ABSTRACT

Title of Dissertation: CONFLICT AND COMPETITION BETWEEN MODEL-BASED AND MODEL-FREE CONTROL

Yuqing Lei
Master of Science, 2021

Thesis Directed by: Professor Alec Solway
Department of Psychology

There are two learning systems behind human decision-making: the model-based (MB) system and the model-free (MF) system. While they both contribute to decision-making behaviors, it is not clear how the two systems interact to formulate a single decision, especially when they are in conflict. This present thesis defines decision conflict between the systems in two popular binary-choice tasks: Daw's Two-step task and Kool's Rocket Task. We used hierarchical modeling to identify conflict-related changes during decision process using the Drift-Diffusion Model (DDM). Evidence showed that the MB system compromises when there is a conflict with the MF system, whether the conflict is on the valuational level or action level. We also looked at how a key component of obsessive compulsive disorder (OCD), the cognitive self-consciousness (CSC), affects the two learning systems during decision.

CONFLICT AND COMPETITION BETWEEN MODEL-BASED AND MODEL-FREE CONTROL

by

Yuqing Lei

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Master of Science
2021

Advisory Committee:
Professor Alec Solway, Chair/Advisor
Professor Matthew Roesch
Professor James Grand

© Copyright by
Yuqing Lei
2021

Table of Contents

Table of Contents ii

List of Tables iv

List of Figures v

Chapter 1: Introduction 1
- 1.1 Decision Conflicts between Learning Systems . . . 1
- 1.2 Reinforcement Learning . . . 9
  - 1.2.1 Model-free Learning . . . 9
  - 1.2.2 Model-based Learning . . . 10
- 1.3 Cognitive Control and Conflicts . . . 14
- 1.4 Computational Modeling of Decisions . . . 19
- 1.5 Bayesian Modeling . . . 22
- 1.6 . . . 23
  - 1.6.1 Computational Modeling of Decisions . . . 27
- 1.7 Metacognition on Anxiety . . . 30

Chapter 2: Methods 36
- 2.1 Datasets . . . 36
  - 2.1.1 Unpublished Dataset . . . 37
- 2.2 Learning algorithms . . . 38
- 2.3 Decision-Making Model . . . 40
  - 2.3.1 Softmax Models . . . 41
  - 2.3.2 Drift Diffusion Models . . . 42
  - 2.3.3 Adding Individual Traits as Regressors . . . 46
  - 2.3.4 Model Fitting . . . 48
  - 2.3.5 Correlation Analysis . . . 49

Chapter 3: Results 50
- 3.1 Overview . . . 50
- 3.2 Effects of Valuation Conflicts and Action Conflicts . . . 51
- 3.3 Correlational Caveat . . . 54
- 3.4 Response Perseveration Influence Decisions . . . 55
- 3.5 Individual Traits . . . 56

Chapter 4: Discussion 58

4.1 Overview 58
4.2 Process Model of MB Control and MB-MF Integration 60
4.3 Caveats 63
4.4 Choice Perseveration 64
4.5 Metacognition and Decision Making 65

Appendix A: Conflict Model Parameter Fits 67

Appendix B: Individual Differences Model Parameter Fits 79

Bibliography 88

List of Tables

2.1 Sources and sample sizes of the four datasets. 36

A.1 Kool,2016/RK Group Parameters 67
A.2 Solway,2019/TS Group Parameters 70
A.3 Unpublished/TS Group Parameters 73
A.4 Kool, 2016/TS Group Parameters 76

B.1 Unpublished/TS Group Parameters with MCQCSC Regressors, Controlling for State Anxiety 79
B.2 Unpublished/TS Group Parameters with MCQCSC Regressors, Without Controlling for State Anxiety 84

List of Figures

1.1 From Daw et al., 2011. (a) Timeline of events in trial. A first-stage choice between two options (green) probabilistically lead to a second-stage choice (pink).(b) Probabilistic State transition. Each first-stage choice had a more 'associated' outcome. Upon making the specific choice, 70% of the time the participant was directed to the more associated state and was directed to the less associated state the remaining 30% of the time. 16

1.2 From Kool et al.,2016. The transition structure of the task. In each trial, one pair of rockets (from a set of two rocket pairs) randomly appeared and the participant was asked to choose one of the rockets from the apparent pair. The rocket transitioned to the matching planet, where the participant needed to press the spacebar to collect either space treasure (positive points) or antimatter (negative points). The points on each planet changed independently and slowly. Subjects were asked to get as many points as possible. 17

1.3 Drift Diffusion Model and it's parameters. Evidence towards either decision boundary is collected at the speed of the drift rate v, with a start point bias z. The reaction time of the decision is the duration of information collection in addition to the non-decision time. 21

1.4 From Kool et al.,2016. The transition structure of the task. In each trial, one pair of rockets (from a set of two rocket pairs) randomly appeared and the participant was asked to choose one of the rockets from the apparent pair. The rocket transitioned to the matching planet, where the participant needed to press the spacebar to collect either space treasure (positive points) or antimatter (negative points). The points on each planet changed independently and slowly. Subjects were asked to get as many points as possible. 25

1.5 Drift Diffusion Model and it's parameters. Evidence towards either decision boundary is collected at the speed of the drift rate v, with a start point bias z. The reaction time of the decision is the duration of information collection in addition to the non-decision time. 29

3.1 Model fits of conflict-related drift rate parameters in all four datasets. The top and middle panels represent the MB and the MF drift rate changes related to Q value conflict, action conflict, and conflict with interaction (the sum of the Q value conflict term and the interaction term) from left to right. The bottom panel presents the substraction between the two systems' drift rate changes. 51
3.2 Model fits of conflict-related boundary parameters in all four datasets. 53
3.3 Correlation coefficients of the three Q quantities (within-MF-system difference, within-MB-system difference, and between-system difference) in Kool2016/rk. 54
3.4 Response perseveration across four datasets. 56
3.5 Model fits of MCQCSC regressors, the MB drift rate change (left) and the MB drift rate change related to Q value conflicts (right). . . . 57

Chapter 1: Introduction

1.1 Decision Conflicts between Learning Systems

How do people make decisions based on past experiences? Imagine a man who won a big prize from his very first lottery ticket. On one hand, this pleasant outcome could motivate him to buy another lottery ticket, promoting a habitual drive "to buy". On the other hand, his rational thinking may tell him "not to buy" another ticket, since the chance of winning again is fairly low. This simultaneous coexistence of the former neural system, habitual control, with the latter neural system, goal-directed control, has been demonstrated in prior research (e.g., reviewed by [1]). However, it remains unclear how these two controllers interact to generate a single decision, for example to buy or not to buy. How do the two systems adapt and resolve their apparent conflict with one other?

The dual controllers of decisions [1], goal-directed control and habitual control, together generate two learning mechanisms in parallel: the model-based (MB) learning and the model-free (MF) learning [2]. Both processes allow the human mind to consistently evaluate actions in a certain environment, and have long been described in the computational reinforcement learning framework. Conventionally, habitual action has been understood as learned behavior patterns after trial-and-

error explorations. Reward conditioning of habitual learning allows agents to build an internal association between reward stimuli and responses. Such a learning process creates a stimulus-value map that informs the agent of the best action to take (i.e., the action most likely to receive a reward) after a given stimulus. This can occur without active planning or a structural knowledge representation of the environment, often described as "model-free learning." Meanwhile, agents may also plan and direct their actions using a structural representation of the task environment via a top-down, goal-directed approach [1]. In multi-step tasks, different actions may lead to different succeeding situations. This transition information is crucial for agents to choose the route with the best outcome or reward. This type of learning—coined "model-based" learning—has the capacity to combine this transition structure with reward outcomes. Although MB and MF learning don't match to the goal-directed and habitual controls in a one-to-one fashion, the two learning strategies are closely related to the two separate but interacting circuits. Previous review showed that the learning of top-down knowledge is largely related to the prefrontal cortex and the medial striatum, while the habitual, MF learning strategy relies more so on the lateral striatum (reviewed by [3]). While MB and MF learning both rely on the basal ganglia [4], particularly the dopaminergic system [5], MB learning incorporates signals from prefrontal cortex, making this learning system similar to goal-directed control. However, unlike the goal-directed control which generally refers to an overriding signal that switches attention to goal-relevant information, MB learning specifically refers to a plan-forward strategy that utilizes environment or task structure.

Each of the two learning systems, MB and MF, has their advantages and disadvantages. MF learning provides a quick, generalized estimation of action reward and thus does not require significant cognitive resources; it relies purely on past experiences to estimate the reward outcome for each action. Consequently, MF learning is more prone to biases, for example, caused by rare events. Comparatively, MB learning often provides a more objective estimation of action-outcomes, taking advantage of structural information, such as the state transitions of the environment. This means it requires more attention resources to be allocated to learning this information and planning successful actions. Together, the two systems compensate each other with their complementary features to optimize both the efficiency and the accuracy of learning.

Many studies have addressed the need to have both systems (e.g. [1, 5]), how the two systems work independently (e.g. [6]), and how the two learning processes potentially accommodate under the same framework (e.g. [7–9]). However, there is little work on how the two systems cooperate or compete in order to generate an incorporated response, especially when the two systems have different valuation of choices.

How do the two systems interact when favoring different choices? Typically, conflict between two cognitive systems is monitored and resolved by cognitive control [10]. Since habitual actions (i.e., being model-free) and goal-directed control (i.e., being model-based) are operated by two different neural circuits and consume different cognitive resources (e.g. [11]), cognitive conflicts occur when the two systems generate contradicting responses.

There are different ways that conflict monitoring and control over the the two systems can be examined computationally. We approached them using the Drift Diffusion Model (DDM; see [12,13]), a model that includes latent decision variables under the Evidence-Integration framework (EA; [14]). EA theory proposes that a decision is made by incrementally collecting information over time until reaching a decision threshold. DDM represents the decision between two options using the gradual accumulation of evidence towards either choice, and has been widely validated across many experimental data, including perceptual discrimination [15,16], go-no-go tasks [17], and value-based choice tasks [18].

Still, it remains unclear how exactly the conflict between the MB and the MF learning systems is resolved. The majority of research on cognitive conflicts, including the example above, examine goal-directed versus habitual control through incongruent stimuli. In these tasks, stimuli often contain task-irrelevant information that activate habitual control only to impede goal-directed behaviors. For example, in the Stroop Color and Word Test, participants are presented a set of words which are names of colors and are asked to say the font color of each word. In non-conflict trials, the color of the ink is congruent with the color name (e.g., the word "green" also has a font color of green); in conflict trials, the color of the ink is incongruent with the color name (e.g., the word "green" has a font color of red). Participants showed prolonged reaction times and higher error rates in conflict trials than in non-conflict trials [19], more so when they are asked to read the color of the word. This is because, although both the literal meanings and the colors of the words are processed autonomously, processing of the former is much faster and more autonomous than

the processing of the latter [20]. When participants are asked to say the color of the word, they need to deploy attention to focus on the color information over the lexical information, making the response slower compared to an autonomous process. In this case, the cognitive conflict happens on the stimuli level.

Conflict between the two learning systems is a massively different product: the two systems evaluate the same information, but generate dissonant value estimations, due to their different strategies of learning (i.e., with and without structural knowledge of the environment), Once there is a conflict between the two systems, they have to compete until the difference settles down. The competition results in one or both systems ultimately being suppressed relative to their baseline performance. Meanwhile, cognitive control pushes back decision "deadline", prolonging reaction time and providing space for alternation of action [21].

Identifying how the two systems react to their conflicts is crucial to understanding the nature of human decisions. Considering all of these factors, it is beneficial to model these conflicts and corresponding changes in the learning process in a computational framework that integrates reinforcement learning framework with the process model of decisions.

