's reward contingency condition (1: state-contingent reward/low MB-RPE, 0: stimulus-contingent reward/ high MB-RPE). For the random effects, the intercept and the slope of the outcome, transition type, and their interaction in the preceding trial are modeled to vary at the single-subject level. And "subID" denotes the identity of each participant. The full model is:

$$isStay \sim isWin{\times}isRare{\times}\left(HighRPE^{mf}_{t-1} + LowSPE^{mb}_{t-1} + LowRPE^{mb}_{t-1}\right)+ \\ (1 + isWin \times isRare|subID)\,,$$

where "×" denotes the main effects and the interaction between each independent variable, and "1" denotes the intercept, that is, the average stay probability for each subject. Trials, where the choice was not made within 2 seconds, were excluded before estimating the regression model. From the regression, we extracted the random effects of the intercept, the outcome, the transition type, and the outcome-transition interaction from each individual to use as features in the clustering algorithm.

As for the reaction time measure for the clustering, we also used a mixed-effect logistic regression to characterize how individuals' reaction times (RT) varied as a function of the current choice (stay or switch), the preceding trial's outcome (win vs. no-win), the preceding trial's transition (common vs. rare), and all the two-way and three-way interactions between these regressors. The same were modeled as random effects for each individual.

$$log(RT_t) \sim Stay_t \times Reward_{t-1} \times isRare_{t-1} + (1 + Stay_t \times Reward_{t-1} \times isRare_{t-1} | subID)\,, \tag{3.1}$$

where we log-transformed the RT, and "×" denotes the main effects and the interaction between each independent variable; and "1" denotes the intercept. Trials, where the choice was not made within 2 seconds, were excluded before estimating the regression model. We also extracted each individual's random effects of the intercept, the outcome, the transition type, and the outcome-transition interaction for their clustering features.

In total, from both the choice measures (summary statistics and regression estimates) and the reaction time measures (regression estimates), there are 19 behavioral features for each participant. We relied on the centroids discovered in an external dataset with a much larger sample size (Cockburn et al., n.d.) to perform cluster allocation for this dataset. In the external dataset, a much larger group of participants (N=678) completed the same two-step task online, and the k-means clustering algorithm was used on the same 19 behavioral features to classify all the participants, and the clusters' centroids were obtained. The optimal number of clusters was found as four using the Gap statistical method (Tibshirani, Walther, and Hastie, 2001) in the external dataset, which gave sensible classification to separate behaviors generally as RL strategies vs. non-RL strategies as well as gave finer-grained strategy characterization for participants using RL strategy - 1) Mixture of MF and MB control, 2) pure MF control, and 3) pure MB control. In the current dataset, we assign each individual to the group label of the individual's closest centroid in terms of Euclidean distance. The labels were defined according to behavioral phenotypes in the external dataset: 1) Mixture Group, 2) MF Group, 3) MB Group, and 4) Other Group (non-RL strategies).

The Computational Model of Reinforcement Learning

The computational model we used to characterize the behavior was a hybrid reinforcement learning model composed of a model-free module and a model-based module. As the space miner task only requires action at the initial stage, which leads to two subsequent intermediate states before reaching the outcome, both the MF and MB modules only learn the action values of the two stimuli at the initial stage, which were guided through learned the downstream state value within each trial. As the position of the two stimuli at the initial stage was fixed, choosing the left (or right) action always led to the yellow (or blue) spaceship, and thus the action values learned were effectively the same as the stimulus values.

Model-Free Module

The model-free module is composed of a slow learning and a fast learning component, which corresponds to the learning over outcomes of multiple past trials ("slow") and learning upon the outcome of the preceding trial ("fast"). The slow learning component is described first.

As mentioned above, since there is only one action needed in the task and the action values at the initial state are equivalent to the stimulus values, the MF module essentially learns the value of potential states $Q_{MF}^{nth}(s)$ (n as the stage order in a sequence, s as the potential states within the nth stage). Hence, for example, the Q-value of the yellow spaceship could be denoted as $Q_{MF}^{1st}(yellow)$, and the Q-value of the red planet and the Q-value of the north pad on the red planet could be denoted as $Q_{MF}^{2nd}(red)$ and $Q_{MF}^{3rd}(\textit{red-north})$, respectively. The reward prediction error experienced at the nth stage could be denoted as RPE_{MF}^{nth} ($n = 2$ when at the planet stage, $n = 3$ when at the landing pad stage, and $n = 4$ when observing the outcome).

The reward prediction error at each stage and the learning rule for Q-values in previous stages (with eligible traces) were defined as the following (the subscripts of MF and the chosen/experienced state variable s for Q-values and RPE are not shown for simplicity):

$$RPE^{(n+1)th} = r_t^{(n+1)th} + Q_t^{(n+1)th} - Q_{t-1}^{nth}, n = 1, 2, 3, \tag{3.2}$$

$$Q_t^{nth} = Q_{t-1}^{nth} + \alpha_{MF} \times RPE^{(n+1)th}, n = 1, 2, 3, \tag{3.3}$$

$$Q_t^{(n-1)th} = Q_{t-1}^{(n-1)th} + \alpha_{MF} \times \lambda \times RPE^{(n+1)th}, n = 2, 3, \tag{3.4}$$

$$Q_t^{(n-2)th} = Q_{t-1}^{(n-2)th} + \alpha_{MF} \times \lambda^2 \times RPE^{(n+1)th}, n = 3, \tag{3.5}$$

where the $\alpha_{MF}(0 < \alpha_{MF} < 1)$ is the learning rate of the Q-values and shared across all stages, $n = 1, 2, 3$, within the trial; the trace decay parameter λ is set to be 1, meaning that an equal amount of value updates was conducted to a proximal and to a distal state. All-stage Q-values at the beginning of the session were initialized to 0 ($n = 1, 2, 3$). $r_t^{nth}(0 \leq r_t^{nth} \leq 1)$ denotes the magnitude of the outcome (scaled down by 100) and is only relevant when $n = 3$, which is when observing the actual outcome, but otherwise $r_t^{nth} = 0$ given no reward delivery in the intermediate states.

Note that $Q_t^{(n+1)th} = 0$ when $n = 3$, as there are no learnable states besides the outcome at that stage.

As for the fast learning component, a value bias $W(s_{t-1}, r_{t-1})$ was added to the learned Q-values of the first-stage chosen stimuli s_{t-1} in the preceding trial dependent on the outcome of the preceding trial $r_{t-1} \in \{0 : no\ reward, 1 : reward\}$. It essentially builds a stay or a switch action tendency bias into the associated stimulus based upon a preceding reward or no-reward event. The incorporation of the fast learning component into the overall utility of the chosen stimulus $U_{MF}^{1st}(s)$ is specified as follows:

$$U_t^{1st}(s) = Q_t^{1st}(s) + W(s_{t-1}, r_{t-1}), s = s_{t-1}, \tag{3.6}$$

$$W(s_{t-1}, r_{t-1}) = \begin{cases} W_{MF}^{stay} & \text{if } r_{t-1} = 1 \\ W_{MF}^{switch} & \text{if } r_{t-1} = 0 \end{cases}, s_{t-1} \in \{yellow, blue\}, \tag{3.7}$$

where W_{MF}^{stay} and W_{MF}^{switch} are the two parameters fitted for each participant, and both parameters could be positive or negative to capture all possible action adjustment policies.

Model-Based Module

The model-based module is a forward learner using the dynamic programming approach, and it also consists of a slow learning component that updates the Q-values on a multi-trial basis and a fast learning component that considers the trial experience in the most recent trial. The slow learning component is first described below.

As described for the model-free module, the Q-value learned in the MB module could be similarly denoted as $Q_{MB}^{nth}(s)$ (n as the stage order in a sequence, s as the potential states within the nth stage as listed for the MF module). To carry out dynamic programming to compute the Q-values of upper-level states, the reward information (binary) was essentially learned in the MB module through RPE_{MB}^{4th} at the outcome stage for the Q-value of the experienced landing pads Q_t^{3rd}:

$$RPE_{MB}^{4th} = r_t^{4th} - Q_{t-1}^{3rd}(s), s \in \{\textit{red-north}, \textit{red-south}, \textit{green-north}, \textit{green-south}\}, \tag{3.8}$$

$$Q_t^{3rd}(s) = Q_{t-1}^{3rd}(s) + \alpha_{MB} \times RPE^{(n+1)th}, s \in \{\textit{red-north}, \textit{red-south}, \textit{green-north}, \textit{green-south}\}, \tag{3.9}$$

where α_{MB} ($0<\alpha_{MB}<1$) is the learning rate for the MB module. r_t^{4th} is a binary variable of 1 when rewarded and 0 when unrewarded in the current trial.

Besides learning the binary reward information, the MB module determines the value of the terminal states $Q^{3rdMag}(s)$ by integrating the magnitude of the reward experienced $Mag(s)$ into the value component $Q^{3rd}(s)$ learned through the binary reward information:

$$Q_t^{3rdMag}(s) = Q_t^{3rd}(s) + Mag_t(s), s \in \{\textit{red-north}, \textit{red-south}, \textit{green-north}, \textit{green-south}\}. \tag{3.10}$$

The MB module utilizes two sets of state-transition probability to compute the upper-level Q-values from the learned Q-value of the landing pads. The first state-transition probability is hard-coded instead of through learning, as assumed by the fact that participants were instructed about the common vs. rare transition structure from the spaceship and the planet and had enough time to practice before the main experiment. The first state-transition probability $T_1(s, s')$, $s \in \{yellow, blue\}$, is specified as below:

$$T_1(yellow, s') = [0.7, 0.3], s' = \{red, green\}; \tag{3.11}$$

$$T_1(blue, s') = [0.3, 0.7], s' = \{red, green\}. \tag{3.12}$$

In contrast, the second state-transition probability $T_2(s, s')$ was learned through state prediction errors (SPE) experienced through the transition from state s to the subsequent s' (i.e., from a planet to either the north or the south landing pad on that planet). The second state-transition probability ($T_2(s, s')$) is updated as below:

$$SPE = 1 - T_2(s, s'), s \in \{red, green\}, s' \in \{north, south\}, \tag{3.13}$$

$$T_2(s, s') = T_2(s, s') + \eta \times SPE, s \in \{red, green\}, s' \in \{north, south\}, \tag{3.14}$$

$$T_2(s, s'') = (1 - \eta) \times T_2(s, s''), s \in \{red, green\}, s'' \in \{north, south\}, s'' \neq s', \tag{3.15}$$

where η is the state-transition learning rate and is a fixed value as it is not a recoverable parameter through model fitting, and thus set as a median value of 0.5. For potential but non-experienced states s'', the transition probability was scaled by (1- η) to normalize the transition probabilities associated with the state s.

With the Q-value of the terminal states after incorporating the learned magnitude component $Q_t^{3rdMag}(s)$, and the second state-transition probabilities, the Q-values of the planets $Q_{MB}^{2nd}$and spaceships Q_{MB}^{1st}could be computed through dynamic programming as below (the subscripts MB are not shown for simplicity):

$$Q_t^{2nd}(s) = \sum_{s'} T_2(s, s') \times Q_t^{3rdMag}(s'), s \in \{red, green\}; \tag{3.16}$$

$$Q_t^{1st}(s) = \sum_{s'} T_1(s, s') \times Q_t^{2nd}(s'), s \in \{yellow, blue\}, s' \in \{red, green\}. \quad (3.17)$$

In addition to the slow learning component, the MB module also integrates information from the preceding trial to rapidly adjust the action tendency of staying with the selected option versus switching to the other option, which is referred to as the fast learning component. Specifically, to facilitate the stay vs. switch action selection tendency, the fast learning component utilizes the task model, considering the outcome and the first-stage transition type in the preceding trial, to add either a positive or negative value bias $W(s_{t-1}, r_{t-1}, c_{t-1})$ to the option selected in the preceding trial, where $s_{t-1} \in \{yellow, blue\}$ denotes the selected spaceship, $r \in \{0 : \textit{no-reward}, 1 : reward\}$ denotes the outcome, and $c \in \{0 : rare, 1 : common\}$ denotes the transition type. The value integration of slow and fast learning into the overall utility $U_{MB}^{1st}(s)$ is specified as the following:

$$U_{MB}^{1st}(s) = Q_{t-1}^{1st}(s) + W_t(s_{t-1}, r_{t-1}, c_{t-1}), s = s_{t-1}, \quad (3.18)$$

$$W_t(s_{t-1}, r_{t-1}, c_{t-1}) = \begin{cases} W_{MB}^{stay} & \text{if } (r_{t-1}, c_{t-1}) = (1,1) \text{ or } (0,0) \\ W_{MB}^{switch} & \text{if } (r_{t-1}, c_{t-1}) = (1,0) \text{ or } (0,1) \end{cases}, s_t \in \{yellow, blue\}, \quad (3.19)$$

where W_{MB}^{stay} and W_{MB}^{switch} are the two parameters fitted for each participant, and both of the parameters are unbounded and could be either positive or negative to capture all possible action adjustment policies.

The Hybrid of MF and MB Module

With the learned first-stage stimulus utility from both the MF and MB modules, integration of the utilities from the two modules was implemented to obtain the mixed utility of first-stage stimulus, which was used for action selection. The learned MF stimulus utility $U_{MF}^{1st}(s)$ and MB stimulus utility $U_{MB}^{1st}(s)$ are combined through a weighting parameter wMF that decides the weight of utility from the MF module for combination and $(1 - wMF)$ is the weight of the MB stimulus utility. A semi-arbitration mechanism is introduced to calculate the value of wMF for trials within certain condition blocks. Considering the three manipulations of prediction error signal throughout the experiment, as hypothesized by the reliability-based arbitration theory (S. W. Lee, Shimojo, and O'doherty, 2014), we set three free parameters that correspond to the weight adjustment of the MF system in response to the three manipulations:

$$\Delta wMF_t^{mfRPE} = \begin{cases} a_{mfRPE} & \text{if low reward magnitude on trial t} \\ -a_{mfRPE} & \text{if high reward magnitude on trial t} \end{cases}, \tag{3.20}$$

$$\Delta wMF_t^{mbRPE} = \begin{cases} a_{mbRPE} & \text{if stimulus-contingent reward on trial t} \\ -a_{mbRPE} & \text{if state-contingent reward on trial t} \end{cases}, \tag{3.21}$$

$$\Delta wMF_t^{SPE} = \begin{cases} a_{SPE} & \text{if high state-transition uncertainty on trial t} \\ -a_{SPE} & \text{if low state-transition uncertainty on trial t} \end{cases}, \tag{3.22}$$

where a_{mfRPE}, a_{mbRPE}, and a_{SPE} are the three free parameters. Based upon what the manipulated reliability condition a given trial t is in, the weight adjustments are integrated into the baseline weight of the MF module:

$$wMF_t^{raw} = wMF_t^{baseline} + \Delta wMF_t^{mfRPE} + \Delta wMF_t^{mbRPE} + \Delta wMF_t^{SPE}. \tag{3.23}$$

The adjusted raw MF weight is then passed into the sigmoid function to transform the raw weight onto the scale of (0,1):

$$wMF_t = \frac{1}{1 + e^{-wMF_t^{raw}}}. \tag{3.24}$$

The weight assigned to the MB module on trial t would then be $1 - wMF_t$. Hence the mixture of the MF stimulus utility and the MB stimulus utility is achieved by:

$$U_t^{Mix}(s) = wMF_t \times U_t^{MF}(s) + (1 - wMF_t) \times U_t^{MB}(s), s \in \{yellow, blue\}. \tag{3.25}$$

Once the mixed stimulus utility is computed, the probability of choosing a specific stimulus s is then calculated through a softmax function:

$$P_t(s) = \frac{e^{\beta \times U_t^{Mix}(s)}}{e^{\beta \times U_t^{Mix}(s)} + e^{\beta \times U_t^{Mix}(s')}}, \quad s, s' \in \{yellow, blue\}, s' \neq s. \tag{3.26}$$

With the calculated choice probability for each participant, we fit the specified parameters to maximize the summed negative log-likelihood of participants' choices across all trials (with no-response trials excluded):

$$negLLE = -\sum_t \log(P_t(s)), \quad s \in \{yellow, blue\}, \tag{3.27}$$

where s denotes the participant's actual choice on trial t.

For each individual, the model parameters were fitted with a Bayesian inference method using the cbm (computational and behavioral modeling) toolbox (Piray et al., 2019) with the non-hierarchical specification. Each parameter was fit using a normally distributed prior with a mean of zero and a variance of 6.25 that ensured the cover of a large range of parameters with no excessive model complexity penalty.

Functional MRI Acquisition

The fMRI data were acquired at the Caltech Brain Imaging Center (Pasadena, CA) using a Siemens Prisma 3T scanner with a 32-channel radiofrequency coil. The functional scans were conducted using a multi-band echo-planar imaging (EPI) sequence with 72 slices, -30 degrees slice tilt from AC-PC line, 192 mm × 192 mm field of view, 2 mm isotropic resolution, repetition time (TR) of 1.12 s, echo time (TE) of 30ms, multi-band acceleration of 4, 54-degree flip angle, in-plane acceleration factor 2, echo spacing of 0.56 ms, and EPI factor of 96. Following each run, both positive and negative polarity EPI-based field maps were collected using similar parameters to the functional sequence but with a single band, TR of 5.13 s, TE of 41.40 ms, and 90-degree flip angle. T1-weighted and T2-weighted structural images were also acquired for each participant with 0.9 mm isotropic resolution and 230 mm × 230 mm field of view. For the T1-weighted scan, TR of 2.55 s, TE of 1.63 ms, inversion time (TI) of 1.15 s, flip angle of 8 degrees, and in-plane acceleration factor 2 were used. The T2-weighted scan was acquired with TR of 3.2 s, TE of 564 ms, and in-plane acceleration factor of 2.

Functional MRI Data Preprocessing

Results included in this manuscript come from preprocessing performed using *fMRIPrep* 23.1.3(Esteban, Christopher J Markiewicz, et al., 2019; Esteban, Christopher J. Markiewicz, et al., 2023; RRID:SCR_016216), which is based on *Nipype* 1.8.6 (Gorgolewski et al., 2011; RRID:SCR_002502).

Anatomical data preprocessing

A total of 1 T1-weighted (T1w) image was found within the input BIDS dataset. The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection (Tustison et al., 2010, distributed with ANTs (Avants et al., 2008, RRID:SCR_004757), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a *Nipype* implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as the target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white matter (WM), and gray matter (GM) was performed on the brain-extracted T1w using fast (FSL, RRID:SCR_002823, Zhang, Brady, and Smith, 2001). Brain surfaces were reconstructed using recon-all (FreeSurfer 7.3.2, RRID:SCR_001847, Dale, Fischl, and Sereno, 1999), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (RRID:SCR_002438, Klein et al., 2017). Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs), using brain-extracted versions of both the T1w reference and the T1w template. The following template was selected for spatial normalization and accessed with *TemplateFlow* (23.0.0): *ICBM 152 Nonlinear Asymmetrical template version 2009c* (Fonov et al., 2009, RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym).

Functional data preprocessing

A *B0*-nonuniformity map (or *fieldmap*) was estimated based on two (or more) echo-planar imaging (EPI) references with topup (Andersson, Skare, and Ashburner, 2003; FSL). The estimated *fieldmap* was then aligned with rigid-registration to the target EPI (echo-planar imaging) reference run.

For each of the 6 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer) which implements boundary-based registration (Greve and Fischl, 2009). Co-registration was configured with six degrees of freedom. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL, Jenkinson et al., 2002). BOLD runs were slice-time corrected to 0.52s (0.5 of slice acquisition range 0s-1.04s) using 3dTshift from AFNI (Cox and Hyde, 1997, RRID:SCR_005927). The BOLD time-series were resampled onto the following surfaces (FreeSurfer reconstruction nomenclature): *fsaverage*. The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying a single, composite transform to correct for head-motion and susceptibility distortions. These resampled BOLD time-series will be referred to as *preprocessed BOLD in original space*, or just *preprocessed BOLD*. The BOLD time-series were resampled into standard space, generating a *preprocessed BOLD run in MNI152NLin2009cAsym space*. Several confounding time-series were calculated based on the *preprocessed BOLD*: framewise displacement (FD), DVARS, and three region-wise global signals. FD was computed using two formulations following Power (absolute sum of relative motions, Power et al., 2014) and Jenkinson (relative root mean square displacement between affines, Jenkinson et al., 2002). FD and DVARS are calculated for each functional run, both using their implementations in *Nipype* (following the definitions by Power et al., 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors was extracted to allow for component-based noise correction (*CompCor*, Behzadi et al., 2007). Principal components are estimated after high-pass filtering the *preprocessed BOLD* time-series (using a discrete cosine filter with 128s cut-off) for the two *CompCor* variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 2% variable voxels within the brain mask. For aCompCor, three probabilistic masks (CSF, WM, and combined CSF+WM) are generated in anatomical space. The implementation differs from that of Behzadi et al.(2007) in that instead of eroding the masks by 2 pixels on BOLD space, a mask of pixels that likely contains a volume fraction of GM is subtracted from the aCompCor masks. This mask is obtained by dilating a GM mask extracted from the FreeSurfer's *aseg* segmentation, and it ensures components are not extracted from voxels containing a minimal fraction of GM. Finally, these masks are resampled into BOLD space and binarized by thresholding at 0.99 (as in the original implementation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the *k* components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite et al., 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardized DVARS were annotated as motion outliers. Additional nuisance time-series are calculated by means of principal components analysis of the signal found within a thin band (*crown*) of voxels around the edge of the brain, as proposed by Patriat, Reynolds, and Birn (2017). All resamplings can be performed with *a single interpolation step* by composing all the pertinent transformations (i.e., head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos, 1964). Non-gridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer).

Many internal operations of *fMRIPrep* use *Nilearn* 0.10.1 [Abraham et al., 2014, RRID:SCR_001362], mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in *fMRIPrep*'s documentation (`https://fmriprep.org/en/latest/workflows.html`).

Functional MRI Data Analysis

The SPM12 package was used for the GLM analysis on the fMRI data (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK). The fMRI data were slice-timing corrected and were applied with a high-pass filter of 180 seconds to remove low-frequency drifts potentially caused by physiological and physical noise. The fMRI data were corrected for motion, warped to the standard Montreal Neurological Institute (MNI) template, and smoothed with a Gaussian kernel (8mm FWHM) to mitigate individual anatomical differences.

We set up a general linear model (GLM) to perform voxel-wise statistical modeling on the BOLD activity. The two blocks of the fMRI data were concatenated into one longer sequence, and there are, in total, seven event-related regressors with their associated parametric modulators in the GLM. The event-related regressors are modeled as a stick function with zero duration and are followed by the z-scored parametric modulators if they have any, as below:

1. Fixation Onset;
2. Stimulus Onset: 1) MF Chosen Utility, 2) MF Rejected Utility, 3) MB Chosen Utility, 4) MB Rejected Utility;
3. Left Response Onset;
4. Right Response Onset;
5. Planet-Pad Onset: 1) State Prediction Error (Note: The event onset regressor is a combined regressor from the two events of interest, and the SPE signals at the planet and pad stage are combined into one parametric modulator);
6. Planet-Pad-Outcome Onset: 1) MF Reward Prediction Error, 2) MB Reward Prediction Error (Note: The event onset regressor is a combined regressor from the onsets of the three events, and both MF RPEs and MB RPEs at the three stages are combined into one parametric modulator);
7. Outcome Onset: 1) Reward Magnitude.

To control for motion and non-neuronal fMRI signals, we include the following nuisance regressors: six rigid-body motion regressors (three translations and three rotations), framewise displacement (FD; quantification of the estimated bulk-head motion) (Power et al., 2012), the averaged signal within cerebrospinal fluid (CSF) mask, the average signal within the white matter mask, the average signal within both the cerebrospinal fluid and white matter mask, and the average signal within the whole brain mask and the derivative of the average signal within the whole brain mask. Besides, the scan volumes with an FD larger than the threshold of 0.77mm were set as regressors of motion spikes. The threshold of 0.77mm is determined by first calculating the FD threshold of 1.5 times the interquartile range plus the third quartile of each participant's FD across all scans, then performing the calculation of 1.5 times the interquartile range plus the third quartile of the entire FD threshold distribution across all individuals.

To obtain the statistical beta maps of decision utility from the MF system, we computed at the contrast level the beta of the MF chosen utility minus that of the MF rejected utility. It is computed the same way for the contrast of the MB decision utility. To assess how decision utility and reward prediction error signals from the MF and MB system are represented by groups using varying degrees of the MF/MB strategy, we extracted the corresponding first-level beta estimates from the corresponding behavioral group and performed second-level random effects modeling across individual betas to test significance with one-sample t-test. To determine whether there is any significantly different encoding of the variable of interest (i.e., decision utility and reward prediction error) between the MB group and MF group, a two-sample t-test between the MF group and MB group was conducted at the whole-brain level on the first-level contrasts of the variables of interest.

Regions of Interest and Small Volume Correction

For the small volume correction conducted for the contrasts of MF RPE and MF RPE > MB RPE in Figure 3.2a, the ROI was defined as the bilateral 10mm spheres centered on the coordinates of (6,2,10) and (-6,2,10), where the coordinates of (-6,2,10) are peak coordinates of an outcome prediction error contrast from the work of Jessup & O'Doherty (2011), and we mirrored this coordinate to the other hemisphere to get the coordinates of (6,2,10).

The small volume correction for the contrasts of MB RPE is a hand-drawn ventral striatum mask that is composed of bilateral nucleus accumbens and bilateral ventral putamen (parts of posterior ventral putamen were not included to create correspondence to rodents literature — posterior putamen is analogous to dorsolateral striatum in rodents). The region labels here were identified based on the Harvard-Oxford Subcortical Structural Atlas (RRID: SCR_001476).

The ROI analysis on vmPFC in Figure 3.3b & Figure 3.3c is a vmPFC mask published from a meta-analysis on neural correlates of subjective value (Bartra, McGuire, and Kable, 2013).

Statistical Analysis

Logistic Regression

For all the mixed-effect logistic regression analyses, "fitglme" function in Matlab was used. To quantify how the previous reward, previous transition type, and their interactions affected the stay choices, we used a mixed-effect logistic regression to model the probability of staying with the option chosen in the previous trial ($isStay_t$) as a function of the previous outcome ($Reward_{t-1}$, 1: reward, 0: no-reward), the previous transition type ($Rare_{t-1}$, 1:rare, 0:common) and their interactions. The fixed effects include the previous outcome, the previous transition type, and their interactions. The intercept and the slope of the previous outcome, the previous transition type, and their interaction were modeled as random effects that could vary at the single-subject level (indicated by $subID$). The regression model is:

$$\begin{aligned}&isStay \sim isWin \times isRare + \\ &(1 + isWin \times isRare|subID)\,,\end{aligned}$$

where "×" denotes the main effects and the interaction between each independent variable, and "1" denotes the intercept, which is the average stay probability for each subject. Trials, where the choice was not made within 2 seconds, were excluded before estimating the regression model.