The dual controllers of decisions [1], goal-directed control and habitual control, together generate two learning mechanisms in parallel: the model-based (MB) learning and the model-free (MF) learning [2]. Both processes allow the human mind to consistently evaluate actions in a certain environment, and have long been described in the computational reinforcement learning framework. Conventionally, habitual action has been understood as learned behavior patterns after trial-and-

error explorations. Reward conditioning of habitual learning allows agents to build an internal association between reward stimuli and responses. Such a learning process creates a stimulus-value map that informs the agent of the best action to take (i.e., the action most likely to receive a reward) after a given stimulus. This can occur without active planning or a structural knowledge representation of the environment, often described as "model-free learning." Meanwhile, agents may also plan and direct their actions using a structural representation of the task environment via a top-down, goal-directed approach [1]. In multi-step tasks, different actions may lead to different succeeding situations. This transition information is crucial for agents to choose the route with the best outcome or reward. This type of learning—coined "model-based" learning—has the capacity to combine this transition structure with reward outcomes. Although MB and MF learning don't match to the goal-directed and habitual controls in a one-to-one fashion, the two learning strategies are closely related to the two separate but interacting circuits. Previous review showed that the learning of top-down knowledge is largely related to the prefrontal cortex and the medial striatum, while the habitual, MF learning strategy relies more so on the lateral striatum (reviewed by [3]). While MB and MF learning both rely on the basal ganglia [4], particularly the dopaminergic system [5], MB learning incorporates signals from prefrontal cortex, making this learning system similar to goal-directed control. However, unlike the goal-directed control which generally refers to an overriding signal that switches attention to goal-relevant information, MB learning specifically refers to a plan-forward strategy that utilizes environment or task structure.

Each of the two learning systems, MB and MF, has their advantages and disadvantages. MF learning provides a quick, generalized estimation of action reward and thus does not require significant cognitive resources; it relies purely on past experiences to estimate the reward outcome for each action. Consequently, MF learning is more prone to biases, for example, caused by rare events. Comparatively, MB learning often provides a more objective estimation of action-outcomes, taking advantage of structural information, such as the state transitions of the environment. This means it requires more attention resources to be allocated to learning this information and planning successful actions. Together, the two systems compensate each other with their complementary features to optimize both the efficiency and the accuracy of learning.

Many studies have addressed the need to have both systems (e.g. [1, 5]), how the two systems work independently (e.g. [6]), and how the two learning processes potentially accommodate under the same framework (e.g. [7–9]). However, there is little work on how the two systems cooperate or compete in order to generate an incorporated response, especially when the two systems have different valuation of choices.

How do the two systems interact when favoring different choices? Typically, conflict between two cognitive systems is monitored and resolved by cognitive control [10]. Since habitual actions (i.e., being model-free) and goal-directed control (i.e., being model-based) are operated by two different neural circuits and consume different cognitive resources (e.g. [11]), cognitive conflicts occur when the two systems generate contradicting responses.

There are different ways that conflict monitoring and control over the the two systems can be examined computationally. We approached them using the Drift Diffusion Model (DDM; see [12,13]), a model that includes latent decision variables under the Evidence-Integration framework (EA; [14]). EA theory proposes that a decision is made by incrementally collecting information over time until reaching a decision threshold. DDM represents the decision between two options using the gradual accumulation of evidence towards either choice, and has been widely validated across many experimental data, including perceptual discrimination [15,16], go-no-go tasks [17], and value-based choice tasks [18].

Still, it remains unclear how exactly the conflict between the MB and the MF learning systems is resolved. The majority of research on cognitive conflicts, including the example above, examine goal-directed versus habitual control through incongruent stimuli. In these tasks, stimuli often contain task-irrelevant information that activate habitual control only to impede goal-directed behaviors. For example, in the Stroop Color and Word Test, participants are presented a set of words which are names of colors and are asked to say the font color of each word. In non-conflict trials, the color of the ink is congruent with the color name (e.g., the word "green" also has a font color of green); in conflict trials, the color of the ink is incongruent with the color name (e.g., the word "green" has a font color of red). Participants showed prolonged reaction times and higher error rates in conflict trials than in non-conflict trials [19], more so when they are asked to read the color of the word. This is because, although both the literal meanings and the colors of the words are processed autonomously, processing of the former is much faster and more autonomous than

the processing of the latter [20]. When participants are asked to say the color of the word, they need to deploy attention to focus on the color information over the lexical information, making the response slower compared to an autonomous process. In this case, the cognitive conflict happens on the stimuli level.

Conflict between the two learning systems is a massively different product: the two systems evaluate the same information, but generate dissonant value estimations, due to their different strategies of learning (i.e., with and without structural knowledge of the environment), Once there is a conflict between the two systems, they have to compete until the difference settles down. The competition results in one or both systems ultimately being suppressed relative to their baseline performance. Meanwhile, cognitive control pushes back decision "deadline", prolonging reaction time and providing space for alternation of action [21].

Identifying how the two systems react to their conflicts is crucial to understanding the nature of human decisions. Considering all of these factors, it is beneficial to model these conflicts and corresponding changes in the learning process in a computational framework that integrates reinforcement learning framework with the process model of decisions.

1.2 Reinforcement Learning

1.2.1 Model-free Learning

MF learning, like other cognitive heuristics, is a simple and straightforward strategy for individuals to understand a complex problem [22]. It helps us to quickly

estimate the desirability of countless trivial choices in daily events, with a reasonable balance between efficiency and accuracy. Although such heuristic estimations are vulnerable to biases, this automatic process cost minimal cognitive resources and thus support fast reaction time compared to controlled, attention process [23]. It allows us to process potentially numerous objects in the environment without becoming overwhelmed with information. MF learning is typically more accurate and effective than simple heuristics. While most heuristics only give a simple estimation of the environment (e.g., using a one-dimensional judgment such as similarity or availability), MF learning can be closer to normative given enough trials or explorations. The mapping between actions and expected rewards is easy to access without much cognitive cost, and the reward estimations are constantly updated based on experience.

1.2.2 Model-based Learning

MB learning utilizes the knowledge of environmental structure. It combines reward-mapping with the structural transitions of states, and allows the agent to "look ahead" for rewards in future steps. For example, consider the act of having lunch. In this scenario, there are two states: a "no-food state" and a "with-food" state. The state-specific action "get food" operates as a transition into the "with-food" state. Now we can conduct the action that is only available from the "with-food" state: eat. Although the first action ("get food") does not directly provide the satisfaction of lunch, it is still gratifying because it leads to a rewarding state

("with-food"). In this case, the MB system assigns the reward of the first action based on the reward of the second state which the action transitions to. MB learning and the state-to-state association it makes, albeit overlapping with the MF system in their involved brain circuits, uses a process of associations that gives MB learning better flexibility but makes it more computationally costly.

Computational Details Both MB and MF learning have been formalized within the framework of reinforcement learning. Typically, a reinforcement learning model simulates a learning scenario with four components: the states of the environment, the actions, the transition function from one state to another, and the reward of each state (see [24]). The actions provide transitions from one state to another. The value of an action, or the Q value, is defined by its long-term discount cumulative reward. The action values may consider a combination of implicit and explicit factors such as personal preference, external rewards, beliefs and principles, etc. Reinforcement learning processes encode the acquired value specific to each state-action pair and associated updating of said value every time a choice is made. The value of an action in a certain state is updated based on its previous value, and plus the product of learning rate and the prediction error. The prediction error is the difference between the actual reward gained after the action and the previously estimated action value. An agent's learning process of the environment, which is the computation of values of actions at each given state, is updated over time; this typically occurs in a trial-by-trial manner in most cognitive experiments. Agents update the expected value of choices using the feedback of previous choices. This

process is dependent on the learning rate, which determines the influence each previous feedback has on future choices. Both the MB system and the MF system require the agent to learn action values under each state, represented by $Q(s, a)$, where Q is the value of the action a under specific states. During MF learning, the Q value of an action under state s at trial t is updated through the following equation, according to State–Action–Reward–State–Action (SARSA) learning, a classic reinforcement-learning algorithm [25]:

$$Q_{mf}(s_t, a_t) = Q_{mf}(s_t, a_t) + \alpha * (r_t + Q_{mf}(s_{t+1}, a_{t+1}) - Q_{mf}(s_t, a_t))) \tag{1.1}$$

t refers to the trial number and r refers to the reward at the next time step, after taking action at time t. The Q value is updated at the $t+1$ time-step (when the action at is taken), knowing the difference between $Q(s_{t+1}, a_{t+1})$ and $Q(s_t, a_t)$. Here, we use a bootstrapping method to estimate the true value of $Q_{mf}(s_t, a_t)$ from the current reward as well as the value of the subsequent state $Q_{mf}(s_{t+1}, a_{t+1})$, compared to using the real reward observation. Bootstrapping allows for faster convergence of expected reward estimation. This model maximizes the expected long-term value of the total reward over all successive steps, thus leading the agent to better rewards.

In contrast to the MF approach, the Bellman equation allows agents look up the reward of the consequence state one step ahead, via the transition functions of the environment, to estimate Q values in a MB approach :

$$V_{mb}(s_t, a_t) = \sum_{s'} p(s'|s_t, a_t) * max_{a'}(r_t + V_{mb}(s', a')) \tag{1.2}$$

$$Q_{mb}(s_t, a_t) = \sum_{s'} p(s'|s_t, a_t) * max_{a'}(r_t + Q_{mb}(s', a')) \tag{1.3}$$

Here, s' and a' are the consequence state and action, similar to s_{t+1} to a_{t+1} but in a forward approach. $p(s'|s_t, a_t)$ represents the probability to transit into s' at state s_t , after taking the action a_t. The value of $Q_{mb}(s_t, a_t)$ takes into account this future reward, $Q_{mb}(s', a')$, since the agent is able to predict s' given the transition probabilities. This equation would be applied iteratively if there is more than one step. Although different algorithms schedule the MB value update differently, these values still rely on the basic value estimation from the MF system. Given enough number of trials, the Q value estimation in the two systems would eventually agree in a stable environment.

It is worth mentioning that full-fledged MB control is near impossible in real-life since the branches of a decision tree would expand so quickly that evaluating the transitions of the whole decision tree, even for just a few steps of decisions, would become impractical, especially considering the limited capacity of our working memory (see [26,27]). The complexities of real-life problems, along with the insubstantial rewards associated with most of them, make the MB control non-optimal for some trivial decisions, especially due to the amount of cognitive resources it would re-

quire. For example, agents may use a hybrid of both MF and MB learning, but the contributions of the MB portion may decay when there are not enough incentives to warrant more MB learning [28]. In addition, the level of cognitive resources we can employ is limited by the capacity of working memory [29, 30]. In this study, both tasks involve choices across two steps, so the MB control only has to look ahead one step as to not demand excessive resources while still being more costly compared to MF decision-making.

1.3 Cognitive Control and Conflicts

Cognitive control refers to a set of cognitive mechanisms that are responsible for deliberately adjusting the configuration of our cognitive systems in order to enhance performance in complicated tasks and to maximize the attainment of reward [31], especially when there is response competition across dissociable cognitive systems [32]. There are roughly three aspects of cognitive control: (a) conflict monitoring, which is the (ability to) access task information and current situations; (b) conflict specification, which is the initial identification of a conflict; and (c) conflict regulation, which is the deployment of cognitive resources to resolve conflict [31].

Cognitive control monitors conflicts and assigns control signals to the engagement of a task when needed. When conflicts are detected and specified, the dorsal anterior cingulate cortex (dACC) and related brain circuits [10, 33] trigger top-down cognitive control over information processing to push back and re-examine responses. Meanwhile, the subthalamic nucleus (STN) increases the decision threshold to allow

for more reaction time [34]. If necessary, goal-directed processing will dominate over autonomous processing at the cost of additional cognitive resources.

Although the demand for control is likely evaluated through the dACC's monitoring for conflict and specification [35], it is not yet clear what deploys such additional control. In tasks where conflicts are triggered by a stimulus' perceptual properties (e.g. the Stroop test), conflict resolution relies on the prefrontal cortex (PFC) through mechanisms such as biased processing of perceptual information [19] or cortical amplification of task-relevant information [36]. However, these studies do not address the resolution of conflict between the MB and the MF learning systems directly, because the conflict happens on the stimulus level.

There are several two-step decision-making tasks that are able to differentiate the MB and the MF systems by creating different valuations between the two systems using careful task structures. We will discuss decision conflicts in the widely adopted Two-step task designed by Daw and colleagues daw2011model and a later adaptation of it, the Rocket task [37], to illustrate the cognitive conflict and potential competitions between MB and MF learning in valuation and decision.

In the Daw Two-Step task ([11]; Fig.1.1), participants needed to make two choices, represented by Tibetan characters, in each trial. The first stage included two choices, and each choice probabilistically lead to one of the two second-stage states. Each second stage included two choices and each of the four choices provided an independent probabilistic reward. The reward probability of each second-stage choices drifted slowly.

In the trials with rare transitions where the choice leads to a transition to the

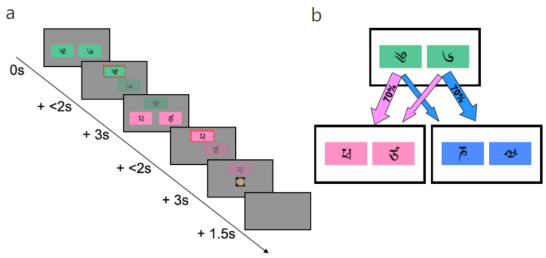

Figure 1.1: From Daw et al., 2011. (a) Timeline of events in trial. A first-stage choice between two options (green) probabilistically lead to a second-stage choice (pink).(b) Probabilistic State transition. Each first-stage choice had a more 'associated' outcome. Upon making the specific choice, 70% of the time the participant was directed to the more associated state and was directed to the less associated state the remaining 30% of the time.

less associated state, the MF Q value attributes the outcome to the selected first stage choice, while the MB Q value, knowing what the second stage state is usually associated with, attributes the outcome to the other first stage choice. Thus, on average there is a conflict in the Q values between the MB decision and the MF decision in the trials right after a rare-transition trial. For example, if the agent chooses the left character of the green pair in the first step and transits to the blue character pair, MF learning will backpropagate the reward from the blue pair to the left character of the green pair. In contrast, MB learning system would instead update the estimated Q value of the right character more than it would update the left character of the green pair, knowing that the blue pair is more commonly associated with the right character of the green pair. Q value conflict is represented by the difference between the estimated Q values of first-stage options from MB and MF learning. In some cases, the difference is so large that the MB system prefers one choice while the MF system prefers the other. The two systems will then have an action conflict due to their opposite "choices".

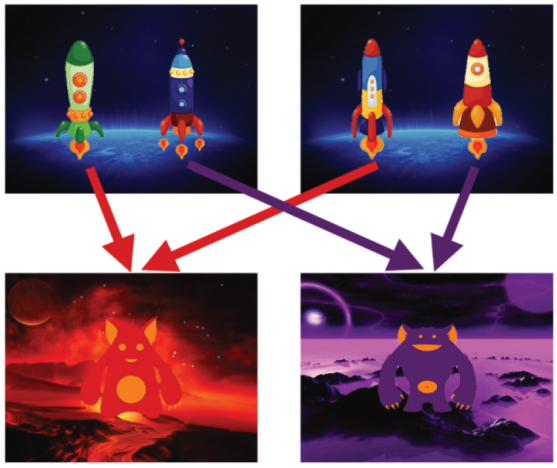

Figure 1.2: From Kool et al.,2016. The transition structure of the task. In each trial, one pair of rockets (from a set of two rocket pairs) randomly appeared and the participant was asked to choose one of the rockets from the apparent pair. The rocket transitioned to the matching planet, where the participant needed to press the spacebar to collect either space treasure (positive points) or antimatter (negative points). The points on each planet changed independently and slowly. Subjects were asked to get as many points as possible.

To quantify these Q value conflicts and action conflicts, we calculate Q value conflict as:

$$Conflict_{value} = |(Q_{mb}(choice_1) - Q_{mb}(choice_2)) - (Q_{mf}(choice_1) - Q_{mf}(choice_2))| \tag{1.4}$$

while the action conflict exists when:

$$(Q_{mb}(choice_1) - Q_{mb}(choice_2)) * (Q_{mf}(choice_1) - Q_{mf}(choice_2)) < 0 \tag{1.5}$$

In the reinforcement-learning simulation of the well-established Daw's Two-step task, Kool and colleagues found no correlation between the strength of MB

control and the actual reward given the reward structure [37, 38]. Participants showed a hybrid strategy of using both MB and MF learning in this task. However, increasing the weight of MB learning did not enhance the averaged reward earned over different values of other learning parameters. This simulation showed a lack of incentive to use MB control in the original Two-step task [37, 38].

As a solution, Kool and colleagues developed the rocket task, in which using more MB learning gives more reward and increases the reward rate, providing incentive for greater use of the MB system. Each trial of the task presented the participant with a pair of rockets, one of which they have to choose. The pair of rockets was randomly presented from a set of two pairs of rockets and the position of the rocket (left versus right) was random and equally probable. Each rocket deterministically leads to one of two planets (Fig.1.4). The planets gave a probabilistic payoff undergoing an independent slow drift.

Similar to the Two-step task, the conflicts between the MB and the MF systems lie in the valuation and choices in the first stage. Since the two rocket pairs showed up randomly across trials, some trials had the same rocket pair as the one in the previous trial, while some trials had a switched rocket pair. In the trials with a switched first-stage choice pair compared to the previous trials, the MB system evaluates the current first-stage choices differently from the MF system. This is because the MB Q values follow the updated second-stage Q value (the Q value of the planet) and include it to the valuation of first-stage options, but the MF Q values are not shared between pairs of rockets, thus the Q_{mf} not updated from the previous trial. This difference potentially leads to a Q value conflict between the

two learning systems and in some cases, evolves into an action conflict of the final decision in some trials.

Similar to the original Two-step task, we define Q value conflicts as:

$$Conflict_{value} = |(Q_{mb}(state_{stage1}, choice_1) - Q_{mb}(state_{stage1}, choice_2)) - (Q_{mf}(state_{stage1}, choice_1) - Q_{mf}(state_{stage1}, choice_2))| \quad (1.6)$$

Notice that there are two possible states in the first stage of the decision, representing two choice pairs.

Accordingly, the action conflict exists when:

$$(Q_{mb}(state_{stage1}, choice_1) - Q_{mb}(state_{stage1}, choice_2)) * (Q_{mf}(state_{stage1}, choice_1) - Q_{mf}(state_{stage1}, choice_2)) < 0 \quad (1.7)$$

1.4 Computational Modeling of Decisions

Decision-making involves both valuation and the conversion of value to choice. While reinforcement learning assigns and updates values to a set of actions, these values convert into a choice. Static models of discrete choices are used to explain the decision of a choice from a set of distinct alternatives, given their relative value or utilities [39]. The model links quantitative continuous utilities of choices to discrete

behaviors of choosing through a squashing function, usually a logistic function [32], and is usually computed as follows in a single binary-choice system:

$$P(choice_1) \sim logistic(V_1 - V_2) \tag{1.8}$$

Here, V_1 and V_2 represent the quantified values of the two choices in a binary decision. The squashing function takes in the differences between the values, and reduces the difference into a choice probability from 0 to 1. Discrete choice models have had many successful applications beyond simple choices, such as to simulate social influence on decision making [40], or to explain how criminals make the decision to attack a particular location [41]. However, since the model focuses on explaining the outcome of decision, but not the decision process itself, it does not provide more detailed information on the latent variables of a decision process, and certainly does not account for reaction time. To overcome these limitations, the Evidence-Integration (EA) theory of decisions was developed to describe the decision process over choices [14]. The EA model proposes that a decision is made by incrementally collecting information over time until reaching a decision threshold, and has received increased attention in the past decade for its capacity to explain accuracy and reaction time, and even neural data (e.g. [42]).

One of the quantitative models inspired by EA, the Drift Diffusion Model (DDM; see [12, 13], represents the decision between two options amidst the gradual accumulation of evidence for one versus the other (Fig.1.5). This model has

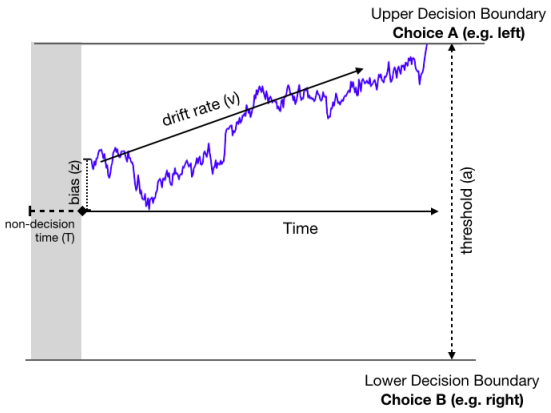

Figure 1.3: Drift Diffusion Model and it's parameters. Evidence towards either decision boundary is collected at the speed of the drift rate v, with a start point bias z. The reaction time of the decision is the duration of information collection in addition to the non-decision time.

been widely validated across many types of experimental data, including perceptual discrimination [15, 16], go-no-go tasks [17], and value-based choice tasks [18]:

$$\Delta E = v * \Delta t + cdW \tag{1.9}$$

Here, E represents the quantity of relative evidence between two choices, and v represents the drift-rate, a free parameter representing the speed of evidence collection. cdW is the Gaussian noise during the drifting process. A decision is made when E reaches one of the two decision thresholds in a binary choice. In Figure:1.5, evidence accumulated and hit the upper decision boundary, upon which the agent will select Choice A.

The drift rate of information collection is proportional to the relative value of the choices [43]. The further apart the values of the two options, the faster relative

evidence is collected towards the more valuable option:

$$v = v' * (V_1 - V_2) \tag{1.10}$$

Again, V_1 and V_2 represent the quantified values of the two choices. v' is the baseline coefficient drift rate.

1.5 Bayesian Modeling

Another methodological advance in the modeling of evaluation, decision, and learning is the use of Bayesian modeling [44]. Computational cognitive modeling allows for specification and testing of various mathematical formulations of mental processing in defined decision tasks, and provides quantitative fitting of a wide range of, sometimes latent, cognitive features. Bayesian modeling [45] predicts the distribution of parameters rather than point-wise value. The concept of Bayesian modeling is based on Bayes' rule. It computes the posterior probability of an event based on a prior probability and a likelihood function, derived from the observed data. The probability distributions are updated according to observations. Compared to traditional statistical methods that estimate one-dimensional metrics, Bayesian inference estimates the probability distribution of the parameters values conditional on the data. For example, in the simple case of a Bayesian linear model, specified priors of the coefficient of interest, along the likelihood of data calculated using the Bayes rule, together decides the posterior distribution of the coefficient. Hierarchical

modeling (or multi-level modeling) is widely used to characterize group parameters in the Bayesian modeling of cognition and behaviors.

Hierarchical modeling has shown profound advantages over single-level modeling in understanding individual observations that are drawn from a joint probability distribution, when such distribution is formed by both the population distribution and the subject-level distribution [46]. Compared to traditional modeling that assumes the observations are independent of each other, adding the group-level distribution benefits our interpretation of cognitive processes in many ways, such as to combine multiple levels of different cognitive processes to account for a single observed behavior, or to conceptually involve and integrate multiple-levels of separate cognitive models [47]. In addition to wider application, hierarchical modeling also shows better performance than individual-level models, including better parameter recovery from simulation, and better prediction accuracy (e.g. [48, 49]).

1.6

In the Daw Two-Step task ([11]; Fig.1.1), participants needed to make two choices, represented by Tibetan characters, in each trial. The first stage included two choices, and each choice probabilistically lead to one of the two second-stage states. Each second stage included two choices and each of the four choices provided an independent probabilistic reward. The reward probability of each second-stage choices drifted slowly.

In the trials with rare transitions where the choice leads to a transition to the

less associated state, the MF Q value attributes the outcome to the selected first stage choice, while the MB Q value, knowing what the second stage state is usually associated with, attributes the outcome to the other first stage choice. Thus, on average there is a conflict in the Q values between the MB decision and the MF decision in the trials right after a rare-transition trial. For example, if the agent chooses the left character of the green pair in the first step and transits to the blue character pair, MF learning will backpropagate the reward from the blue pair to the left character of the green pair. In contrast, MB learning system would instead update the estimated Q value of the right character more than it would update the left character of the green pair, knowing that the blue pair is more commonly associated with the right character of the green pair. Q value conflict is represented by the difference between the estimated Q values of first-stage options from MB and MF learning. In some cases, the difference is so large that the MB system prefers one choice while the MF system prefers the other. The two systems will then have an action conflict due to their opposite "choices".