Exclusion Criteria

Participants for the main analysis were excluded based on the proportion of fMRI volumes that exceed a manually specified motion threshold to eliminate potential neural confounds due to too much motion (for threshold selection, see the section of Functional MRI Data Analysis). Participants who had over 15% of the total volumes whose framewise displacement (FD) exceeded the threshold of 0.77mm in both of the two fMRI runs were excluded. In the fMRI analysis, we also excluded any runs with over 15% volumes that exceeded the threshold of 0.77mm.

References

Abraham, Alexandre et al. (2014). “Machine learning for neuroimaging with scikit-learn”. In: *Frontiers in neuroinformatics* 8, p. 71792.

Andersson, Jesper LR, Stefan Skare, and John Ashburner (2003). “How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging”. In: *Neuroimage* 20.2, pp. 870–888.

Avants, Brian B et al. (2008). “Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain”. In: *Medical image analysis* 12.1, pp. 26–41.

Balleine, Bernard W and Anthony Dickinson (1998). “Goal-directed instrumental action: contingency and incentive learning and their cortical substrates”. In: *Neuropharmacology* 37.4-5, pp. 407–419.

Balleine, Bernard W and John P O’doherty (2010). “Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action”. In: *Neuropsychopharmacology* 35.1, pp. 48–69.

Bartra, Oscar, Joseph T McGuire, and Joseph W Kable (2013). “The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value”. In: *Neuroimage* 76, pp. 412–427.

Behzadi, Yashar et al. (2007). “A component based noise correction method (CompCor) for BOLD and perfusion based fMRI”. In: *Neuroimage* 37.1, pp. 90–101.

Beierholm, Ulrik R et al. (2011). “Separate encoding of model-based and model-free valuations in the human brain”. In: *Neuroimage* 58.3, pp. 955–962.

Berns, Gregory S et al. (2001). “Predictability modulates human brain response to reward”. In: *Journal of neuroscience* 21.8, pp. 2793–2798.

Cockburn, Jeffrey et al. (n.d.). “Characterizing heterogeneity in human reinforcement learning and the arbitration of behavioral control”. In: ().

Cox, Robert W and James S Hyde (1997). “Software tools for analysis and visualization of fMRI data”. In: *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo* 10.4-5, pp. 171–178.

Dale, Anders M, Bruce Fischl, and Martin I Sereno (1999). “Cortical surface-based analysis: I. Segmentation and surface reconstruction”. In: *Neuroimage* 9.2, pp. 179–194.

Daw, Nathaniel D et al. (2011). “Model-based influences on humans’ choices and striatal prediction errors”. In: *Neuron* 69.6, pp. 1204–1215.

Dezfouli, Amir and Bernard W Balleine (2013). “Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized”. In: *PLoS computational biology* 9.12, e1003364.

Dezfouli, Amir, Nura W Lingawi, and Bernard W Balleine (2014). “Habits as action sequences: hierarchical action control and changes in outcome value”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1655, p. 20130482.

Dickinson, Anthony and Bernard Balleine (2002). “The role of learning in the operation of motivational systems”. In: *Stevens’ handbook of experimental psychology* 3, pp. 497–533.

Esteban, Oscar, Christopher J Markiewicz, et al. (2019). “fMRIPrep: a robust preprocessing pipeline for functional MRI”. In: *Nature methods* 16.1, pp. 111–116.

Esteban, Oscar, Christopher J. Markiewicz, et al. (June 2023). *fMRIPrep: a robust preprocessing pipeline for functional MRI*. Version 23.1.3. DOI: 10.5281/zenodo.8076450. URL: https://doi.org/10.5281/zenodo.8076450.

Feher da Silva, Carolina and Todd A Hare (2020). “Humans primarily use model-based inference in the two-stage task”. In: *Nature Human Behaviour* 4.10, pp. 1053–1066.

Feher da Silva, Carolina, Gaia Lombardi, et al. (2023). “Rethinking model-based and model-free influences on mental effort and striatal prediction errors”. In: *Nature Human Behaviour* 7.6, pp. 956–969.

Fonov, Vladimir S et al. (2009). “Unbiased nonlinear average age-appropriate brain templates from birth to adulthood”. In: *NeuroImage* 47, S102.

Gläscher, Jan et al. (2010). “States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning”. In: *Neuron* 66.4, pp. 585–595.

Gorgolewski, Krzysztof et al. (2011). “Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python”. In: *Frontiers in neuroinformatics* 5, p. 12318.

Greve, Douglas N and Bruce Fischl (2009). “Accurate and robust brain image alignment using boundary-based registration”. In: *Neuroimage* 48.1, pp. 63–72.

Hampton, Alan N, Peter Bossaerts, and John P O’doherty (2006). “The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans”. In: *Journal of Neuroscience* 26.32, pp. 8360–8367.

Houk, James C, James L Adams, and Andrew G Barto (1994). “A model of how the basal ganglia generate and use neural signals that predict reinforcement”. In.

Iigaya, Kiyohito et al. (2019). “Deviation from the matching law reflects an optimal strategy involving learning over multiple timescales”. In: *Nature communications* 10.1, p. 1466.

Jenkinson, Mark et al. (2002). “Improved optimization for the robust and accurate linear registration and motion correction of brain images”. In: *Neuroimage* 17.2, pp. 825–841.

Jessup, Ryan K and John P O’Doherty (2011). “Human dorsal striatal activity during choice discriminates reinforcement learning behavior from the gambler’s fallacy”. In: *Journal of Neuroscience* 31.17, pp. 6296–6304.

Kahneman, Daniel, Shane Frederick, et al. (2002). “Representativeness revisited: Attribute substitution in intuitive judgment”. In: *Heuristics and biases: The psychology of intuitive judgment* 49.49-81, p. 74.

Killcross, Simon and Pam Blundell (2002). “Associative representations of emotionally significant outcomes”. In: *Emotional cognition*, p. 13.

Klein, Arno et al. (2017). “Mindboggling morphometry of human brains”. In: *PLoS computational biology* 13.2, e1005350.

Konovalov, Arkady and Ian Krajbich (2016). “Gaze data reveal distinct choice processes underlying model-based and model-free reinforcement learning”. In: *Nature communications* 7.1, p. 12438.

Kool, Wouter, Fiery A Cushman, and Samuel J Gershman (2016). “When does model-based control pay off?” In: *PLoS computational biology* 12.8, e1005090.

Lanczos, Cornelius (1964). “Evaluation of noisy data”. In: *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis* 1.1, pp. 76–85.

Lee, Sang Wan, Shinsuke Shimojo, and John P O’doherty (2014). “Neural computations underlying arbitration between model-based and model-free learning”. In: *Neuron* 81.3, pp. 687–699.

Loewenstein, George and Ted O’Donoghue (2004). “Animal spirits: Affective and deliberative processes in economic behavior”. In: *Available at SSRN 539843*.

McClure, Samuel M, Gregory S Berns, and P Read Montague (2003). “Temporal prediction errors in a passive learning task activate human striatum”. In: *Neuron* 38.2, pp. 339–346.

Montague, P Read, Peter Dayan, and Terrence J Sejnowski (1996). “A framework for mesencephalic dopamine systems based on predictive Hebbian learning”. In: *Journal of neuroscience* 16.5, pp. 1936–1947.

O’Doherty, John P et al. (2003). “Temporal difference models and reward-related learning in the human brain”. In: *Neuron* 38.2, pp. 329–337.

Otto, A Ross et al. (2013). “Working-memory capacity protects model-based learning from stress”. In: *Proceedings of the National Academy of Sciences* 110.52, pp. 20941–20946.

Pagnoni, Giuseppe et al. (2002). “Activity in human ventral striatum locked to errors of reward prediction”. In: *Nature neuroscience* 5.2, pp. 97–98.

Patriat, Rémi, Richard C Reynolds, and Rasmus M Birn (2017). “An improved model of motion-related signal changes in fMRI”. In: *Neuroimage* 144, pp. 74–82.

Piray, Payam et al. (2019). “Hierarchical Bayesian inference for concurrent model fitting and comparison for group studies”. In: *PLoS computational biology* 15.6, e1007043.

Power, Jonathan D et al. (2014). “Methods to detect, characterize, and remove motion artifact in resting state fMRI”. In: *Neuroimage* 84, pp. 320–341.

Rangel, Antonio, Colin Camerer, and P Read Montague (2008). “A framework for studying the neurobiology of value-based decision making”. In: *Nature reviews neuroscience* 9.7, pp. 545–556.

Satterthwaite, Theodore D et al. (2013). “An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data”. In: *Neuroimage* 64, pp. 240–256.

Schultz, Wolfram, Peter Dayan, and P Read Montague (1997). “A neural substrate of prediction and reward”. In: *Science* 275.5306, pp. 1593–1599.

Sloman, Steven A (1996). “The empirical case for two systems of reasoning.” In: *Psychological bulletin* 119.1, p. 3.

Smittenaar, Peter et al. (2013). “Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans”. In: *Neuron* 80.4, pp. 914–919.

Sutton, Richard S (2018). “Reinforcement learning: An introduction”. In: *A Bradford Book*.

Tibshirani, Robert, Guenther Walther, and Trevor Hastie (2001). “Estimating the number of clusters in a data set via the gap statistic”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2, pp. 411–423.

Tustison, Nicholas J et al. (2010). “N4ITK: improved N3 bias correction”. In: *IEEE transactions on medical imaging* 29.6, pp. 1310–1320.

Wunderlich, Klaus, Peter Smittenaar, and Raymond J Dolan (2012). “Dopamine enhances model-based over model-free choice behavior”. In: *Neuron* 75.3, pp. 418–424.

Zhang, Yongyue, Michael Brady, and Stephen Smith (2001). “Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm”. In: *IEEE transactions on medical imaging* 20.1, pp. 45–57.

Chapter 4

THE ENHANCED ENCODING OF MODEL-BASED REWARD PREDICTION ERROR IN ANTERIOR CINGULATE CORTEX AND THE MODEL-FREE BEHAVIORAL BIAS AMONG INDIVIDUALS WITH HIGH OBSESSIVE-COMPULSIVE TENDENCY

ABSTRACT

Obsessive-compulsive behaviors are repetitive, unhelpful behaviors that are difficult to inhibit despite the behaviors' adverse consequences. It has been suggested that a psychological model of imbalance between the habitual (model-free) and goal-directed (model-based) control could characterize obsessive-compulsive behaviors: people with high obsessive-compulsive tendencies express bias towards model-free behavior. However, it is unclear what specific computations go wrong across the dual controllers that cause the observed behavioral abnormality. In the current study, by leveraging individual differences in reward prediction error encoding and self-reports of obsessive-compulsive tendency, we demonstrate that a neurocomputational model of the anterior cingulate cortex (ACC) performing error/conflict monitoring could explain the abnormal imbalance between the dual controllers within the theoretical framework of reliability-based arbitration. The over-reliance on the habitual (model-free) control in highly obsessive-compulsive individuals is found to be caused by the enhanced encoding of the reward prediction error (RPE) signal from the model-based system in ACC, as the exaggeratedly encoded RPE could over-signal the unreliability of the model-based system, biasing the arbitrator to rely more on the model-free system to guide behavior. Our finding suggests it is the computation of the reliability signal in the arbitration process that causes the dual-control abnormality, and ACC could be one of the potential neural targets for symptom alleviation in obsessive-compulsive disorder.

4.1 Introduction

Compulsive behaviors, as manifested in Obsessive-Compulsive Disorder (OCD), are featured as conducting repetitive but pointless behaviors with the difficulty of controlling them (Robbins et al., 2012). A dual-controller framework capturing the learning process (Balleine and O'doherty, 2010) offers insights into how compulsive behavior might arise. The framework consists of the inflexible and history-dependent habitual controller, which associates action with a stimulus upon reward history, and the flexible, outcome-driven goal-directed controller, which plans and executes actions to achieve desirable outcomes. It has been experimentally shown that OCD patients display impairments in goal-directed behavioral control and instead increased habit expression in a task with the outcome-devaluation procedure (Gillan, Papmeyer, et al., 2011), consistent with the nature of inflexible and senseless behavior in OCD.

A computational framework in reinforcement learning (RL) postulated a pair of algorithms, the model-free (MF) and the model-based (MB), to describe the psychological pair of habitual and goal-directed control. In essence, the MF system learns "cached" values of actions through rewards and errors in history with no explicit encoding of the task environment; in contrast, the MB system represents the task environment by learning the state-transition model and computes the action values by deploying the model of the environment. The relationship between compulsiveness tendency and deficits in goal-directed control holds with the computational characterization of the MF and MB systems. With the two-step task to evaluate the participants' tendency to express MB control and with self-report questionnaires to assess OCD levels, it is shown that having a strong psychological trait of "compulsiveness" is associated with having deficits in MB control (Gillan, Kosinski, et al., 2016).