To quantify these Q value conflicts and action conflicts, we calculate Q value conflict as:

$$Conflict_{value} = |(Q_{mb}(choice_1) - Q_{mb}(choice_2)) - (Q_{mf}(choice_1) - Q_{mf}(choice_2))| \tag{1.11}$$

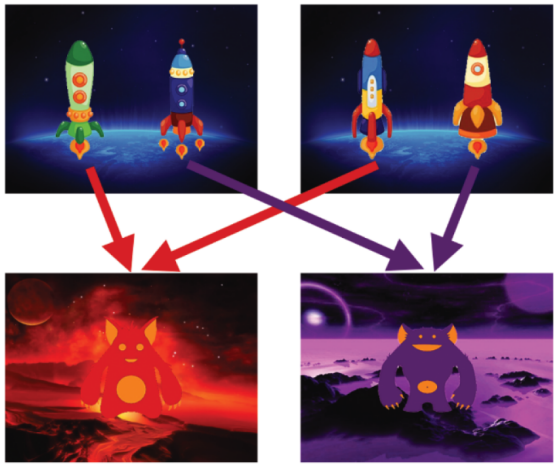

Figure 1.4: From Kool et al.,2016. The transition structure of the task. In each trial, one pair of rockets (from a set of two rocket pairs) randomly appeared and the participant was asked to choose one of the rockets from the apparent pair. The rocket transitioned to the matching planet, where the participant needed to press the spacebar to collect either space treasure (positive points) or antimatter (negative points). The points on each planet changed independently and slowly. Subjects were asked to get as many points as possible.

while the action conflict exists when:

$$(Q_{mb}(choice_1) - Q_{mb}(choice_2)) * (Q_{mf}(choice_1) - Q_{mf}(choice_2)) < 0 \qquad (1.12)$$

In the reinforcement-learning simulation of the well-established Daw's Two-step task, Kool and colleagues found no correlation between the strength of MB control and the actual reward given the reward structure [37, 38]. Participants showed a hybrid strategy of using both MB and MF learning in this task. However, increasing the weight of MB learning did not enhance the averaged reward earned over different values of other learning parameters. This simulation showed a lack of incentive to use MB control in the original Two-step task [37, 38].

As a solution, Kool and colleagues developed the rocket task, in which us-

ing more MB learning gives more reward and increases the reward rate, providing incentive for greater use of the MB system. Each trial of the task presented the participant with a pair of rockets, one of which they have to choose. The pair of rockets was randomly presented from a set of two pairs of rockets and the position of the rocket (left versus right) was random and equally probable. Each rocket deterministically leads to one of two planets (Fig.1.4). The planets gave a probabilistic payoff undergoing an independent slow drift.

Similar to the Two-step task, the conflicts between the MB and the MF systems lie in the valuation and choices in the first stage. Since the two rocket pairs showed up randomly across trials, some trials had the same rocket pair as the one in the previous trial, while some trials had a switched rocket pair. In the trials with a switched first-stage choice pair compared to the previous trials, the MB system evaluates the current first-stage choices differently from the MF system. This is because the MB Q values follow the updated second-stage Q value (the Q value of the planet) and include it to the valuation of first-stage options, but the MF Q values are not shared between pairs of rockets, thus the Q_{mf} not updated from the previous trial. This difference potentially leads to a Q value conflict between the two learning systems and in some cases, evolves into an action conflict of the final decision in some trials.

Similar to the original Two-step task, we define Q value conflicts as:

$$Conflict_{value} = |(Q_{mb}(state_{stage1}, choice_1) - Q_{mb}(state_{stage1}, choice_2)) - \quad (1.13)$$
$$(Q_{mf}(state_{stage1}, choice_1) - Q_{mf}(state_{stage1}, choice_2))|$$

Notice that there are two possible states in the first stage of the decision, representing two choice pairs.

Accordingly, the action conflict exists when:

$$(Q_{mb}(state_{stage1}, choice_1) - Q_{mb}(state_{stage1}, choice_2)) * \quad (1.14)$$
$$(Q_{mf}(state_{stage1}, choice_1) - Q_{mf}(state_{stage1}, choice_2)) < 0$$

1.6.1 Computational Modeling of Decisions

Decision-making involves both valuation and the conversion of value to choice. While reinforcement learning assigns and updates values to a set of actions, these values convert into a choice. Static models of discrete choices are used to explain the decision of a choice from a set of distinct alternatives, given their relative value or utilities [39]. The model links quantitative continuous utilities of choices to discrete behaviors of choosing through a squashing function, usually a logistic function [32], and is usually computed as follows in a single binary-choice system:

$$P(choice_1) \sim logistic(V_1 - V_2) \tag{1.15}$$

Here, V_1 and V_2 represent the quantified values of the two choices in a binary decision. The squashing function takes in the differences between the values, and reduces the difference into a choice probability from 0 to 1. Discrete choice models have had many successful applications beyond simple choices, such as to simulate social influence on decision making [40], or to explain how criminals make the decision to attack a particular location [41]. However, since the model focuses on explaining the outcome of decision, but not the decision process itself, it does not provide more detailed information on the latent variables of a decision process, and certainly does not account for reaction time. To overcome these limitations, the Evidence-Integration (EA) theory of decisions was developed to describe the decision process over choices [14]. The EA model proposes that a decision is made by incrementally collecting information over time until reaching a decision threshold, and has received increased attention in the past decade for its capacity to explain accuracy and reaction time, and even neural data (e.g. [42]).

One of the quantitative models inspired by EA, the Drift Diffusion Model (DDM; see [12, 13], represents the decision between two options amidst the gradual accumulation of evidence for one versus the other (Fig.1.5). This model has been widely validated across many types of experimental data, including perceptual discrimination [15, 16], go-no-go tasks [17], and value-based choice tasks [18]:

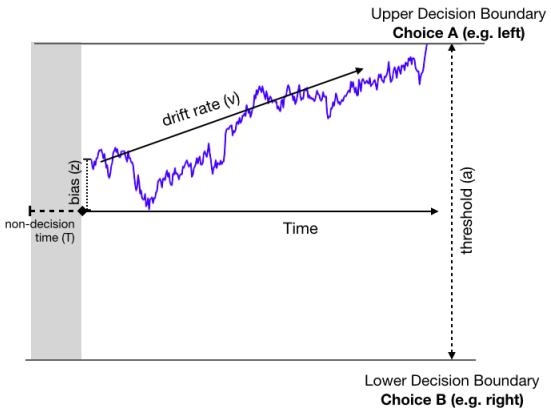

Figure 1.5: Drift Diffusion Model and it's parameters. Evidence towards either decision boundary is collected at the speed of the drift rate v, with a start point bias z. The reaction time of the decision is the duration of information collection in addition to the non-decision time.

$$\Delta E = v * \Delta t + cdW \tag{1.16}$$

Here, E represents the quantity of relative evidence between two choices, and v represents the drift-rate, a free parameter representing the speed of evidence collection. cdW is the Gaussian noise during the drifting process. A decision is made when E reaches one of the two decision thresholds in a binary choice. In Figure:1.5, evidence accumulated and hit the upper decision boundary, upon which the agent will select Choice A.

The drift rate of information collection is proportional to the relative value of the choices [43]. The further apart the values of the two options, the faster relative evidence is collected towards the more valuable option:

$$v = v' * (V_1 - V_2) \tag{1.17}$$

Again, V_1 and V_2 represent the quantified values of the two choices. v' is the baseline coefficient drift rate.

1.7 Metacognition on Anxiety

In addition to looking at the competition between the MB control and the MF control, the data we use in this thesis presents an opportunity to understand how the competition between the two systems interacts with individual differences in OCD symptoms. Although previous work has suggested OCD symptoms are related to reduced MB control, goal-directed learning (reviewed by [50]) and decision-making (e.g. [51, 52]), it is not clear which specific aspects of OCD drive this relationship.

In a recent model of OCD [50], compulsivity modulated the deficits in the MB system. High compulsivity was specifically related to a preference towards the automatic MF system for choices over the deliberative MB system [53]. In order to identify the individual trait that captures the unique cognitive compulsivity component that distinguishes OCD from other anxiety disorders, Wells and colleagues proposed the metacognitive model of OCD [54–56]. In this model, compulsivity is induced and maintained by the metacognitive beliefs about the meaning and consequences of intrusive thoughts. Accordingly, the Meta-Cognitions Questionnaire [57] was designed to measure theses metacognitive beliefs and became a popular trans-

diagnostic inventory. Its five subscales address different aspects of metacognitive traits including anxiety, worry, and proneness to obsessional symptoms. Among all five subscales, it is identified that the monitoring of one's own thinking, as measured by the Cognitive self-consciousness (CSC) subscale, is the crucial part that differentiate people with OCD diagnosis from people with General Anxiety Disorder diagnosis [57]. This constant, real-time self-awareness of thinking process may cause prolonged decision-making due to deprived attention resources.

In this study, we used an unpublished dataset that includes participants' responses to five questionnaires related to worry: Metacognition Questionnaire-30 (MCQ30), Intolerance to Uncertainty Scale (IUS), The State-Trait Anxiety Inventory - Trait (STAI_T), The State-Trait Anxiety Inventory - State (STAI_S), and Penn State Worry Questionnaire (PSWQ). These questionnaire answers allowed me to conduct an exploratory analysis, to look into the mechanism between metacognition and MB learning, and to ask the following questions: a) what is it about OCD that drives worse decision performance? and b) how does OCD relate to MB learning and decision boundary?

Aims This thesis presents a computational model that examines conflict adaptation between the two learning systems to understand how the MB controller and the MF controller react to valuation dissonance. I used an integrated decision model to combine both reinforcement learning algorithms and DDM in order to explain the conflict-resolving mechanism of the two learning systems behind their incorporated decisions, particularly to ask the following questions:

1) How does the decision process adapt when there is a conflict

between MB and MF learning? In traditional conflict tasks such as the Stroop test and the Flanker test, the data often shows prolonged reaction time and increased errors in trials with high conflict than it does in trials with low conflict. The decrease in performance often declines after training, once conflict adaptation kicks in and promotes goal-directed processing. For example, experiments have shown extended decision thresholds related to mPFC activity during conflicts [34] to allow extra processing time and to reduce the error rate. This extension allows a slower and more controlled attentional processing to happen.

However, conflicts in these experiments happen at the stimulus level, and not the value level. In other words, these experiments are essentially looking at selective processing, but not the interaction between the two learning systems during their conflict. In this study, we aimed to address the response competition between the habitual and goal-directed learning systems directly. Several existing models (e.g. [7, 8]) have proposed that the MB system, closely related to goal-directed control, benefits from extra allocation of attention resources when needed and guides model-free, habitual behaviors in a top-down approach. However, studies of individual differences suggest the otherwise: the MB system almost always yield in while the MF system tends to stay in a more stable way (e.g. [58, 59]). Those findings of compromised MB learning go along with the idea that MF learning is the default and autonomous response and thus is less likely to change.

This study quantifies the two systems' propensities in the decision process using a combination of RL model and DDM, allowing us to look at the shifts in latent decision parameters in both learning systems during a conflict. Among the

four DDM parameters, decision drift rate is modeled separately for the MB and the MF controllers, representing separate evidence collection speed, each proportional to the Q value differences between choices in their corresponding system; decision boundary is shared across the two systems, and controls the decision "deadline". There are three possibilities of conflict adaptation within the decision parameters: change in drift rate in one or both systems, change in decision boundary, or no change in either parameters. While previous work does not allow us to identify the specific changes, it is intuitive to expect some decrease in at least one of the two systems' drift rates, especially the MB drift rate given how it is prone to changes, in a trial with action conflict. We also expected to see an increase in decision boundary to provide more time for conflict resolution [34]. Additionally, we tested if Q value conflicts alone without action conflict can induce conflict adaptation in the two systems.

2) How do the two learning systems react differently during conflicts in the Rocket Task v.s. the original Two-step Task? Previous studies using the Two-step task have found that participants use a hybrid strategy of both MB and MF learning (e.g. [60, 61]), even though the MB system is not additionally incentivized [37]. Kool and colleagues' Rocket task, an adapted version of the Two-step task, provided more rewards to MB decisions [37]. Their experiments showed that participants had a higher propensity of MB learning in the Rocket task than in the Two-step task [38]. This study aimed to investigate a further question: does the mechanism of conflict adaptation change when the task structure differ? When the MB learning becomes more rewarding in the Rocket task (compared to the Two-step

task), will the conflict-related changes in the two systems differ?

3) How do previous choices bias a decision? Choice biases towards the previous chosen option affect decisions in a way that is independent of the MB controller and the MF controller. Such biases are the result of response "stickiness" or perseveration, the measure of participants' preference to stick with or alternate their previous response in the last trial. Quite a few studies have looked into factors that determine the strength of this choice perseveration, such as anxiety (e.g. [62]). Failure to specify choice perseveration may result in biased parameter estimations [63]. While researchers commonly taken this effect into consideration when they model a decision, no one has closely identified its mechanism within the decision process. Previous studies often modeled the perseveration effect as an extra term in the updating of Q value in RL algorithms or as an addition to the value difference between two choices in a static discrete choice model(e.g. [38, 60, 61]). Unlike these studies, we directly modeled the perseveration in the two latent decision parameters, starting point bias and drift rates. The starting point bias represents how much less evidence is needed for one choice, relative to the other choice, while the drift rate, again, represents the speed of information accumulation. These two parameters allowed us to ask more specifically about decision biases from previous choices: do they influence future choices through biased speed of information collection, or a change in starting point evidence towards one direction?

4) How does metacognition on anxiety modulate conflict adaptation? We conducted an exploratory analysis on the modulation relationship between metacognition on anxiety and the conflict adaption in the two systems. We

expected that CSC trait of metacognition, the crucial trait of OCD, will influence decision-making, via one or multiple decision parameters and their conflict adaptation. Previous studies have consistently shown that the MB system, but not the MF system, are prone to changes caused by OCD and compulsivity [64]. Along these lines, we expected a decrease in the MB drift rate in participants with higher CSC subscale scores of the Meta-Cognitions Questionnaire (MCQCSC), as a result of slower evidence collection due to cognitive overload caused by excessive metacogition. Considering that OCD symptoms are often associated with prolonged reaction time but undermined performance accuracy [65], we expected that higher CSC results in smaller conflict adaption, such as a smaller conflict-related drift rate change. Additionally, we expected to see extended decision boundaries related to compulsive urges and regret [50] in participants with higher MCQCSC scores.

Chapter 2: Methods

2.1 Datasets

We included four datasets (Table 2.1) in this study, three from previous studies using the two tasks [37, 66], and one unpublished dataset previously collected by Dr. Alec Solway. The [66] dataset and the unpublished datasets were collected on campus where participants performed the decision task on a lab computer, and the two [37] datasets were collected through Amazon Mechanical Turk. In all datasets, participants who have more than 20% trials with reaction time ¡ 0.15 sec or a missing choice in either stage were excluded from the analysis. All trials needed to be complete in both stages to be included in the learning process.

Two-step task	[66]	[37]	Unpublished data
Original Sample Size	(n=119)	(n=206)	(n=92)
Effective Sample Size	(n=115)	(n=175)	(n=84)
Rocket task	[37]	/	/
Original Sample Size	(n=199)		
Effective Sample Size	(n=179)		

Table 2.1: Sources and sample sizes of the four datasets.

2.1.1 Unpublished Dataset

92 participants were recruited around Princeton University campus completed the Two-step task. The experiment protocol was similar to previous studies that used this task (e.g. [1,37,38,66]). Participants first went through a tutorial session with practice trials, where they were instructed to find the best choice to gain more rewards in this two-stage binary decision task. Each first-stage choice lead to each of the two second-stage choice pairs with a pre-determined probability. The transition probabilities were set to be 0.7/0.3, identical with the original experiment. Each trial started with a pair of choices represented by two boxes with Tibetan characters. Once participants selected a choice from the first-stage choice pair, the chosen box was highlighted for one second. Then, the resulting second-stage choice pair appeared below the first-stage choice while the non-chosen first-stage choice box disappeared. Once participants performed the second-stage choice, their choice was again highlighted for one second. Finally, the reward of that trial, either 0 or 1 showed up for one second. The reward rate of each second-stage choices drifted with a Gaussian random walk. There was no time limitation for participants to choose between presented choices. After the tutorial, each participant completed 201 trials of the experiment and then filled out a set of questionnaires on the computer, including Metacognition Questionnaire-30 (MCQ30), Intolerance to Uncertainty Scale (IUS), The State-Trait Anxiety Inventory - Trait (STAI_T), The State-Trait Anxiety Inventory - State (STAI_S), and Penn State Worry Questionnaire (PSWQ).

2.2 Learning algorithms

Both the MB and the MF learning systems have been formalized within the framework of reinforcement learning. In both tasks, we used the classic SARSA algorithm to simulate the MF learning process, a standard way of modeling Two-step task and Rocket task (e.g. [11, 28, 37, 66, 67]). First, we updated the value estimation of state-action pairs in the first stage of the tasks based on the difference between the following second-stage Q value and the current estimation of first-stage Q value:

$$Q_{mf}(state_{stage1}, choice_{stage1}) = Q_{mf}(state_{stage1}, choice_{stage1}) + \alpha * [Q_{mf}(state_{stage2}, choice_{stage2} - Q_{mf}(state_{stage1}, choice_{stage1})] \quad (2.1)$$

Then, the value update took a further step to propagate back the reward value at the end of the second stage, and included it in the estimation of first stage Q values at a discounted rate λ:

$$Q_{mf}(state_{stage1}, choice_{stage1}) = Q_{mf}(state_{stage1}, choice_{stage1}) + \lambda * \alpha * (rew - Q_{mf}(state_{stage2}, choice_{stage2})) \quad (2.2)$$

Finally, the Q values of state-action combinations in the second stage of the

tasks are updated as:

$$Q_{mf}(state_{stage2}, choice_{stage2}) = Q_{mf}(state_{stage2}, choice_{stage2}) + \alpha * (rew - Q_{mf}(state_{stage2}, choice_{stage2})) \tag{2.3}$$

Here, learning rate *alpha* and discount rate *lambda* originally ranged from $-inf$ to inf, and were transformed using an logistic function during model fitting to ensure they had values between 0 and 1. In the first stage of the task, there were two possible first-stage states (the two rocket pairs), each having two choices (two rockets), while the Two-step task had one first-stage state (one pair of Tibetan characters), with two choices (two Tibetan characters). In the second stage of the task, the rocket task had two possible second-stage states (the two planets), but only one choice (pressing the space bar), while the Two-step task had two second-stage state (two pairs of Tibetan characters), each with two choices (two characters).

The first stage Q values in the MB system was directly linked to the second stage Q values through use of the Bellman equation:

$$Q_{mb}(state_{stage1}, choice_{stage1}) = \sum p' * Q_{mb}(state'_{stage2}, choice'_{stage2}) \tag{2.4}$$

$state_{stage2'}$ and p' represents the possible second-stage states given the first stage state and choice, and their corresponding transition probability p'. $choice_{stage2'}$

represents the optimal choice in the given state.

Then, since the second stage is last stage, there was no more information to plan ahead. In this case, the MB system shared the same second-stage Q values with the MF system, :

$$Q_{mb}(state_{stage2}, choice_{stage2}) = Q_{mf}(state_{stage2}, choice_{stage2}) \tag{2.5}$$

All non-selected choices in each stage decayed towards initial Q values $Q_{initial}$ at a speed of α:

$$Q(state', choice') = Q(state', choice') + \alpha * (Q_{initial} - Q(state', choice'))) \tag{2.6}$$

2.3 Decision-Making Model

We fitted all models in a hierarchical setup, implemented in the statistical programming language Stan [68] in C++ environment. The program performs Markov chain Monte Carlo (MCMC) using the No-U-Turn sampler to find inferences for all parameters. As previously discussed, there are two types of conflict between the two learning systems: the Q value conflict and the action conflict. The former addresses the difference between Q value estimations, or the difference between the two systems in how much more Q value one choice has over the other. The latter addresses the difference in the two systems' preferred choices. In both models, we took into

account the switching of first stage stimuli in the Rocket task by adding the sw parameters, where sw is 1 when the rocket pair in the previous trial is different from the current pair in the first stage of the task; sw is 0 when the rocket pair stays the same. In the Two-step task, sw is 0. This switch parameter adjusts for participants' possible change in decision due to the repetition of first-stage stimuli.

Stan relies on a smooth gradient to generate MCMC proposals, which is broken by the presence of the action conflict indicator variable (see below). In order to optimize the model fitting process, we used a two-stage fitting procedure: first we determined the learning rate α and the discount rate λ of each participants using a softmax decision model, and then we identified other parameters using the drift diffusion model.

2.3.1 Softmax Models

We first ran a set of static choice models to determine the learning rates α and the discount rate λ at an individual level:

$$\begin{aligned} choice_{stage1} \sim & bernoulli_logit((beta_{mb} + beta_{mb,sw} * sw)* \\ & [Q_{mb}(state_{stage1}, choice_1) - Q_{mb}(state_{stage1}, choice_2])+ \\ & (beta_{mf} + beta_{mf,sw} * sw)* \\ & [Q_{mf}(state_{stage1}, choice_1) - Q_{mf}(state_{stage1}, choice_2])+ \\ & p_{action}) \end{aligned} \tag{2.7}$$

In this model, the continuous Q value differences in both systems are squashed into binary choices. *beta* parameters adjust the shape of the squashing curve. p_{action} models action stickiness, or the participant's preference to stay with the previous choice, which is commonly observed in this kind of decision tasks. This model does not fit reaction time or latent decision parameters. We took the median of fitted α and λ values of each participant to pass them into the following reaction time models.

2.3.2 Drift Diffusion Models

The integrated DDM model is designed to quantitatively understand decision boundaries, evidence collecting speed and start-point bias of decisions with choice valuations estimated through reinforcement learning algorithms. In this model, when the MB and the MF learning processes decide towards a consistent direction, their drift rates add up to reach a faster decision. However, when the two systems prefer different choices (i.e. there is an action conflict), the drift rates would cancel out one another.

Combining with the classical DDM model, we have the following model using the *rstan* package in R environment to fit parameters with participants' choices and reaction time through the *Stan* program [68]:

$$rt \sim wiener(boundary, \tau, p_{bias}, \upsilon_{total}) \tag{2.8}$$

Here, $boundary$, τ, p_{bias} and υ_{total} are boundary separation, non-decision time, bias towards the previous chosen option (or "response stickiness") in starting point, and total drift rate, respectively. Q value is updated at $t+1$ timestep (when the action is taken), knowing the difference between $Q(state_{t+1}, choice_{t+1})$ and $Q(state_t, choice_t)$. Similar to p_{action} in the softmax model, p_{bias} adjust for response perseveration effects on the starting point bias towards the last chosen action. The Wiener function implements the Gaussian random walk process that simulates the decision process of evidence integration, except that the function is on continuous timescale, and does not consider variances in drift rate or non-decision time across trials [43].

υ_{total} is the linear summation of MB drift rate υ_{mb} and MF drift rate υ_{mf}. Independent drift rates υ_{mb} and υ_{mf}, as well as conflict and switch related drift rate changes, are calculated as parameters proportional to the difference between the correspondingly Q values of two presented choices [43]:

$$\begin{aligned} \upsilon_{total} =& \upsilon_{mb} * [Q_{mb}(state_{stage1}, choice_1) - \\ & Q_{mb}(state_{stage1}, choice_2)] + \\ & \upsilon_{mf} * [Q_{mf}(state_{stage1}, choice_1) - \\ & Q_{mf}(state_{stage1}, choice_2)] + p_\upsilon * ap \end{aligned} \tag{2.9}$$

Here, p_υ represents response stickiness in drift rates similar to the p_{bias} above, but it adjusts for the drift rate bias. ap indicates the last chosen first-stage action.