The higher propensity of expressing habitual control (or MF control) in the OCD population has not been understood well in terms of its computational mechanism. The computational framework of using two independent RL algorithms with the two-step task to dissociate each of their role in the expressed behavior offers an opportunity to further elucidate the mechanistic cause of reduced goal-directed (MB) control in RL terms. An uncertainty-based control allocation theory has been proposed (Daw, Niv, and Dayan, 2005), and a form of the theory's RL realization, reliability-based arbitration, was shown to act as a good computational model of arbitration in the two-step task, and the key computational variables were found to be represented neurally in a model-based fMRI approach (Lee, Shimojo, and O'doherty, 2014). Given the imbalance of MB control and MF control in the populations with high obsessive-compulsive tendencies, it is worth testing the abnormality of what components within the reliability-based arbitration process might cause the observed impairment in MB control. Specifically, the reliability-based arbitration framework has the MF controller learning values through reward prediction errors (RPEs), where the level of RPEs indicates the "reliability" of the MF system, and smaller RPEs indicate higher reliability; on the other hand, the MB controller learns the state-transition model of the task through state prediction errors, which serve as the reliability of the MB system, and smaller state prediction errors indicate good reliability. In addition, another source of learning signal, model-based reward prediction error (MB RPE), was shown to serve as another source of reliability of the MB system as well (Cockburn et al., n.d.). An arbitrator, neurally found to be located in the inferior frontal gyrus (Lee, Shimojo, and O'doherty, 2014; Kim et al., 2024), monitors these multiple forms of reliability signals from either the MF or the MB system and allocates weight to the more reliable controller. In the case of a more manifested MF control in OCD patients, there has been behavioral evidence that OCD patients experienced larger MB uncertainty signals than healthy controls, causing the arbitrator to engage the MB controller to a lesser extent, a potential computational cause of the control imbalance (Kim et al., 2024).

Besides a potential computational source, it is helpful to unravel the potential neural causes of the observed control imbalance in people scoring high on obsessive-compulsive symptoms. As the neural arbitrator (i.e., inferior frontal gyrus) represents the controller reliability given the prediction error signals from the two controllers, it modulates the value region MF system via negative coupling to gate the MF controller (Lee, Shimojo, and O'doherty, 2014). Consequently, it could be the case that the connectivity between the arbitrator region and the MF value region is attenuated in OCD patients compared to the healthy controls, which was indeed the case experimentally (Kim et al., 2024). However, it is an unanswered question whether the upstream neural process in monitoring the reliability signals in the uncertainty-based competition is distinct in the highly obsessive-compulsive group. It is thus a necessary task to unravel the underlying neurocomputational mechanism by probing the neural encoding of multiple forms of reliability signals in high vs. low OC tendency: 1) MF reward prediction error (MF RPE), 2) state prediction error (SPE), and 3) MB reward prediction error (MB RPE). The hypothesis is that the neural encoding of MB reliability signals in people with high obsessive-compulsive tendencies would be stronger than that of people with low obsessive-compulsive tendencies, as the stronger neural signature of uncertainty in the MB control would lead the arbitrator to put more weight on the MF control. Conversely but otherwise similarly, the neural encoding of MF reliability signals could be weaker in people with higher obsessive-compulsive tendencies.

As for the brain regions tracking the reliability signals, the literature has shown the striatum encodes the reward prediction error signal (O'Doherty et al., 2003; McClure, Berns, and Montague, 2003; Jessup and O'Doherty, 2011), and the intraparietal sulcus and the lateral prefrontal cortex encodes the SPE signal (Gläscher et al., 2010); thus these regions would be of interest in testing whether their representations of prediction error signals differ as a function of obsessive-compulsive levels. Importantly, as obsessive-compulsive behavior can be understood as the repetitive behavioral correction to mitigate the mismatch between the expected state and the actual perceived state upon an executed action, the neural region that supports the conflict/error detection would be of specific interest to examine for a relationship between conflict/error processing and the downstream behavioral control (Pitman, 1987; Fitzgerald and Taylor, 2015). Since the anterior cingulate cortex (ACC) has been found to serve as an error/conflict monitor, it is probable the neural information in ACC consequently leads to the expression of obsessive-compulsive behavior. Specifically, the anterior cingulate cortex (ACC) has been found to play a critical role in monitoring the conflict between habitual and goal-directed control (Watson, Wingen, and Wit, 2018). The abnormality in ACC's activation (Del Casale et al., 2011) and in its connectivity with other regions among OCD patients further indicate ACC's important role in maintaining an adaptive instrumental control. Hence, it is a natural question whether a distinct neural profile of ACC, when monitoring the error signals from MF and MB controllers, directs OCD patients towards the inflexible habitual (MF) control more than the healthy controls.

With participants (N=238, including controls and patients) self-reporting their obsessive-compulsive scores (OCI-R, Foa et al., 2002) and performing the two-step task while scanned by functional magnetic resonance imaging (fMRI), we could use the model-based fMRI approach to establish how different neural regions support the tracking of reliability signals differently as a function of various degrees of obsessive-compulsive tendencies. The use of the two-step task provides a scope to validate the relationship between the degree of MF vs. MB control and the level of OCD and, more importantly, to extract the controller-specific reliability signals to test the neurocomputational hypotheses. Specifically, we would use a semi-arbitration reinforcement learning model (Cockburn et al., n.d., see Methods) to characterize participants' MF (or MB) tendency and derive various prediction error signals from their corresponding MF/MB controller. Our main interest would be to investigate whether the individual obsessive-compulsive level measured through self-reports would correlate with the encoding strength of prediction error signals across individuals in the region of interest that we just described, which from the neurocomputational point of view could explain the overreliance on habitual (MF) system of high OC-tendency people through their high susceptibility to the "error" signal (the inverse of reliability) experienced through the dual controllers.

4.2 Results

The Reinforcement Learning Strategy across Various Levels of Obsessive-Compulsive Tendency

First, to ensure participants with all levels of OC tendency comprehended the two-step task and engaged in effective learning of the reward dynamics, we ran 1) a statistical test between the task performance of all participants and that of a random agent and 2) a correlational analysis between the participant's task performance, as indicated by the proportion of rewarded trials among all trials, and the self-report measure of OCI-R collected before performing the task. A good learning of the task across various levels of OCI-R score would be indicated by a significantly better performance than a random agent and by no correlation between task performance and the OCI-R measure. Indeed, we observed that the performance of all participants (reward rate, the average proportion of rewarded trials: 47.16%) was significantly better than a random agent (reward rate: 42.96%; $p < 0.001$), and we did not find a significant correlation between task performance and the OCI-R score (Spearman correlation $r = 0.0258, p = 0.6925$, Figure 4.1a), suggesting good and comparable task learning across all participants with varying levels of obsessive-compulsive tendency. The result holds when further decomposing the OCI-R score into the "obsessiveness" and "compulsiveness" components from the overall OCI-R questionnaire, as the participant's task performance was not either correlated with the "obsessiveness" component ($r = -0.0078, p = 0.9050$) or with the "compulsiveness" component ($r = -0.0305, p = 0.6392$). One note is that given this variant of the two-step task was designed in such a way that the MF strategy and the MB strategy would be equally optimal in reward earning, no correlation between task performance and OCI-R score does not imply the degrees of MF strategy engagement are comparable across all levels of obsessive-compulsive tendencies, which was further investigated below. Establishing a comparable level of task engagement across all levels of OCI-R score is critical as this precludes the possibility that any behavioral results and difference in neural profiles across different levels of obsessive-compulsive tendency are due to differences in meta-task factors such as task comprehension and learning efficacy, which might introduce interpretation difficulty in any correlation observed between the OC tendency and the measured behavioral variables.

To evaluate the change of behavioral strategy, specifically, the reinforcement learning strategy of MF and MB control, used by participants across multiple levels of OCI-R score, we conducted the correlational analysis between the OCI-R score and the degree of MF control across all participants. As for the degree of MF control, we extract the weighting parameter (i.e., wMF) indicative of the degree of MF control usage from a hybrid reinforcement learning model, where wMF was used to proportionally combine the option utility learned in parallel from the MF module and the MB module (see Methods for model details). We found a significant positive correlation between the OCI-R score and the degree of MF control ($r = 0.1708, p = 0.0083$, Figure 4.1b). Again, if further decomposing the OCI-R score into the "obsessiveness" and "compulsiveness" components, the degree of MF control was correlated with both the "obsessiveness" component ($r = 0.1385, p = 0.0327$) and the "compulsiveness" component ($r = 0.1594, p = 0.0138$). Overall, the results are consistent with the previous work on the impairment of behavioral control in OCD (Gillan, Papmeyer, et al., 2011; Gillan, Kosinski, et al., 2016), suggesting an impaired MB control and potentially goal-directed control in people with high obsessive-compulsive tendencies.

To explore if the reliability-based arbitration theory could account for the overweight of the MF system in high OC-tendency participants, we ran a few post-hoc correlational analyses between the OC-severity (OCI-R score and its "obsessiveness" & "compulsiveness" sub-components) and the computational variables of absolute prediction error signals derived from the hybrid reinforcement learning model: the absolute MF RPE signal, the absolute MB RPE signal, and the SPE signal (the SPE signal is always positive) from the MB system. All prediction errors experienced at different stages were concatenated and averaged over the entire experiment for each participant before the correlational analysis was conducted. Overall, we did not find any significant correlation between the OC-tendency and the experienced absolute

MB RPE signal (Raw OCI-R, Spearman $r = -0.0987, p = 0.1290$; obsessiveness, Spearman $r = -0.0842, p = 0.1954$; compulsiveness,$r = -0.1026, p = 0.1143$), or any significant correlation between the OC-tendency and the experienced SPE signal (Raw OCI-R, Spearman $r = -0.0857, p = 0.1878$; obsessiveness, Spearman $r = -0.0851, p = 0.1910$; compulsiveness,$r = -0.0719, p = 0.2695$). Interestingly, we found a trending effect of negative correlation between the OC-tendency, specifically the sub-component of compulsiveness, and the experienced absolute MF RPE (Raw OCI-R, Spearman $r = -0.1243, p = 0.0556$; obsessiveness, Spearman $r = -0.0708, p = 0.2768$; compulsiveness,$r = -0.1255, p = 0.531$). This suggests that the lower level of the absolute MF RPEs experienced by the high OC-tendency participants could signal the higher reliability of the MF system, whereby the participants' behaviors are guided with a higher weight.

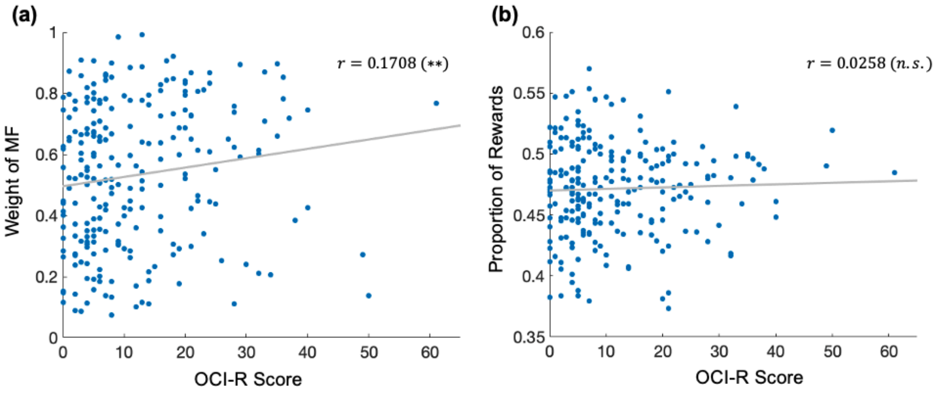


Figure 4.1: The correlations between the measured OC tendency (OCI-R score) and behavioral measures of task performance and MF weight estimated from the model (a) The correlation between the OCI-R score and the task performance: average proportion of rewarded trials. (b) The correlation between the OCI-R score and MF weight (wMF) that estimated from each participant through the hybrid reinforcement learning model. The MB weight ($1 - wMF$) would be perfectly anti-correlated with the MF weight in the current model.

Obsessive-compulsive tendency and the neural encoding of prediction errors

After establishing the relationship that participants with high obsessive-compulsive tendencies tended to exhibit a higher degree of MF control, we sought to examine how such behavioral correlation was manifested through a neurocomputational mechanism. Through the lens of a reliability-based arbitration framework, we approached the behavioral variation of MF tendency as a consequence of adapting behavioral controls according to the various levels of the perceived prediction error signals encoded by the brain. Specifically, according to the reliability-based theory, if the prediction error signals, a proxy of the system reliability, from the MB system were found to be neurally encoded more strongly in people with high obsessive-compulsive tendency, the perceived low reliability of the MB system would drive people's behavior towards the MF end of the spectrum. We thus used the model-based fMRI approach to examine how prediction error signals were neurally encoded across individuals with various levels of the OCI-R score.

First, to estimate the neural encoding of the prediction error signal from each individual, we extracted the prediction error signal from the computational model fitted to each individual. As the reliability measure of the MB system, MB reward prediction errors (MB RPEs) and state prediction errors (SPEs) were extracted, whereas MF reward prediction errors (MF RPEs) were extracted for the reliability measure of the MF system. All three extracted prediction error signals were built into an event-related design matrix to model the modulation of the prediction errors on each individual's fMRI signal (see Methods for details). Since three RPE signals from the MF and MB system arise on each trial, the derived three RPE variables are set as the parametric modulators at their corresponding time points - planet onset, pad onset, and outcome onset. Importantly, to increase the statistical power of capturing the RPE-specific variance in the BOLD signal, we combine the three event regressors into one chained regressor "planet-pad-outcome onset," which entails three stick functions are parametrically modulated by both the MF RPE and MB RPE at each stage. Consequently, using the combined event regressor for the RPE signal would have three times more observation points than setting three separate RPE regressors at each stage, yet not differentiate RPE encoding at different trial stages. Similarly, as the state prediction error signals arise at two stages within a trial (i.e., planet onset and pad onset), a combined event regressor "planet-pad" parametrically modulated by the SPE signals was built into the design matrix.