For each of the MB and the MF systems, υ_{mb} and υ_{mf} are a combination of base υ_{mb} and υ_{mf}, $\Delta\upsilon_{mb,conflict,value}$ and $\Delta\upsilon_{mf,conflict,value}$ responding to between-system Q value conflict, $\Delta\upsilon_{mb,conflict,action}$ and $\Delta\upsilon_{mf,conflict,action}$ responding to the between-system action conflict, $\Delta\upsilon_{mb,conflict,interaction}$ and $\Delta\upsilon_{mf,conflict,interaction}$ responding to the interaction between Q value conflict and action conflict, and $\Delta\upsilon_{mb,sw}$ and $\Delta\upsilon_{mf,sw}$ responding to stimuli switch, as discussed above:

$$\begin{aligned}\upsilon_{mb} = \upsilon_{mb,base} + &\Delta\upsilon_{mb,conflict,value} * conflict_{value} + \\ &\Delta\upsilon_{mb,conflict,action} * conflict_{action} + \\ &\Delta\upsilon_{mb,conflict,interaction} * interaction + \\ &\Delta\upsilon_{mb,sw} * sw\end{aligned} \tag{2.10}$$

$$\begin{aligned}\upsilon_{mf} = \upsilon_{mf,base} + &\Delta\upsilon_{mf,conflict,value} * conflict_{value} + \\ &\Delta\upsilon_{mf,conflict,action} * conflict_{action} + \\ &\Delta\upsilon_{mf,conflict,interaction} * interaction + \\ &\Delta\upsilon_{mf,sw} * sw\end{aligned} \tag{2.11}$$

As previously discussed, $conflict_{action}$ is a binary variable that is set to 0 in when there is no between-system action conflict, and 1 when there is. $conflict_{value}$ is calculated as:

$$
\begin{aligned}
conflict_{value} = & |[Q_{mb}(state_{stage1}, choice_1) - \\
& Q_{mb}(state_{stage1}, choice_2)] - \\
& [Q_{mf}(state_{stage1}, choice_1) - Q_{mf}(state_{stage1}, choice_2)]|
\end{aligned} \tag{2.12}
$$

Similar to the drift rates, the first-stage decision boundaries are calculated as

$$
\begin{aligned}
bound = bound_{base} + & \Delta bound_{conflict,value} * conflict_{value} + \\
& \Delta bound_{conflict,action} * conflict_{action} + \\
& \Delta bound_{conflict,interaction} * conflict_{value} * conflict_{action} + \\
& \Delta bound_{sw} * sw
\end{aligned} \tag{2.13}
$$

For the Two-step task, the first stage shares the same model with the rocket task, except that the sw parameter is always set to 0. We fitted the second-stage choices using:

$$
rt2 \sim wiener(bound_{stage2}, \tau, 0.5, \upsilon_{stage2}) \tag{2.14}
$$

The model is structured in a hierarchical form. The group parameters include the decision boundary *bound*, the drift rates in the two systems and each of their conflict-related regressors υ and $\Delta\upsilon$, and the non-decision time τ. Learning rate α

and discount rate λ from the softmax models are passed into the model. Although the efficiency of the system and the propensities of the two systems are confounded in this model, it still provides the relative weights of the MB system and the MF system in the decision-making, represented by the magnitude of their drift rates.

2.3.3 Adding Individual Traits as Regressors

Previous work has shown that the CSC subscale in the MCQ inventory is the key aspect that distinguishes OCD from GAD [57], indicating its crucial contribution towards compulsivity. While metacognition is an overloaded term that has different meanings in different contexts, here we are talking about metacognition as defined by the Cognitive Self-consciousness subscle (MCQCSC) in the MCQ30 questionnaire [57], which represents a tendency to be aware of and monitor thinking.

High compulsivity is largely linked to a deficit in MB learning [53]. To explore how metacognition affects the two learning systems during decision conflicts, we added the MCQCSC regressors as coefficients for participants' z-scored MCQCSC subscale scores to the following parameters: *bound*, υ_{mb}, υ_{mf}, and each of their conflict-related parameters.

It is worth noting that state anxiety, often induced by psychological stress, also attenuates the contribution of MB system, but not the MF system, in learning and decision-making [69,70]. To control for the state effect, we ran another model with a set of state anxiety regressors in addition to the metacognition parameters, to isolate metacognition-induced changes. The state anxiety regressors are modeled as the

coefficients of z-scored STAI_S scores, from the well-established STAI inventory [71] that assesses the presence and severity of participants' current state of anixiety in a reliable and sensitive way.

Both regressors contributed directly to the base parameters and conflict-related parameters, e.g. in the model without controlling for anxiety, we have the base drift rate and valuation conflict related changes:

$$\upsilon'_{mb,base} = \upsilon_{mb,base} + \Delta\upsilon_{mb,mcqcsc} * MCQCSC \tag{2.15}$$

$$\Delta\upsilon'_{mb,Qconflict,value} = \Delta\upsilon_{mb,Qconflict,value} + \Delta\upsilon_{mb,Qconflict,value,mcqcsc} * MCQCSC \tag{2.16}$$

In the model with the STAI_S regressors, we have:

$$\upsilon''_{mb,base} = \upsilon_{mb,base} + \Delta\upsilon_{mb,mcqcsc} * MCQCSC + \Delta\upsilon_{mb,stai_s} * STAI_S \tag{2.17}$$

$$\Delta\upsilon''_{mb,Qconflict,value} = \Delta\upsilon_{mb,conflict,value} + \Delta\upsilon_{mb,Qconflict,value,mcqcsc} * MCQCSC + \Delta\upsilon_{mb,Qconflict,value,stai_s} * STAI_S \tag{2.18}$$

Similarly, we computed for the action conflict regressors and the interaction regressors.

2.3.4 Model Fitting

For each dataset, we fitted a Bayesian hierarchical model implemented in the Stan statistical programming language (Stan Development Team). Each fit had three chains, 10000 iterations each, with the first 1000 iterations as warmup. The Gelman-Rubin $\hat{R}$ statistic was ensured to be less than 1.1 for all variables. A significant effect was determined by if the parameter's central 95% credible interval excludes the critical value of interest (e.g. 0). Stan runs a Markov Chain Monte Carlo on the model specification using the No-U-Turn sampler and provides the posterior estimations of parameters, given behavioral data and specified priors. All model fits had reasonable effective sample sizes given the number of chains we had (Bulk_ESS ¿700 and Tail_ESS¿1200). The priors had an hierarchical structure given that both models were hierarchically set up. All parameters had weakly informative priors. In all four model fits, the priors for group level parameters of $\Delta\upsilon_{mb,conflict,value}$, $\Delta\upsilon_{mb,conflict,action}$, $\Delta\upsilon_{mb,conflict,interaction}$, $\Delta\upsilon_{mb,conflict,value,mcqcsc}$, $\Delta\upsilon_{mb,conflict,action,mcqcsc}$, $\Delta\upsilon_{mb,conflict,interaction,mcqcsc}$, $\Delta\upsilon_{mb,conflict,value,stai_s}$, $\Delta\upsilon_{mb,conflict,action,stai_s}$, $\Delta\upsilon_{mb,conflict,interaction,stai_s}$, $\Delta\upsilon_{mf,conflict,value}$, $\Delta\upsilon_{mf,conflict,action}$, $\Delta\upsilon_{mf,conflict,interaction}$, $\Delta\upsilon_{mf,conflict,value,mcqcsc}$, $\Delta\upsilon_{mf,conflict,action,mcqcsc}$, $\Delta\upsilon_{mf,conflict,interaction,mcqcsc}$, $\Delta\upsilon_{mf,conflict,value,stai_s}$, $\Delta\upsilon_{mf,conflict,action,stai_s}$, $\Delta\upsilon_{mf,conflict,interaction,stai_s}$ were $\mu \sim N(0, 200)$ and $\sigma \sim N(0, 200)$; the priors for group level parameters of $\upsilon_{mb,base}$, $\Delta\upsilon_{mb,sw}$,$\Delta\upsilon_{mb,mcqcsc}$,

$\Delta\upsilon_{mb,stai_s}$, $\upsilon_{mf,base}$, $\Delta\upsilon_{mf,sw}$,$\Delta\upsilon_{mf,mcqcsc}$, $\Delta\upsilon_{mf,stai_s}$, $\Delta bound_{sw}$, p_υ and υ_{stage2} were $\mu \sim N(0, 20)$ and $\sigma \sim N(0, 20)$; the priors for group level parameters of $\Delta bound_{conflict,value}$, $\Delta bound_{conflict,action}$, $\Delta bound_{conflict,interaction}$,$\Delta bound_{conflict,value,mcqcsc}$, $\Delta bound_{conflict,action,mcqcsc}$, $\Delta bound_{conflict,interaction,mcqcsc}$, $\Delta bound_{conflict,value,stai_s}$, $\Delta bound_{conflict,action,stai_s}$, and $\Delta bound_{conflict,interaction,stai_s}$ were $\mu \sim N(1, 20)$ and $\sigma \sim N(0, 20)$; the priors for group level parameters of $bound_{base}$, $\Delta bound_{mcqcsc}$, $\Delta bound_{stai_s}$ and $bound_{stage2}$ were $\mu \sim N(1, 5)$ and $\sigma \sim N(0, 5)$; the priors for group level parameters of p_{bias} was $\mu \sim N(0, 5)$ and $\sigma \sim N(0, 5)$; the priors for group level parameters of τ were $\mu \sim N(0.5, 1)$ and $\sigma \sim N(0, 1)$, with a lower boundary of τ at 0.

2.3.5 Correlation Analysis

A potential confound of the model-fitting process was correlation between the within system value differences v.s. between system conflicts. To address this concern, we conducted a correlation analysis to look at our ability to distinguish these quantities. With the learning rates α and the discount rate λ fitted in the softmax model, we simulated Q values trial-by-trial using the same SARSA algorithm for both the MB and the MF controllers in both tasks, using all four datasets. Then, we performed a correlation analysis of between-system Q value conflicts and within-system Q value differences using the Hmisc package in R.

Chapter 3: Results

3.1 Overview

Previous studies have consistently identified both MB and MF reinforcement learning in the Two-step task and the Rocket task. Both systems produce their respective value estimations for choices, and deferentially weigh into an incorporated decision. We used four published datasets that are comparable in their experimental setup and reward structure, three from the Two-step task (solway2019loss,kool2016does,unpublished data) and one from the rocket task [37]. We used Bayesian estimation with weakly informative priors to fit all free parameters in the Integrated DDM reinforcement learning model to participants' choices and reaction times in the Two-step task and the rocket task. In our model, the MB controller and the MF controller shared the same decision boundaries and action biases, but drifted towards the boundaries separately according to their unique value predictions.

All four hierarchical model fits converged, with $\hat{R}$ of all parameters smaller than 1.05. Supplement A. reports the estimated drift rates and other parameters in the models. Both MB and MF had significant baseline drift rates (Kool2016/rk: 3.30 [2.68 3.92] and 1.01 [0.33 1.69]; Kool2016/ts: 7.18 [5.11 9.24] and 7.07 [5.70 8.58]; Solway2019/ts: 4.41 [2.44 6.35] and 7.25 [5.82 8.96]; Unpublished/ts: 6.10 [4.18 8.05]

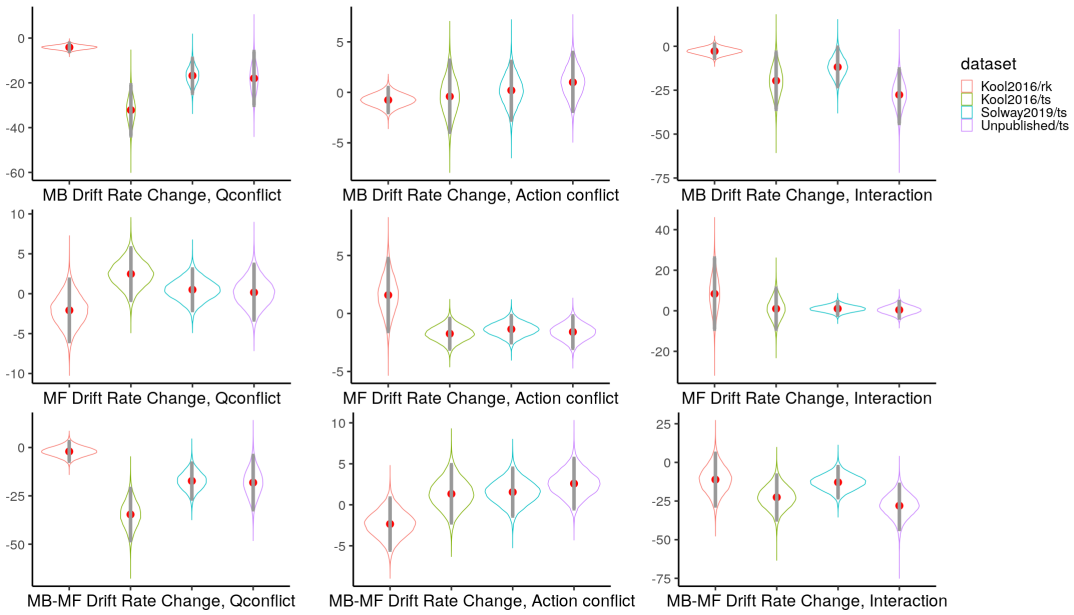

Figure 3.1: Model fits of conflict-related drift rate parameters in all four datasets. The top and middle panels represent the MB and the MF drift rate changes related to Q value conflict, action conflict, and conflict with interaction (the sum of the Q value conflict term and the interaction term) from left to right. The bottom panel presents the substraction between the two systems' drift rate changes.

and 5.40 [4.12 7.03], respectively), confirming that both systems contribute to the decision process during both tasks, across all datasets.