The first-level general linear model (GLM) was fitted to each individual's fMRI data to estimate the encoding strength (i.e., beta maps) of the MF RPE, MB RPE, and SPE at the whole brain level. Once the beta coefficients of the three prediction error signals were obtained, at the second-level group analysis, each individual's OCI-R score was entered as a covariate in the model to estimate how the OCI-R score correlates with the encoding strength of the three prediction error signals across individuals. To ensure the specificity of any observed correlation comes from the obsessive-compulsive disorder per se, we also included the self-report score on the measure of anxiety (STAI-Trait,Spielberger, 1983) and depression (BDI-II, Beck, Steer, and Brown, 1996), as the covariates at the second-level group analysis. Interestingly, we found that the encoding strength of MB RPE signal in the anterior cingulate cortex (ACC) was significantly correlated with the OCI-R score across individuals (p = 0.046, cluster-level FWE, peak voxel (-12,50,16), Figure 4.2), suggesting that the higher the OCI-R score an individual reports, the stronger encoding of MB RPE signal in ACC in that individual. As for the hypothesized correlations between the OCI-R score and the encoding of the MF RPE signal (i.e., negative correlations) or the SPE signal (i.e., positive correlations), no significant clusters survived FWE corrections across the brain.

The significant correlation between MB RPE encoding in ACC and the individual's OCI-R score is consistent with the reliability-based arbitration framework to account for the observed high MF expression among people with high obsessive-compulsive tendencies in this two-step task. Among people with higher obsessive-compulsive tendencies, the perceived strong MB RPE signal indicates a less reliable MB system during the task, hence the arbitrator driving behavior to rely more on the MF system. Importantly, in spite of the significant correlations between obsessive-compulsive tendency and other psychiatric traits such as anxiety (Spearman correlation, r = 0.5438, $p < 0.001$) and depression (Spearman correlation, r = 0.4706, $p < 0.001$), our results are robust to such confounds and the observed correlation of prediction error encoding strength and the OCI-R score comes above and beyond the variance that could be accounted for by the anxiety and depression measures.

After establishing the overall correlation between the OCI-R measure and the encoding strength of the MB RPE signal in ACC, we conducted a post-hoc decomposition of the OCI-R measure (21 items) into the components of the obsessiveness (3 items) and compulsiveness (18 items) to further interrogate the source of the observed correlation. In the second-level group analysis, both scores associated with the obsessiveness and compulsiveness components were entered as the covariates. Not too surprisingly, it was found that it was the trait of compulsiveness within the OCI-R measure that was correlated with the encoding strength of the MB RPE signal in ACC (p = 0.026, cluster-level FWE, peak voxel (-12,50,16), SVC, see Methods), which depicts the specific psychiatric trait that is influenced through the prediction errors in the neural arbitration mechanism.

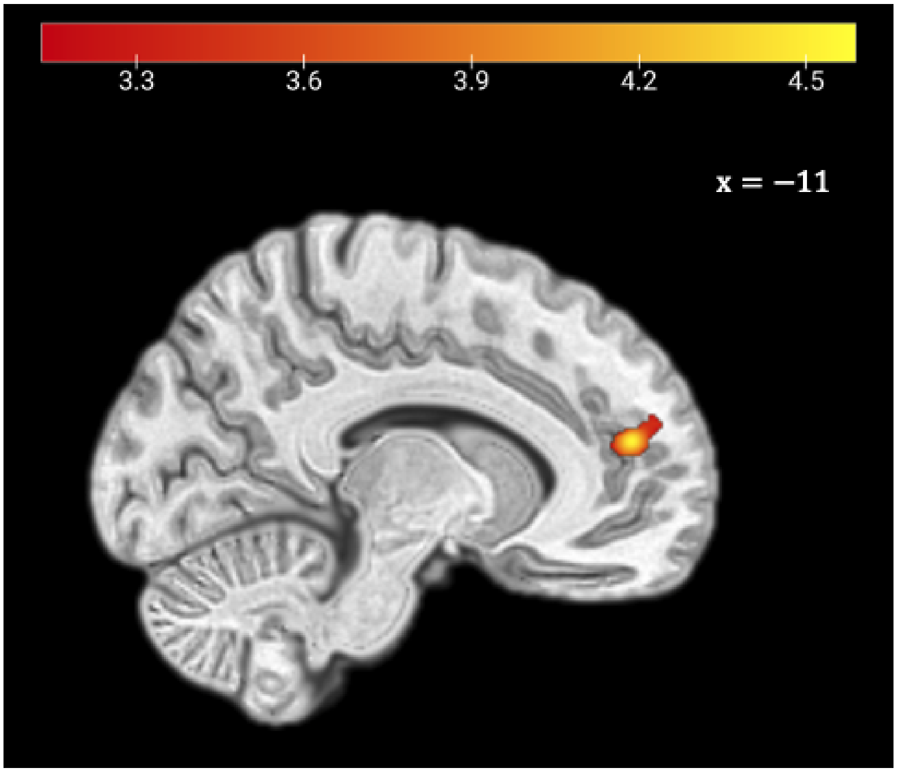


Figure 4.2: The T-map of the cluster showing the correlational effect in ACC between neural measures of encoding strength of MB RPE and the obsessive-compulsive tendency (OCI-R score). The cluster was formed with a threshold of p<0.001, uncorrected, k=169, peak voxel (-12, 50, 16), T=4.68; k denotes the number of voxels in the cluster.

4.3 Discussion

By scanning a large number of participants while they were doing the two-step task and by taking their self-reported scores of the OCI-R questionnaire, we were able to have enough across-individual variance in the behavioral measures to study its relationship with the variation in participants' obsessive-compulsive tendency as well as to investigate potential neural sources that potentially facilitate such behavioral relationship across individuals. Specifically, we first examined whether the degree of obsessive-compulsive tendency in our participants has any relationship with the performance in the two-step task or whether it has any correlation with the MF/MB tendency expressed throughout the task measured through a reinforcement learning model. Given no discrimination against any specific reinforcement learning strategy in this version of the two-step task, the result that there was no correlation between the OC tendency and task performance provides evidence that there are no adverse effects of high OC tendency on task comprehension or task engagement. This is an important fact for the validity of all correlational arguments made on OC tendency using the current task, as it suggests any observed behavioral or neural correlations related to OC tendency could not be attributed to meta-task factors that could potentially be indirectly caused by the psychiatric symptoms. Also, although OCD patients suffer constant subjective feelings of doubt and unsatisfactory behavior, they exhibit unimpaired performance in cognitive tasks (Galderisi et al., 1995). Thus, to study real-world obsessive-compulsive behavior from an ecologically valid point of view, it is critical to have a decision-making task that facilitates the investigation of psychological models of obsessiveness and compulsiveness while maintaining the capacity to enable unimpaired task performance even in people with high OC tendencies as in the real world. The current version of the two-step task seems to serve well in this regard.

Using a computational model to characterize participants' behavior in the two-step task, we found a significant positive correlation between the participant's MF tendency in the task and the self-reported OCI-R score. In detail, both the "obsessiveness" and "compulsiveness" components in the OCI-R score were found to be positively correlated with the MF tendency, which provides better specificity in explaining the sub-component of obsessive-compulsive behavior through the reinforcement learning model. The current finding is consistent with previous findings that the OCD populations tend to have impaired goal-directed control (approximated by the MB algorithm) and rely excessively on a habitual system modeled through the MF system (Gillan, Papmeyer, et al., 2011; Sanne de Wit et al., 2012). Specifically, the positive correlation between both the degree of "obsessiveness" & "compulsiveness" sub-component and the MF behavioral tendency echoes the previous finding that the psychiatric dimension of "compulsive behavior and intrusive thoughts" relates to the deficits in goal-directed control with clinical specificity (Gillan, Kosinski, et al., 2016). From the perspective of having a correct psychological model of obsessive-compulsive behavior, it is important to learn whether OC behavior is 1) a product of enhanced habitual (MF) learning or 2) a consequence of compromised goal-directed (MB) learning. However, since we approached the behavior with a reinforcement learning hybrid model where the weight of MF and MB modules sum up to one, it is not feasible to differentiate whether the increase in the degree of MF (or MB) tendency (i.e., the weight of the MF system, wMF) in high OC-tendency participants was due to the enhanced learning of stimulus-action association through the MF system or the impairment of action-outcome association through the MB system, and therefore it is hard to draw evidence for either psychological account. It needs further modeling work, potentially by orthogonalizing the model weights, to dissociate the independent contribution of the MF and MB modules to the overall behaviors.

Although we could not rely on the current behavioral model to differentiate whether it is the abnormality from the MF or from the MB system that causes the overreliance on the MF system among high OC-tendency participants, the neural measures indeed gave us the means to investigate whether the arbitration process between the dual systems might underlie the observed behavioral shifts across individuals. With the model-based fMRI approach, we found a significant positive correlation between the severity of obsessive-compulsive tendency, specifically the component of "compulsiveness," and the encoding strength of the MB reward prediction errors in anterior cingulate cortex across all participants (both controls and patients). Such result supports the account that the bias towards MF behavior in high OC tendency participants could be attributed to an abnormal reliability-based arbitration process that the reliability signal of the MB system, MB RPE, was exaggeratedly represented in ACC so that the MB system was signaled as unreliable according to the arbitrator and the MF system was instead relied on. Here we want to make a further clarification on the interpretation of this neural finding with regards to the distinction of this arbitration account at the algorithmic level vs. at the neural implementation level (Marr, 2010). Considering specifically the role of MB RPE in the reliability-based arbitration framework, a high degree of expressed MF behavior in high OC-tendency participants could be 1) driven by higher-level of absolute MB RPEs experienced in the algorithmic MB computations throughout the experiment on average, or 2) driven by a relay of information from an upstream neural region (e.g., ACC) that exaggeratedly encodes the algorithmic MB RPE signal to a downstream region (i.e., arbitrator) that takes this input to decide the system weight for behavioral output accordingly, or 3) both the first and the second scenario. Our current neural finding potentially supports the second scenario, a mechanism at the level of neural implementation, but did not directly speak to the first scenario at the algorithmic level. The post-hoc correlational analysis between OC-tendency and the MB RPE provided some tentative evidence. The result of no significant correlation suggests it is not the algorithmic MB RPEs experienced by the participants but the relay of exaggeratedly represented neural information of MB RPEs in ACC that elicits the MF bias in behavior in participants with higher levels of OC tendency.

Besides the role of MB RPE, the role of MF RPE and SPE were also examined at the algorithmic level (in a post-hoc way) and at the neural implementation level in terms of driving the behavior to be more MF in high OC-tendency participants. Our results suggest that the bias towards the MF behavior could also be partially attributed to the lower MF RPE signal experienced by the high OC-tendency participants, implicated by the trending negative correlation between absolute MF RPE and the OC-tendency. Yet there was a lack of evidence indicating that the MF RPE signal was disproportionally represented in the striatal regions in high OC-tendency participants. Also, we did not find any SPE-related effects at either the algorithmic level or the neural implementation level (in cortical regions of interest; Gläscher et al., 2010) on mediating the system balance across individuals with varying OC tendencies. As the correlational behavioral analysis regarding the experienced prediction error signal is, in essence, post-hoc, cautions are needed when interpreting the reported algorithmic cause of the increased usage of the MF system in high OC-tendency participants. More tailored investigations are needed to pinpoint how the prediction errors as the reliability proxy would account for the overreliance on the MF system at the algorithmic level. One potentially informative analysis would be to measure the shift of the degree of MF (or MB) behavior when the levels of prediction error signal are manipulated to be high vs. low and to examine how the degree of the measured shift varies as a function of one's OC tendency. The participants with higher OC tendencies should have a higher degree shift to engage the MF system when the MF RPE is deemed low or when the MB RPE is deemed high compared to those with less severe OC tendencies. This would potentially also be a behavioral corroboration of the neural finding on the correlation between MB RPE encoding and OC tendency in the current study.

One potential confound in the correlational analysis between the OCI-R score and the encoding strength of MB RPE comes from the "comorbidity" at the trait level (Gillan, Kosinski, et al., 2016), specifically from the correlation between the obsessive-compulsive measure and the depression-related or anxiety-related psychiatric measures, bringing into question whether the observed correlation was specific to the obsessive-compulsive tendency. In our group-level fMRI analysis, both the STAI-Trait and the BDI-II scores of the participants were entered as the covariates along with the OCI-R score; we could argue the enhanced MB RPE encoding was specific to participants with high OC tendency rather than influenced by levels of anxiety or depression, although we could not rule out the possibility that variance from other psychiatric symptoms besides OC, anxiety, and depression measures could contribute to the observed variation of MB RPE encoding strength, which requires more strictly-controlled study to narrow down the contributing psychiatric sources.