3.2 Effects of Valuation Conflicts and Action Conflicts

We defined the conflict between the MB and the MF systems as two kinds: the valuation conflict and the action conflict. The Q value conflicts calculate the absolute difference in relative value-based preference of the two first-stage choices between MB learning and MF learning, while the preference is determined by the difference between the value estimations of the two choices of each system. In other words, it represents how different the relative valuation of the two choices are in the two systems in a continuous scale. In contrast, the action conflict captures a

binary confrontation between the two systems: when one system values a choice higher than the other, but the other system values the two choices in the opposite direction. When the action conflict occurs, the drift rates of the two systems hold different directions and cancel out each other. For each of the two conflicts and their interaction, we assigned a coefficient parameter in the MB drift rate, the MF drift rate and the (first-stage) decision boundary. The interaction term addresses additional change to value conflict parameters when an action conflict exists. When there is no action conflict in a trial, the total conflict-related change of this trial is the value of the Q conflict regressor. Otherwise, the total change is the sum of the Q conflict regressor and the interaction regerssor.

Figure 3.1 shows the coefficient terms of all conflict-related parameters (using a 95% Credible Interval). Q value conflict suppresses MB decision-making in all four datasets when there is no action conflict (Kool2016/rk: -4.12 [-5.92 -2.31]; Kool2016/ts: -32.10 [-43.61 -20.85]; Solway2019/ts: -16.76 [-24.60 -9.11]; Unpublished/ts: -17.93 [-30.04 -6.02]). All three Two-step datasets showed such suppression in the MB drift rate when there is action conflict (Kool2016/rk: -2.77 [-6.71 1.17]; Kool2016/ts: -19.63 [-35.86 -3.41]; Solway2019/ts: -11.83 [-22.94 -0.53]; Unpublished/ts: -27.46 [-43.91 -12.92]). Participants disengage from the MB system when there is a Q value conflict to let the MF system dominate the decision process in the Two-step task. This suppression was mostly not found in the MF system with action conflict (Kool2016/rk: 8.34 [-9.04 25.98]; Kool2016/ts: 2.95 [-1.12 7.19]; Solway2019/ts: 1.01 [-2.14 4.27]; Unpublished/ts: 0.39 [-3.52 4.55]) or without (Kool2016/rk: 1.58 [-1.55 4.73]; Kool2016/ts: 2.47 [-0.82 5.73]; Solway2019/ts: 0.50

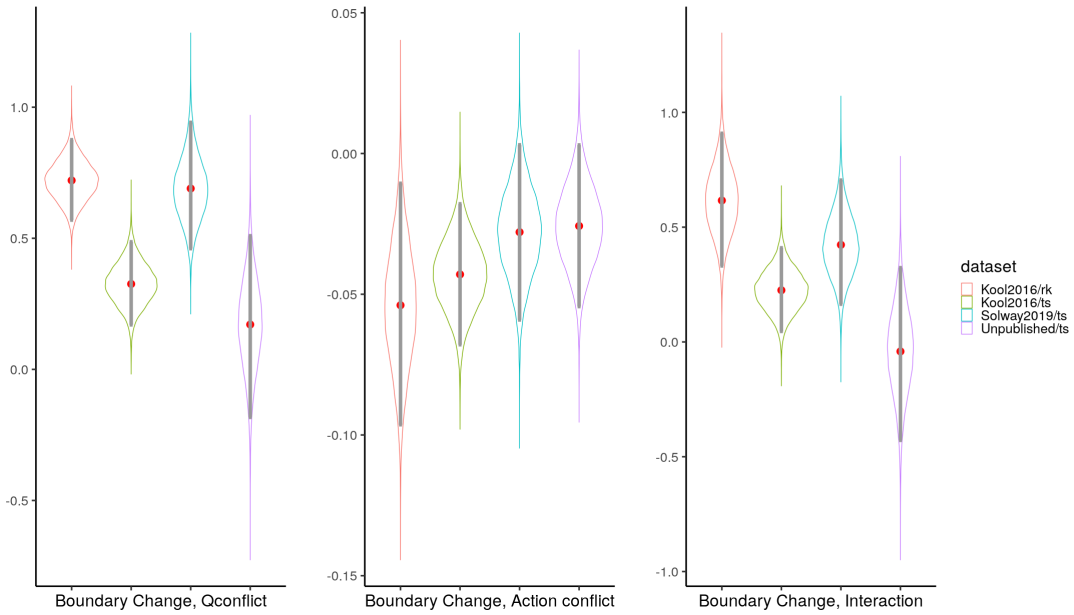

Figure 3.2: Model fits of conflict-related boundary parameters in all four datasets.

[-2.10 3.09]; Unpublished/ts: 0.14 [-3.28 3.68]).

Independent of Q value conflict, the MF system adapts to action conflicts. Positive action conflict decreased the MF drift rate (Kool2016/rk: 10.36 [-7.34 28.28]; Kool2016/ts: -1.74 [-3.06 -0.43]; Solway2019/ts: -1.36 [-2.52 -0.19]; Unpublished/ts: -1.58 [-3.00 -0.22]), but not the MB drift rate (Kool2016/rk: -0.76 [-1.99 0.47]; Kool2016/ts: -0.41 [-3.95 3.18]; Solway2019/ts: 0.19 [-2.74 3.06]; Unpublished/ts: 0.99 [-1.87 3.95]), in all three Two-step datasets.

Q value conflict is also positively correlated with boundary extension across three out of four datasets [37, 66] of the two tasks, both when there is an action conflict (Kool2016/rk: 0.62 [0.33 0.91]; Kool2016/ts: 0.23 [0.05 0.41]; Solway2019/ts: 0.42 [0.16 0.70]; Unpublished/ts: -0.04 [-0.43 0.32]) and when there is not (Kool2016/rk: 0.72 [0.24 1.14]; Kool2016/ts: 0.32 [0.17 0.49]; Solway2019/ts: 0.69 [0.46 0.94]; Unpublished/ts: 0.17 [-0.18 0.51]).

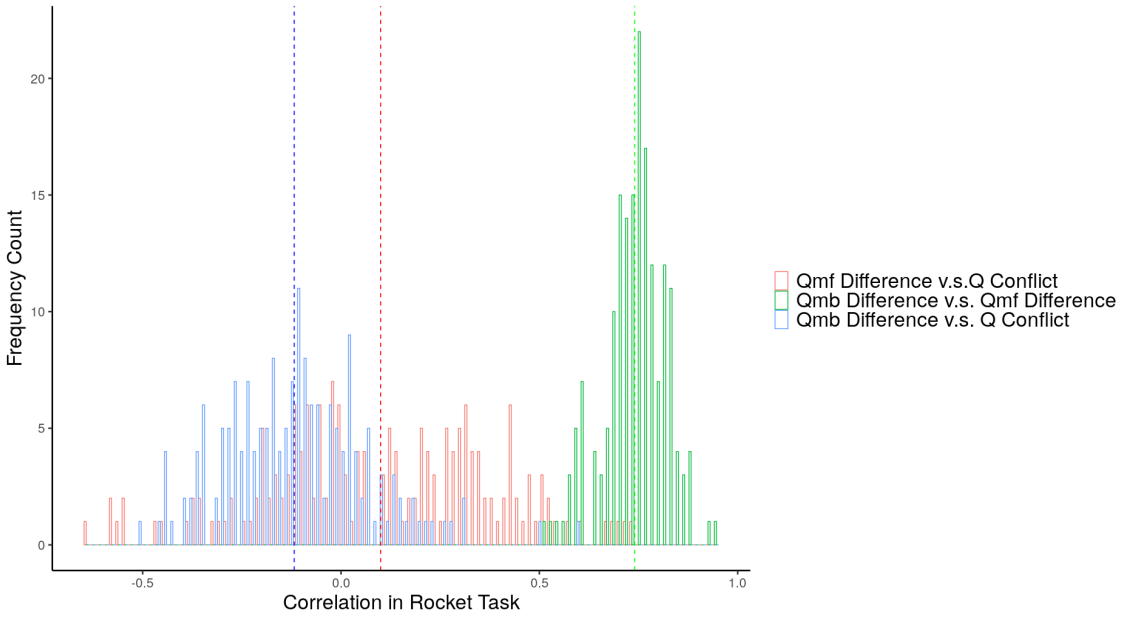

Figure 3.3: Correlation coefficients of the three Q quantities (within-MF-system difference, within-MB-system difference, and between-system difference) in Kool2016/rk.

3.3 Correlational Caveat

One potential caveat that may limit the reliability of model fitting is the correlation between the two within-system Q value differences (Q value differences between the two first-stage choices in each learning system), and between the with-in system difference and the between-system conflicts (Fig. 3.3).

In the Two-step task, between-system conflicts were partially influenced by the MF Q value differences between two presented choices at each trial, indicated by the weak correlation between the MF Q value differences and the value of Q conflicts (Kool2016/ts: mean =0.23 , sd = 0.18 ; Solway2019/ts: mean = 0.29 , sd = 0.37 ; Unpublished/ts: mean = 0.16 , sd = 0.21). This correlation was not identified between the MB Q value differences and the value of Q conflicts (Kool2016/ts:

mean =-0.01 , sd = 0.29 ; Solway2019/ts: mean = 0.01 , sd = 0.23; Unpublished/ts: mean = -0.01 , sd =0.13). This is because the scale of choice differences in the MF system tends to be larger than that in the MB system in the Two-step task, in which the MB Q value estimates in the first stage are taken from the maximum of the second-stage values estimates. Additionally, the MB Q value difference was moderately correlated with the MF Q value difference among most participants (Kool2016/ts: mean =0.40 , sd = 0.08; Solway2019/ts: mean = 0.51 , sd = 0.25; Unpublished/ts: mean = 0.40 , sd =0.23).

In the rocket task, the between-system conflicts were not correlated with within-system value differences (MB and Q conflict: mean = 0.10 , sd = 0.28; MF and Q conflict: mean = -0.12 , sd = 0.18); the two within-system value differences are strongly correlated among participants (mean = 0.74, sd = 0.079).

3.4 Response Perseveration Influence Decisions

We identified both aspects of response perseveration across all four datasets, specifically a starting point farther away (Kool2016/rk: -0.11 [-0.13 -0.08], Kool2016/ts: -0.22 [-0.26 -0.18], Solway2019/ts: -0.31[-0.37 -0.26], Unpublished/ts: -0.20 [-0.24 -0.15]) as well as a higher drift rate (Kool2016/rk: 0.39 [0.35 0.43], Kool2016/ts: 0.98 [0.85 1.11], Solway2019/ts: 1.05 [0.89 1.20], Unpublished/ts: 0.69 [0.57 0.80]) for the choice selected in the previous trial (Figure 3.4).

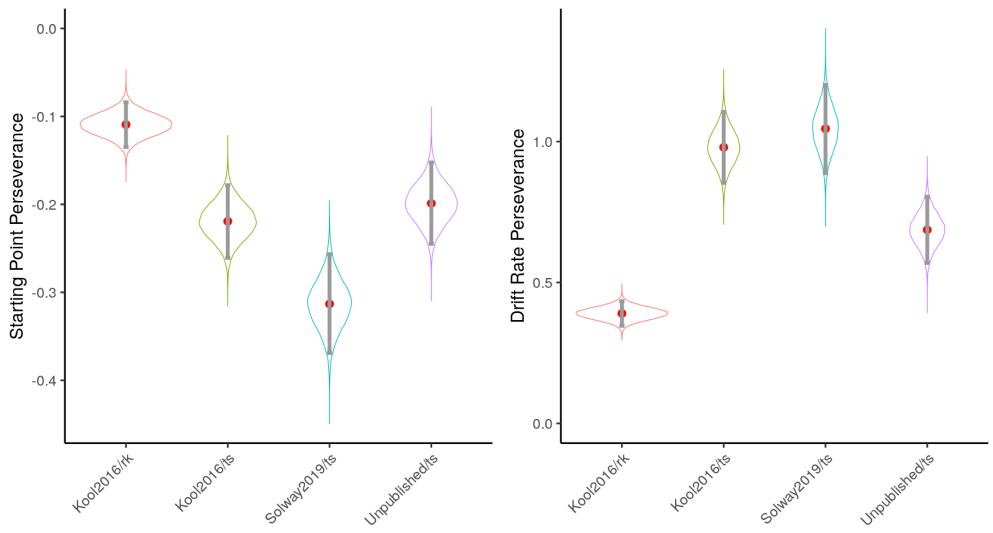

Figure 3.4: Response perseveration across four datasets.