In this chapter, we leveraged the behavior of a large cohort of participants with both controls and patients to investigate the relationship between obsessive-compulsive tendency and the tendency of engaging habitual learning (approximated through model-free reinforcement learning algorithm) as well as the potential neurocomputational mechanism that gave rise to such a behavioral relationship. The significant correlation between OC tendency and degree of MF usage is consistent with previous findings on the overreliance on habitual control in OCD populations (Gillan, Papmeyer, et al., 2011) and justifies the computational characterization of OCD using reinforcement learning algorithms. Critically, we found evidence of the disproportionally encoded prediction error signal in ACC among high OC-tendency people that could explain their behavioral MF bias through the reliability-based arbitration framework. Lastly, with the current insights and potential future findings through the neurocomputational approach, we can study OCD better in terms of its mechanistic causes in the brain, and more importantly, clinical interventions on relevant neural targets could be developed accordingly to better treat OCD.

4.4 Methods

Participants

For participant recruitment, we recruited participants who reside in the United States and who are fluent English speakers and readers. The age range for the studies is from 18 to 65 years. Before the experiment, all participants signed the informed consent approved by the California Institute of Technology's Institutional Review Board under protocol 19-0914. All participants were reimbursed in monetary form for base pay and their performance bonus. There are a total of 238 (158 females) participants with usable data after exclusions, who have a mean age of 30.0084 years (sd = 10.0694). As for recruitment criteria of healthy controls, participants would not have any history of substance/alcohol use disorder, anxiety disorder (Obsessive-Compulsive Disorder, Body Dysmorphia Disorder, generalized anxiety, social anxiety/social phobia), and/or depressive disorders (dysthymia, major depression). Also, participants would not use any medications for any subclinical psychiatric disorder treatment.

Experimental Procedure

See the same section in Chapter 3.

Space Miner Task

See the same section in Chapter 3.

Task Design

See the same section in Chapter 3.

The Computational Model of Reinforcement Learning

See the same section in Chapter 3.

Functional MRI Acquisition

See the same section in Chapter 3.

Functional MRI Data Preprocessing

See the same section in Chapter 3.

Functional MRI Data Analysis

For the design matrix of the 1st-level GLM, see the same section in Chapter 3.

To assess how the encoding strength of MB reward prediction error (RPE) signal varied across individuals with various levels of OCI-R score, we took the 1st-level contrast of MB RPE estimated from the 1st-level GLM as the dependent variable and ran the 2nd-level multiple regression with the OCI-R score (or its sub-components of obsessiveness or compulsiveness), along with the STAI-T and BDI-II scores as covariates. The coefficient estimates of the OCI-R score in the regression were tested for statistical significance.

Regions of Interest and Small Volume Correction

For the small volume correction on the fMRI group-level correlations between the betas of the MB RPE cluster in the cingulate cortex and the sub-components of the OCI-R score (i.e., obsessiveness and compulsiveness), the ROI used is a binarized mask of the paracingulate gyrus extracted from the Harvard-Oxford Cortical Structural Atlas (RRID: SCR_001476).

Exclusion Criteria

See the same section in Chapter 3.

References

Balleine, Bernard W and John P O'doherty (2010). "Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action". In: *Neuropsychopharmacology* 35.1, pp. 48–69.

Beck, Aaron T, Robert A Steer, and Gregory Brown (1996). "Beck depression inventory–II". In: *Psychological assessment*.

Cockburn, Jeffrey et al. (n.d.). "Characterizing heterogeneity in human reinforcement learning and the arbitration of behavioral control". In: ().

Daw, Nathaniel D, Yael Niv, and Peter Dayan (2005). "Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control". In: *Nature neuroscience* 8.12, pp. 1704–1711.

Del Casale, Antonio et al. (2011). "Functional neuroimaging in obsessive-compulsive disorder". In: *Neuropsychobiology* 64.2, pp. 61–85.

Fitzgerald, Kate D and Stephan F Taylor (2015). "Error-processing abnormalities in pediatric anxiety and obsessive compulsive disorders". In: *CNS spectrums* 20.4, pp. 346–354.

Foa, Edna B et al. (2002). "The Obsessive-Compulsive Inventory: development and validation of a short version." In: *Psychological assessment* 14.4, p. 485.

Galderisi, Silvana et al. (1995). "Neuropsychological slowness in obsessive–compulsive patients: is it confined to tests involving the fronto-subcortical systems?" In: *The British Journal of Psychiatry* 167.3, pp. 394–398.

Gillan, Claire M, Michal Kosinski, et al. (2016). "Characterizing a psychiatric symptom dimension related to deficits in goal-directed control". In: *elife* 5, e11305.

Gillan, Claire M, Martina Papmeyer, et al. (2011). "Disruption in the balance between goal-directed behavior and habit learning in obsessive-compulsive disorder". In: *American Journal of Psychiatry* 168.7, pp. 718–726.

Gläscher, Jan et al. (2010). "States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning". In: *Neuron* 66.4, pp. 585–595.

Jessup, Ryan K and John P O'Doherty (2011). "Human dorsal striatal activity during choice discriminates reinforcement learning behavior from the gambler's fallacy". In: *Journal of Neuroscience* 31.17, pp. 6296–6304.

Kim, Taekwan et al. (2024). "Neurocomputational model of compulsivity: deviating from an uncertain goal-directed system". In: *Brain* 147.6, pp. 2230–2244.

Lee, Sang Wan, Shinsuke Shimojo, and John P O'doherty (2014). "Neural computations underlying arbitration between model-based and model-free learning". In: *Neuron* 81.3, pp. 687–699.

Marr, David (2010). *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.

McClure, Samuel M, Gregory S Berns, and P Read Montague (2003). “Temporal prediction errors in a passive learning task activate human striatum”. In: *Neuron* 38.2, pp. 339–346.

O’Doherty, John P et al. (2003). “Temporal difference models and reward-related learning in the human brain”. In: *Neuron* 38.2, pp. 329–337.

Pitman, Roger K (1987). “A cybernetic model of obsessive-compulsive psychopathology”. In: *Comprehensive psychiatry* 28.4, pp. 334–343.

Robbins, Trevor W et al. (2012). “Neurocognitive endophenotypes of impulsivity and compulsivity: towards dimensional psychiatry”. In: *Trends in cognitive sciences* 16.1, pp. 81–91.

Spielberger, CD (1983). “State-trait anxiety inventory for adults”. In: *Mind Garden*.

Watson, P, G van Wingen, and S de Wit (2018). “Conflicted between goal-directed and habitual control, an fMRI investigation”. In: *eneuro* 5.4.

Wit, Sanne de et al. (2012). “Corticostriatal connectivity underlies individual differences in the balance between habitual and goal-directed action control”. In: *Journal of Neuroscience* 32.35, pp. 12066–12075.

Chapter 5

GENERAL DISCUSSION

When biological agents interact with the world, predictions about events and learning an accurate value of the events are critical from an evolutionary perspective, as predictions with accurate values of the real-world states could help biological agents accrue rewards and avoid risks effectively. The prediction errors could serve as the critical learning signal for the organisms to build these necessary predictions in appetitive and aversive domains. Through theoretical modeling work, the temporal-difference (TD) error was shown to be capable of facilitating value learning over an unconstrained temporal structure and explaining well the behaviors observed in both Pavlovian and instrumental conditioning (Schultz, Dayan, and Montague, 1997; Hollerman and Schultz, 1998; Fiorillo, Tobler, and Schultz, 2003; Waelti, Dickinson, and Schultz, 2001; Berns et al., 2001; O'Doherty et al., 2003).

To explain the learning behavior observed in the biological agents through a computational scope, specifically regarding the reflexive stimulus-response learning and the reflective action-outcome learning, the model-free (MF) and model-based (MB) reinforcement learning algorithms have been proposed to characterize these respective learning behaviors. Importantly, the TD error serves as the basic learning module to facilitate the operation of the MF and MB systems by establishing accurate reward predictions in the environment. Devising an appropriate task for the experimental study of the MF and MB systems is critical in that the task should be capable of soliciting the actual MF and MB strategies as expressed in the behavioral signatures when the task is engaged rather than alternative strategies that masquerade as MF, MB, or a mixed strategy. A Markov decision-making task called the two-step task (N. D. Daw, Gershman, et al., 2011) was developed for investigations of the MF and MB systems and the corresponding neural substrates. However, there has been some debate recently about whether the apparent MF and MB behaviors are indeed driven by the well-defined MF or MB computations or whether the strategy identified through the task behavior is actually the consequence of strategy misattribution (Feher da Silva and Hare, 2020; Feher da Silva, Lombardi, et al., 2023). Such debate could potentially cast doubts on the relevant findings regarding the MF and MB strategies and underlying neural systems in studies that used the two-step task (N. D. Daw, Gershman, et al., 2011; C. M. Gillan, Otto, et al., 2015; Wunderlich, Smittenaar, and Dolan, 2012; Smittenaar et al., 2013; Doll et al., 2015; F. Cushman and Morris, 2015). Specifically, the arguments on the limitation of the two-step task are that the mass strategy engaged in the two-step task is model-based and the detected MF components in the behavior could be due to using a wrong model on the task, and this thus negates the legitimacy of studying the MF strategy in the two-step task. Besides, such arguments were presented with neural results showing the absence of MF reward prediction error (TD error) signal among participants while engaging in the two-step task (Feher da Silva, Lombardi, et al., 2023).

In Chapter 3, through recruiting a large sample of participants, we found the neural evidence of reward-independent MF RPE encoded in the striatum as well as the MF decision utility signal in the ventromedial prefrontal cortex, which helped address some of the recent concerns about the validity of using the two-step task to study the MF system. In addition to the investigations on the existence of MF signal in the two-step task, our work in Chapter 3 also presented the evidence of the key signal of learning and decision-making for the MB system, with MB RPE encoded in the ventral striatum and MB decision utility encoded in the ventromedial prefrontal cortex. The co-existence of MF- and MB-related computational signals we found provides the support for using the two-step task to potentially study the behavioral and neural profile of both the MF and the MB systems, which could help obtain insights into the more fundamental stimulus-response and action-outcome learning mechanisms.

A validated two-step task for studying the MF and MB system does not imply that the evidence of MF and MB strategy engagement at the group level we observed consists of such evidence present at each individual level. This is cause for caution when drawing insights from the group average into the individual decision-making process; at the same time, the individual differences in strategy use as observed in the sub-groups are themselves of critical research interest (Charpentier et al., 2024), especially when these research insights are to be used as guidance for individually targeted therapies in psychiatry. Reinforcement learning strategies can largely fall into the category of MF and MB as we discussed, and previous literature has identified people's behavior in the two-step task as a mixture of the two (N. D. Daw, Gershman, et al., 2011; Kool, F. A. Cushman, and Gershman, 2016; Wunderlich, Smittenaar, and Dolan, 2012; Dezfouli and Balleine, 2013; Otto et al., 2013; Smittenaar et al., 2013; Dezfouli, Lingawi, and Balleine, 2014). Therefore, if considering the MF and MB strategies as the two ends of the RL spectrum, each individual's strategy when engaging in the task could lie anywhere in between, with the mixture sitting in the middle point. Of course, when other non-RL decision-making strategies are considered, such as win-stay-lose-switch or gambler's fallacy (Jarvik, 1951; Tversky and Kahneman, 1971), the kinds of possible strategies in use could increase exponentially. Any patterns of individual differences in strategy use might be viewed as insignificant when the number of participants is limited for any given kind of strategy, and hence it is difficult to draw any meaningful insights. Increasing the sample size would resolve such concerns effectively.

In Chapter 3, the behavioral and neural data from a relatively large number of participants provided a good handle on approaching the individual differences of strategy use in the two-step task. The overall group-average strategy was decomposed into four subgroups through cluster classification, which gives a finer-grained delineation of the degree of RL strategy use in the two-step task on the MF-MB spectrum and other potential non-RL strategies. The previously identified apparent group-level strategy of using a mixture of MF and MB algorithms could only describe around a quarter of the actual strategy type deployed by individuals in the participant pool. The apparent mixture strategy was actually made up of participants using pure MB, pure MF strategy, and other potential non-RL strategies (we did not dive into the non-RL group here). With the help of the individual difference approach, we again demonstrated the existence of an MF system engaged in the two-step task by identifying a behavioral group that engages a pure MF strategy. Given the rich profile of the individual-level strategies used in the current two-step task, it also brings up an important follow-up question as neural measure (e.g., fMRI, N. D. Daw, Gershman, et al., 2011) or physiological data (e.g., eye-tracking, Konovalov and Krajbich, 2016) are accompanied with the behavioral measures: how much of the neural or physiological findings at the group-average level are representative of neural and physiological processes of individuals deploying a behavioral strategy that deviates from the overall average strategy in the group?

To investigate this question, we leveraged the fMRI measures on the large participant sample and investigated the individual strategy group's neural processes in Chapter 3. We found the involvement of MF computations on decision-making (i.e., decision utility signal) in the medial prefrontal cortex across all strategy groups, which is consistent with previous group-level findings (Hampton, Bossaerts, and O'doherty, 2006; Beierholm et al., 2011). At the same time, the ubiquitous presence of the MF decision utility signal across distinct strategies (including the potentially non-RL strategies) underscores the essential role that a computationally efficient MF system serves in human Markov decision processes — MF computations are carried out indiscriminately to facilitate downstream value integration and action selection, in cooperation with the MB system. On the other hand, the decision-making computation (i.e., decision utility) from the MB system was found to be present in the medial prefrontal cortex most significantly in the MB strategy group, and the strength reduced gradually as the behavioral strategy shifted to the MF end of the spectrum. Such a finding again is consistent with the previous group-level neural hypothesis that the medial prefrontal cortex performs MB decision-related signal (Beierholm et al., 2011) or MF/MB value comparison (Wunderlich, Dayan, and Dolan, 2012), but it also makes a detailed neural hypothesis regarding the MB-related computations underlying various strategy type in use, which directly benefits from the individual difference analytic approach we took here. Relying on group-level behavioral and neural results to test theoretical hypotheses and draw insights into decision-making has been an adopted practice. Here through our work in Chapter 3, we demonstrate that indiscriminate generalization of group-level findings to each individual could cover intrinsic data variance that, if identified properly, could make the picture of decision-making theory more complete to account for heterogeneity in the biological agents. It highlights the significance of deploying the individual difference approach in the field of learning and decision-making to develop meaningful behavioral characterization (Charpentier et al., 2024) and the corresponding neural theory to achieve high explanatory power on real-world behaviors.

As shown in Chapter 3, various types of controls of MB, MF, or a mixture of both could be elicited across different individuals using the two-step task. It is an interesting question as to why different individuals facing the same decision-making task would deploy different behavioral strategies, which could have important psychiatric implications that we will discuss further later. But for now, if we take the current behavioral profile across individuals as it is, it makes sense to make the assumption that for a given individual, the necessary neural modules to implement a different strategy like what is implemented in other individuals of the same cohort are likely to be also available in this individual. Thus, if we further zoom into each individual's strategy on the timescale of completing the entire task, it is possible that the strategy deployed by an individual could change as the experiment goes on. Having the flexibility of deploying different kinds of computational systems for decisions could benefit the biological agent as the environment keeps changing, and various computations of each strategy would have certain advantages or disadvantages upon the changes. The pros and cons of systems could be theoretically leveraged mainly through 1) the uncertainty evaluation regarding the system in its signature predictions (N. D. Daw, Niv, and Dayan, 2005; Lee, Shimojo, and O'doherty, 2014), 2) the cost-benefit analysis of deploying different strategy systems (Keramati, Dezfouli, and Piray, 2011; Pezzulo, Rigoli, and Chersi, 2013), or 3) the system optimality in reward accruement (Simon and N. Daw, 2011; Kool, Gershman, and F. A. Cushman, 2017; Yi and O'Doherty, 2023). An arbitrator can then effectively adjust the strategy system in use according to these pros and cons when environmental statistics of reward and states shift.

In Chapter 2, we found direct behavioral evidence supporting an arbitration framework centered on system uncertainty: the reliability-based arbitration framework, which makes use of various types of prediction errors to approximate the system reliability (Lee, Shimojo, and O'doherty, 2014). The findings fit well with the recently proposed generalized arbitration framework in decision neuroscience: Mixture of Experts (MoE, O'Doherty et al., 2021), which was adapted from the field of machine learning (Jacobs et al., 1991). As the prediction error signals, both on reward and state learning, have been robustly found to be neurally encoded in classical and instrumental conditioning settings in the literature (Schultz, Dayan, and Montague, 1997; Schultz and Dickinson, 2000; O'Doherty et al., 2003; Gläscher et al., 2010), it is natural to assume that these neural signals could be easily adapted to guide an upper-level decision to make in the brain — which strategy system to use according to the prediction errors the systems generate. It is reassuring that the role of the RPE signal from the MF system and the SPE signal from the MB system serves as the system reliability for the arbitrator's decision, given that the MF system predominantly learns the environmental reward and the MB system learns the state structure by definition (Sutton, 2018). We also found the reward prediction error signal from the MB system, if hypothesized as another source of MB system reliability, also exerted a significant effect on the arbitration of controls. Theoretically, one might consider the reward prediction error a conventional MF learning signal that also contributes to MB value computation, yet in our study, the reward learning for the MF and MB systems were modeled as learning magnitude or binary, respectively; hence the two system's reward prediction error signal can have dissociable contributing effects on control selection during arbitration. This behavioral finding about the role of MB RPE also echoes the neural encoding result we found in Chapter 3: the MB RPE signal was found to be significantly encoded in the ventral striatum, whereas the RPE signal was found to be encoded significantly in the dorsal caudate. The non-overlapping nature of the brain region that encodes these two signals provides evidence that, indeed, in the current two-step task, the MF and MB systems could possibly learn two different aspects of the reward in the environment simultaneously. In the future, the causal role of the RPE signal from the MF and MB system in arbitration could be validated through neural stimulation in regions encoding these RPE signals by observing if the behavioral output could be biased towards MF (or MB) control contingent upon the neurally represented MB (or MF) RPE signal being enlarged by stimulation.

Given the multi-facet nature of environmental statistics of reward, other aspects of the reward, beyond magnitude and contingency, could potentially guide the arbitrator as well. In one study, it has been shown behaviorally that the volatility and noise of reward magnitude would affect the arbitration process according to their influence on the uncertainty of value estimation (Simon and N. Daw, 2011). Specifically, the presence of both low noise and reasonably high volatility in rewards should make the MB system more favorable in this uncertainty-related framework. It makes intuitive sense as the flexible cognitive MB system could react properly to sudden changes in rewards, and the MB system would have little advantage if there are relatively large reward samples needed to average for accurate estimation when there is large reward noise. To account for such observation through the reliability-based arbitration framework, higher MF RPE (magnitude) or lower MB RPE (binary) should be observed when the reward volatility is high and the reward noise is low, compared to other combinations of volatility and noise levels, which needs model simulations and behavioral testing using the two-step task presented in this thesis for further examination. Although the manifested temporal effect on prediction error from reward volatility might be comparable to those imposed from reward magnitude and contingency, it is not hard to imagine there could be other kinds of "volatility" defined on various lengths of unit time. It is an open question how these temporally defined reward features and potentially state features could contribute to the relatively static features of reward and state so far. This brings up the question for future studies to answer: whether the computational model and the corresponding neural implementation of the arbitration model that was found to have explanatory power for the "static" reward feature (Lee, Shimojo, and O'doherty, 2014) could also generalize to cases when the "temporal" reward features are weighed in for system reliability evaluation. Also, in the exploratory analysis of Chapter 2, we found evidence that MF RPE and MB RPE interact as reliability signals during the arbitration process. It is an ecologically interesting research problem how prediction errors estimated over multiple timescales could interact to guide the arbitrator to allocate the controller's weight. Future works with sophisticated task design and computational modeling are needed to tackle this problem which is pertinent to real-world decision-making given the complexity of the environment.

The goal-directed or model-based strategy is considered to be more capable of representing the environment to accrue more rewards, but at the same time, is cognitive resource demanding and requires additional computational time for deliberation, given the working memory and the amount of complex computations needed (Shenhav, Botvinick, and Cohen, 2013). Arbitration schemes that evaluate this cost-benefit trade-off regarding the engagement of the MB system have been proposed (Keramati, Dezfouli, and Piray, 2011; Pezzulo, Rigoli, and Chersi, 2013). In our post-hoc exploratory analysis in Chapter 2, we did not find direct supporting behavioral evidence of such cost-benefit arbitration. In the framework of Mixture of Experts (O'Doherty et al., 2021), the cost-benefit analysis of engaging the more complex MB algorithms could be implicitly reflected in the evaluation of the prediction reliability of the MB algorithm, as high prediction reliability (i.e., low prediction error) would likely lead to better reward accruement, and too complex models that use a lot of cognitive resources would potentially generate a lot of prediction errors (i.e., low prediction reliability), as too complex models could not generalize well to new data samples due to the bias/variance trade-off (Geman, Bienenstock, and Doursat, 1992; Von Luxburg and Schölkopf, 2011). Hence, our current finding is consistent with the potential explanatory power entailed in the MoE framework, and we hope to see more future empirical studies confirm the role of the proposed essential minimal factors (i.e., prediction uncertainty) in characterizing well all aspects of the arbitration process.

We as individuals navigating life could be described with sequences of various types of actions, with these actions expressed by us with different traits and states of mood, and so on and so forth. Daily tasks as simple as driving or cooking could be accomplished differently depending on one's state of mood or the valence of temporally proximal events. As we have previously alluded to in the discussion of individual differences in strategy use in the two-step task, individual differences of strategy use in cognitive tasks identified through computational models could show significant relationships with measures of one's psychiatric traits, especially if the tasks are well chosen, and the behaviors are properly modeled to capture the key symptomatic signatures in the psychiatric disorders (Zou et al., 2022; Wu et al., 2024). As the computational models in such investigations are typically composed of lower-level mechanistic components (e.g., learning, value integration, planning, action selection) that coordinate together to serve the cognitive tasks, it could then provide a detailed scope to study the problematic node that gives rise to the observed abnormality, which would be beyond what could be offered through a descriptive psychological model of the disorder of interest.

In Chapter 4, we found a positive correlation between one's obsessive-compulsive tendency (OCI-R score and its obsessiveness and compulsiveness sub-score) and one's degree of using the MF strategy in the two-step task. Computationally, such a finding, along with previous literature (Voon et al., 2015; C. M. Gillan, Kosinski, et al., 2016), informs us that the bias towards relying on habitual control in populations with obsessive-compulsive disorder (OCD) can be traced back to the computation of state-action value learning and the subsequent action selection in the RL framework. Pinning down the source of how the variation in computations relates to the variation in obsessive-compulsive (OC) tendency would require finer-grained investigations on computational model parameters. Within the framework of a hybrid RL model with a semi-arbitration mechanism discussed throughout the thesis, the learning rate of the MF and MB system or the arbitration parameter to decide the model weight within the model could both potentially contribute to the biased behaviors in participants with high OC tendency. Out of speculations, it is possible, just as part of all possibilities, that the observed overreliance on MF control is due to an impaired MB learning process that could be reflected through a low MB learning rate or due to the weight bias added towards the MF value in the arbitration process (N. D. Daw, Niv, and Dayan, 2005; Lee, Shimojo, and O'doherty, 2014). Although not directly approached computationally in our work in Chapter 4, such modeling efforts to know the specific mechanistic causes of the overexpression of habit-like behavior are critical to curating better therapy for OCD. A problematic value learning process in the goal-directed (or MB) system would suggest treatments emphasizing prompting the opportunity cost of engaging the maladaptive ritualistic behaviors, whereas if the problem sits in the lack of a proper control allocation process, then behavioral intervention such as exposure to symptom-triggering stimuli with inhibition of OC behavior could potentially strengthen the relevant neural "muscle" to execute adaptive arbitration. It should be noted that these are some naive ideas for cognitive behavioral therapy inspired by potential computational psychiatric findings on OCD, and more stringent and robust clinical testing is needed to validate such proposals before any applicative practice.

Understanding the computational mechanism of OCD is valuable in itself for learning and curing the disorder better, and at the same time, it is also helpful to guide the integration of neuroscience knowledge to obtain insights into mechanisms of neural computations underlying OCD. Our finding in Chapter 4 regarding the role of the anterior cingulate cortex (ACC) in compulsive behaviors is an example. It has long been suggested in the cognitive control literature that ACC serves as a performance monitor that is sensitive to behavioral errors and conflicts so that the actions can be adjusted upon needs (Holroyd and Coles, 2002). OCD involves constant feelings of error/conflict as well as the need to adjust, and hence the findings associated OCD with abnormal activation of ACC were inspiring as they led to one neural source of OCD with respect to the error/conflict-monitoring process(Gehring, Himle, and Nisenson, 2000; Ursu et al., 2003; Fitzgerald et al., 2005); however, it still remained unclear what specific computations ACC contributes when monitoring errors/conflicts to generate the behavioral deficits among OCD. Is ACC signaling performance "error" in a hyperactive way? Or is ACC conveying action adjustment information for executive control after an experienced error/conflict OCD? Utilizing a model-based neural analytic approach helped clarify the answer. OCD approached as the imbalance between the MF system and MB system in computational terms, could be explained by an abnormal reliability-based arbitration process in evaluating the absolute prediction error signals as the proxy of system reliability (Lee, Shimojo, and O'doherty, 2014; Kim et al., 2024); the reliability of the MB system is deemed too low so that behaviors are mostly driven by the MF system. The analysis of fitting the dynamics of MB reward prediction error to the fMRI blood-oxygen-level-dependent (BOLD) signal across individuals of varying OC tendency cast light on a potential cause of the overreliance on habitual/MF behavior in OCD: ACC sends input for reliability approximations to the arbitrator in an abnormally enhanced manner so that the neural arbitrator perceives the MB system as very unreliable, the degree of which is so large that it mismatches the actual reliability estimated through the prediction error information in the external environment.

The abnormal activity of ACC underlying OC behaviors was a bit better understood from a scope of impaired arbitration between the MF and MB systems in the RL framework. Yet some additional questions remain to be answered for a complete neurocomputational model of OCD. Although the MB RPE encoding strength in

ACC significantly correlated with an individual's self-reported OC tendency, indicating ACC's role in signaling the neurally-perceived RPE signal for the arbitrator to evaluate, it is a little mysterious why ACC itself does not encode MB RPE at the group level and why we only found MB RPE encoding in the ventral striatum (Chapter 3). Also, there seems to be a functional specificity of ACC that relays information related to the system reliability for arbitration, as it is only ACC that shows MB RPE encoding strength correlation with OC severity, and we did not find such correlations in other PE-encoding regions we found at the group level in Chapter 3 (i.e., MF RPE: dorsal caudate, MB RPE: ventral striatum). One possibility is that the MB RPE encoding in the ventral striatum and MB RPE encoding strength correlation observed in ACC simply reflect two different stages of neural computations when the brain learns the value through reward experience and adjusts action policy based upon behavioral feedback. Holroyd and Coles (2002) proposed a unified account of two neural systems that cooperate for behavioral control: a reinforcement learning system in the basal ganglia and an error-processing system in the anterior cingulate cortex. In this theory, the basal ganglia learn about the state values through reward prediction error signals when interacting with the environment, and the reward prediction error signals are conveyed to ACC for its monitoring and the subsequent adjustment of action policy by commanding the motor module in the brain. Under this framework, the classic error-related negativity (ERN) in ACC is a reflection of RPE relayed from the basal ganglia, and the ERN was used for behavioral policy adjustment. This framework could potentially explain our current findings of the ventral striatum and ACC encoding different information related to MB RPE, with the former encoding a pure learning signal and the latter monitoring the relayed MB RPE signal from the ventral striatum for strategy adjustment at the individual level. The presented work in Chapter 4 potentially only unravels a partial neural mechanism that underlies OC behaviors. Future works are needed to study how the action selection process might be differently implemented in OC vs. "normal" behaviors; also, it is a question to be answered that besides the neural representation of prediction error in ACC, whether the value representation underlying OC behaviors also shows distinct patterns compared to that underlying the "normal" behaviors and how the distinct patterns of value encoding drive the stereotyped action policy in OC behaviors.

Our knowledge of reinforcement learning computations and the corresponding neural regions that implement such computations underlying obsessive-compulsive behaviors could inform us of potentially better neural therapy that mediates symptoms of OCD. Given that ACC could overrepresent the MB RPE signal sent by the ventral striatum in high OC-tendency people, neural intervention that mediates the strength of neural communications between ACC and ventral striatum could help alleviate the MF bias in OCD, as the reliability-based arbitration theory would predict. Additionally, conditional neural inhibition of ACC could also help reduce the maladaptive policy adjustment adopted by the OCD population due to the hypersensitive ACC in delivering motor commands for downstream neural regions. More insights, if any, on how the error signals are abnormally generated in the first place in the brain underlying OCD could also help identify effective clinical interventions correspondingly that help the OCD populations alleviate their subjective feeling of "error" and doubt. Overall, the so-called computational psychiatry approach that relates task measures of certain cognitive process engagement through computational modeling to the self-reported psychiatric measures could have important clinical implications (Huys et al., 2021), as not only a detailed understanding of but also neuroscience-based therapies of the mental disorder could be developed alongside the neurotechnological tools (e.g., transcranial direct current stimulation, Brunelin et al., 2018; transcranial magnetic stimulation, Trevizol et al., 2016; Rapinesi et al., 2019).

In sum, in this thesis, we first focused on the computations of learning, choice valuation, and strategy arbitration in reinforcement learning and their neural implementations across a large sample of healthy participants, with a special focus on individual differences in behavioral strategies. Our findings suggest that these relevant computations from the MF and MB systems, the two fundamentals in the reinforcement learning framework, are present in human Markov decisions with both behavioral and neural evidence. Despite the fact that expressed behavioral strategies vary from individual to individual, the commonality of the underlying neural computations shared by these distinct behavioral strategies is the MF RL computations, highlighting the influence of an evolutionarily early neural system featuring reflectiveness and efficiency on our daily decisions today. In response to moment-to-moment environmental changes when navigating the world, an adaptive arbitration mechanism between these behavioral strategies is critical for our biological fitness. We expanded the dictionary of the factors considered by the arbitrator in a reliability-based framework, and we expect to have more vocabulary not just for arbitration within the RL strategies but also for generalized arbitration between RL and non-RL strategies. Moving from adaptive decision-making to maladaptive behaviors, our work also suggests that the abnormal neural encoding of an arbitration signal related to reward prediction (i.e., MB RPE) could help explain the behavioral compulsiveness, proposing the role of the anterior cingulate cortex as the mediator between learning and strategy adjustment to maintain adaptive decision-making. With the computational modeling approach and fine-grained neural investigations focusing on individual differences, we hope to gain more insights into the homogeneity and heterogeneity of the computations and the neural implementation of adaptive and maladaptive decision-making strategies within and beyond the reinforcement learning framework. Importantly, such insights, in relation to psychiatric disorders, can help us develop individually-targeted clinical interventions to maintain our decision adaptiveness as human beings.

References

Beierholm, Ulrik R et al. (2011). “Separate encoding of model-based and model-free valuations in the human brain”. In: *Neuroimage* 58.3, pp. 955–962.

Berns, Gregory S et al. (2001). “Predictability modulates human brain response to reward”. In: *Journal of neuroscience* 21.8, pp. 2793–2798.

Brunelin, Jérôme et al. (2018). “Transcranial direct current stimulation for obsessive-compulsive disorder: a systematic review”. In: *Brain sciences* 8.2, p. 37.

Charpentier, Caroline J et al. (2024). “Heterogeneity in strategy use during arbitration between experiential and observational learning”. In: *Nature Communications* 15.1, p. 4436.

Cushman, Fiery and Adam Morris (2015). “Habitual control of goal selection in humans”. In: *Proceedings of the National Academy of Sciences* 112.45, pp. 13817–13822.

Daw, Nathaniel D, Samuel J Gershman, et al. (2011). “Model-based influences on humans’ choices and striatal prediction errors”. In: *Neuron* 69.6, pp. 1204–1215.

Daw, Nathaniel D, Yael Niv, and Peter Dayan (2005). “Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control”. In: *Nature neuroscience* 8.12, pp. 1704–1711.

Dezfouli, Amir and Bernard W Balleine (2013). “Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized”. In: *PLoS computational biology* 9.12, e1003364.

Dezfouli, Amir, Nura W Lingawi, and Bernard W Balleine (2014). “Habits as action sequences: hierarchical action control and changes in outcome value”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1655, p. 20130482.

Doll, Bradley B et al. (2015). “Model-based choices involve prospective neural activity”. In: *Nature neuroscience* 18.5, pp. 767–772.

Feher da Silva, Carolina and Todd A Hare (2020). “Humans primarily use model-based inference in the two-stage task”. In: *Nature Human Behaviour* 4.10, pp. 1053–1066.

Feher da Silva, Carolina, Gaia Lombardi, et al. (2023). “Rethinking model-based and model-free influences on mental effort and striatal prediction errors”. In: *Nature Human Behaviour* 7.6, pp. 956–969.

Fiorillo, Christopher D, Philippe N Tobler, and Wolfram Schultz (2003). “Discrete coding of reward probability and uncertainty by dopamine neurons”. In: *Science* 299.5614, pp. 1898–1902.

Fitzgerald, Kate Dimond et al. (2005). “Error-related hyperactivity of the anterior cingulate cortex in obsessive-compulsive disorder”. In: *Biological psychiatry* 57.3, pp. 287–294.

Gehring, William J, Joseph Himle, and Laura G Nisenson (2000). “Action-monitoring dysfunction in obsessive-compulsive disorder”. In: *Psychological science* 11.1, pp. 1–6.

Geman, Stuart, Elie Bienenstock, and René Doursat (1992). “Neural networks and the bias/variance dilemma”. In: *Neural computation* 4.1, pp. 1–58.

Gillan, Claire M, Michal Kosinski, et al. (2016). “Characterizing a psychiatric symptom dimension related to deficits in goal-directed control”. In: *elife* 5, e11305.

Gillan, Claire M, A Ross Otto, et al. (2015). “Model-based learning protects against forming habits”. In: *Cognitive, Affective, & Behavioral Neuroscience* 15, pp. 523–536.

Gläscher, Jan et al. (2010). “States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning”. In: *Neuron* 66.4, pp. 585–595.

Hampton, Alan N, Peter Bossaerts, and John P O’doherty (2006). “The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans”. In: *Journal of Neuroscience* 26.32, pp. 8360–8367.

Hollerman, Jeffrey R and Wolfram Schultz (1998). “Dopamine neurons report an error in the temporal prediction of reward during learning”. In: *Nature neuroscience* 1.4, pp. 304–309.

Holroyd, Clay B and Michael GH Coles (2002). “The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity.” In: *Psychological review* 109.4, p. 679.

Huys, Quentin JM et al. (2021). “Advances in the computational understanding of mental illness”. In: *Neuropsychopharmacology* 46.1, pp. 3–19.

Jacobs, Robert A et al. (1991). “Adaptive mixtures of local experts”. In: *Neural computation* 3.1, pp. 79–87.

Jarvik, Murray E (1951). “Probability learning and a negative recency effect in the serial anticipation of alternative symbols.” In: *Journal of experimental psychology* 41.4, p. 291.

Keramati, Mehdi, Amir Dezfouli, and Payam Piray (2011). “Speed/accuracy tradeoff between the habitual and the goal-directed processes”. In: *PLoS computational biology* 7.5, e1002055.

Kim, Taekwan et al. (2024). “Neurocomputational model of compulsivity: deviating from an uncertain goal-directed system”. In: *Brain* 147.6, pp. 2230–2244.

Konovalov, Arkady and Ian Krajbich (2016). “Gaze data reveal distinct choice processes underlying model-based and model-free reinforcement learning”. In: *Nature communications* 7.1, p. 12438.

Kool, Wouter, Fiery A Cushman, and Samuel J Gershman (2016). “When does model-based control pay off?” In: *PLoS computational biology* 12.8, e1005090.

Kool, Wouter, Samuel J Gershman, and Fiery A Cushman (2017). “Cost-benefit arbitration between multiple reinforcement-learning systems”. In: *Psychological science* 28.9, pp. 1321–1333.

Lee, Sang Wan, Shinsuke Shimojo, and John P O’doherty (2014). “Neural computations underlying arbitration between model-based and model-free learning”. In: *Neuron* 81.3, pp. 687–699.

O’Doherty, John P et al. (2003). “Temporal difference models and reward-related learning in the human brain”. In: *Neuron* 38.2, pp. 329–337.

O’Doherty, John P et al. (2021). “Why and how the brain weights contributions from a mixture of experts”. In: *Neuroscience & Biobehavioral Reviews* 123, pp. 14–23.

Otto, A Ross et al. (2013). “Working-memory capacity protects model-based learning from stress”. In: *Proceedings of the National Academy of Sciences* 110.52, pp. 20941–20946.

Pezzulo, Giovanni, Francesco Rigoli, and Fabian Chersi (2013). “The mixed instrumental controller: using value of information to combine habitual choice and mental simulation”. In: *Frontiers in psychology* 4, p. 92.

Rapinesi, Chiara et al. (2019). “Brain stimulation in obsessive-compulsive disorder (OCD): a systematic review”. In: *Current neuropharmacology* 17.8, pp. 787–807.

Schultz, Wolfram, Peter Dayan, and P Read Montague (1997). “A neural substrate of prediction and reward”. In: *Science* 275.5306, pp. 1593–1599.

Schultz, Wolfram and Anthony Dickinson (2000). “Neuronal coding of prediction errors”. In: *Annual review of neuroscience* 23.1, pp. 473–500.

Shenhav, Amitai, Matthew M Botvinick, and Jonathan D Cohen (2013). “The expected value of control: an integrative theory of anterior cingulate cortex function”. In: *Neuron* 79.2, pp. 217–240.

Simon, Dylan and Nathaniel Daw (2011). “Environmental statistics and the trade-off between model-based and TD learning in humans”. In: *Advances in neural information processing systems* 24.

Smittenaar, Peter et al. (2013). “Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans”. In: *Neuron* 80.4, pp. 914–919.

Sutton, Richard S (2018). “Reinforcement learning: An introduction”. In: *A Bradford Book*.

Trevizol, Alisson Paulino et al. (2016). “Transcranial magnetic stimulation for obsessive-compulsive disorder: an updated systematic review and meta-analysis”. In: *The journal of ECT* 32.4, pp. 262–266.

Tversky, Amos and Daniel Kahneman (1971). “Belief in the law of small numbers.” In: *Psychological bulletin* 76.2, p. 105.

Ursu, Stefan et al. (2003). “Overactive action monitoring in obsessive-compulsive disorder: evidence from functional magnetic resonance imaging”. In: *Psychological science* 14.4, pp. 347–353.

Von Luxburg, Ulrike and Bernhard Schölkopf (2011). “Statistical learning theory: Models, concepts, and results”. In: *Handbook of the History of Logic*. Vol. 10. Elsevier, pp. 651–706.

Voon, Valerie et al. (2015). “Disorders of compulsivity: a common bias towards learning habits”. In: *Molecular psychiatry* 20.3, pp. 345–352.

Waelti, Pascale, Anthony Dickinson, and Wolfram Schultz (2001). “Dopamine responses comply with basic assumptions of formal learning theory”. In: *Nature* 412.6842, pp. 43–48.

Wu, Qianying et al. (2024). “Individual differences in autism-like traits are associated with reduced goal emulation in a computational model of observational learning”. In: *Nature Mental Health* 2.9, pp. 1032–1044.

Wunderlich, Klaus, Peter Dayan, and Raymond J Dolan (2012). “Mapping value based planning and extensively trained choice in the human brain”. In: *Nature neuroscience* 15.5, pp. 786–791.

Wunderlich, Klaus, Peter Smittenaar, and Raymond J Dolan (2012). “Dopamine enhances model-based over model-free choice behavior”. In: *Neuron* 75.3, pp. 418–424.

Yi, Sanghyun and John P O’Doherty (2023). “Computational and neural mechanisms underlying the influence of action affordances on value learning”. In: *bioRxiv*, pp. 2023–07.

Zou, Amy R et al. (2022). “Impulsivity relates to multi-trial choice strategy in probabilistic reversal learning”. In: *Frontiers in Psychiatry* 13, p. 800290.