3.5 Individual Traits

Unlike what we expected, we did not find significant correlation between of MCQCSC or STAI_S score with any of the conflict-related changes in the MB drift rate, the MF drift rate, or boundary separation (see Supplement B). However, MCQCSC score was positively correlated with the MB drift rate (2.21 [0.26 4.21]), opposite to what we expected. When we did not control for state anxiety, the effect of MCQCSC disappeared (1.72 [-0.24 3.70]). There was no significant relationship between the two scores and the MF drift rate, or the two scores and the boundary separation.

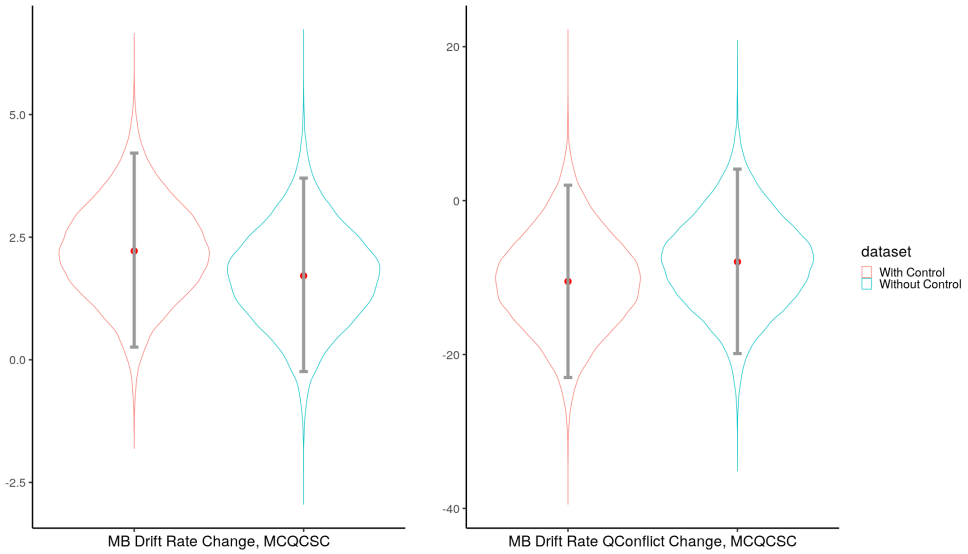

Figure 3.5: Model fits of MCQCSC regressors, the MB drift rate change (left) and the MB drift rate change related to Q value conflicts (right).

Chapter 4: Discussion

4.1 Overview

This project explores the underlying mechanisms of conflict adaption during value-based decisions using the popular computational reinforcement learning framework. Over the last century, it becomes a prevalent belief that our learning system composes two important parts: MB control and MF control. While the majority of studies model their contribution to behaviors independently, it is crucial to understand their interaction, or adaptation to their conflicts. We fitted four datasets using two different experiment paradigm to explore how MB system and MF system respond to a decision conflict. Particularly, we divided the decision conflict into two situations: (a) when the two systems disagree on the optimal choice, in which case when there is an action conflict and a valuation conflict; and (b) when the two systems simply have different value estimations of the choices, showing a Q value conflict. Our results from model fitting revealed that MB system is the system that mostly yields in when there is a conflict between the two systems during a decision, while MF system remains relatively unaffected, especially in the Two-step task. This suppression in MB control does not require the two systems to prefer different decisions; in fact, it is robust even when there is only valuation conflicts between the

two systems' value estimation of choices. When there is no action conflict but only valuation conflict, participants suppressed MB system whether it's optimal or not, potentially because the brain does not know if there will be an action conflict ahead of time. Interestingly, although MB suppression is significant in the Rocket task when there was Q value conflict, the difference between MB and MF changes was not significant, indicating that both systems were suppressed comparably. When action conflict between the two systems exist, all three datasets using the Two-step task showed the same effect as it did in the no-action-conflict trials. While the rocket task dataset did not have the suppression in the MB drift rate during trials with action conflicts, this result may be due to the sampling error: while trials with a positive action conflict constituted roughly 28% to 35% of all trials in all three datasets of Two-step tasks, the rocket task only had $\sim$ 15% trials that has such conflict. This largely reduced the accuracy of model-fitting for the interaction term. Another potential reason behind the different results of the two tasks is the difference in the task structures: the rocket task motivates MB learning more than the the Two-step task, which means that it makes more sense for participants to pick the MB side, when the two systems do prefer different choices. This idea would be consistent with kool2016does's analysis: MB suppression in the Two-step tasks is due to the lack of incentives in the Two-step paradigm. These results further support that MF control is the default system which can hardly be suppressed or turned off. Although one theoretical framework (e.g. dezfouli2013actions) depicted a picture in which MB control is on a higher level and is able to alter the contribution of MF system during cognitive conflicts, our results were more in line with other

work (e.g. eppinger2013goals,otto2013working) which suggests that the propensity for MB control is flexible relative to MF control.

When we look at the latent decision parameters, decision drift rates are not the only parameters that are affected during decision conflicts. Extension of decision boundary is observed in three out of four datasets [37, 66] with or without action conflict. Participants increased their decision thresholds when facing a valuation conflict, whether or not it was strong enough for an action conflict to occur. This result is consistent with previous studies [34] on extension of decision boundary during conflicts. Although we did not find this effect in the unpublished dataset, it is noteworthy that this dataset had the smallest sample size among all datasets, with an effective sample size of 84. More data is needed to further confirm the influence of cognitive conflict on boundary separation.

4.2 Process Model of MB Control and MB-MF Integration

Notice that in this thesis, MB value computation and the decision process is in a serial order. While this is a typical way of modeling in studies looking at both MB learning and MF learning, in reality, it is not likely that we have the MB values computed before making a decision – the drift diffusion process applies exactly because we do not have access to the exact values. Here we are assuming that the MB values are computed first in serial order with the MF estimation for computational purpose, whereas such estimation of the MB values is likely part of the decision process.

More accurate process models of MB control indicate that the MB system samples from noisy distributions of one-step rewards. For example, instead of taking the average of the past rewards as the value estimation of an action, MB deliberation may compare the choice item against a series of references sampled from memory [6, 72]. Moreover, while this project treats first and second-stage value updating separately, previous work has shown that the MB decision process integrates evidence across time, not limited by stage, as competitions between different paths of a decision tree [6, 72]. A different process model of MB control suggests the brain treats it as a Bayesian inference problem, where the decision-making process iteratively computes the posterior over a joint probability distribution composed of transition functions and rewards [72], disparate from the additive computations we do in this project.

Besides the process model of the MB controller itself, we also need a process model that describes how the two controllers integrate to make a singular action. Historically, these systems have been treated independently, each making its own parametric predictions and contributing to choices separately. However, in reality the two controllers also need to interact further to adjust to their conflict. The interaction is either through a third party controller, in which case the conflict is computed in a separate neural system that connects with both controllers, or through direct competition with each other, in which case the predictions from one system would drive the conflict response, or suppression, in the other system.

A confounding issue of this "third-party controller" versus "direct competition" question is the quantity used to calculate adaptation: what quantity results in

suppression? There are two potential answers: conflict monitoring between the two systems uses either the cross-system difference (a.k.a. "conflict"), or the within-MF-system difference between choices as the indicator. In the first case, the monitor needs to have access to real-time value estimations in both the MB and the MF systems, to then convert these estimates into a conflict index. In the second case, the MF system passes its control signal, directly or not, to the MB system, and essentially has "control" over performances in both systems.

No current theory has explained why the value estimation in either system would modulate performance across systems, but a third-party controller hypothesis does have some theoretical support that conflict adaptation uses information from both systems (as well as their conflict). This popular hypothesis is that the dACC integrates task information, particularly the expected rewards and costs, to estimate the expected value of control (shenhav2013expected; EVC). In situations with response competition, the EVC estimates the cognitive cost and expected reward or outcome of a control sign, and determines if more attention-based control should be employed in order to strike a balance between the use of attention resources and the optimization of task performance. However, it is not clear how the EVC framework applies to the current findings, since it is based on brain optimality in terms of resources, which does not seem to be the case in our data, considering that participants engage in the costly MB learning whether it is more rewarding than MF learning (e.g. in the Rocket task) or not (e.g. in the Two-step task). Given that the two tasks have different scales of with-in system value differences, we cannot compare across the datasets either to determine if participants used a

higher arbitration of MB relative to MF when MB is more gainful.

On the other hand, it is also unlikely that the MF controller is completely unconstrained and dominating. Some previous study has introduced the idea that judgements of reliability of the MB estimation and the MF estimation are incorporated to determine the arbitration of the two systems in a particular decision [2]. To take this idea further, ODoherty2015structure proposed an integrated RL framework that incorporates reliability-based arbitration, value estimations, and state inferences. In this model, MB and MF value estimations are both affected by direct interaction with each other and a higher-level metacontroller based on the two systems' respectively reliability. This reliability quantity, just like in our work, is calculated based on trials in the recent past and updated trial-by-trial. Evidence showed that the MF controller is adopted proportionally to its reliability. In other words, the reliability metacontroller can potentially promote the arbitration of MB learning over MF learning in worthy scenarios.

4.3 Caveats

Given the current tasks, we are able to differentiate between the two within-system differences of MB control and MF control, and between each of them and the between-system conflict. Although there is some lack of orthogonality between the two with-in system regressors in the Rocket task, Kool and colleagues' study has successfully recovered separate, simulated MB and MF parameters through model-fitting procedures using the same task, indicting that the task structure allows

researchers to extract the distinguishment between the two systems even with the presence of high value correlation [37]. However, it would be ideal in the future to have a further adapted version of this multi-step valued-based decision task that can tear apart the two within-system estimations better.

Related to task design, it is also worthy to note if the choice competition happens at the stimulus level, or the action level. In this project, we do not make a distinguishment between the target of value assignment: are the MB and the MF controllers assigning values to the stimuli, or are these values assigned to the actions? Previous work has found that the dorsomedial prefrontal cortex (dmPFC) and the intraparietal sulcus (IPS), after the ventromedial prefrontal cortex (vmPFC) encoded stimulus values, coupled with the motor cortex to transform the stimulus values into motor commands [73]. It is not yet clear if the MB value assignment (or choice evaluation) happens before or after that transformation from stimulus values to motor commands. While this project frames MB value estimation in parallel with MF value estimation, it is also possible that the MB controller evaluates the values of motor commands instead of stimulus values.

4.4 Choice Perseveration

Back to the interpretation of results, our findings showed that response stickiness exists across population and operates in both directions: consistent across all four datasets, the starting point biases towards an altered choice other than the last chosen option, while the drift rate biases towards the last chosen option. The two

biases seem to act against each other, one via initial preference and one via faster evidence collection.to balance out each other's influence on decisions. This is the first study that has implemented perseveration parameters in more than one aspect of decision parameters. Although more research is needed to reach a convincing conclusion on this dynamic finding, our results peak into the delicate, comprehensive tuning system in even the most simplistic binary decisions. Previous research has discussed the need to consider gradual perseveration as a less constrained option than immediate perseveration [63]. Although such alternative was not supported by model selection in the original study [63], further investigation is needed to see if different aspects of choice perseveration differ on this their "memory span."

4.5 Metacognition and Decision Making

Regarding individual differences, cognitive self-consciousness and state anxiety were not directly correlated with any of the conflict-related changes. This result indicates that decision conflicts may be processed in more of a lower-level than we thought, invulnerable to higher-level cognitive factors. Interestingly, our results show that higher metacognition of worry is positively correlated with the MB drift rate. Considering the intimate association between general metacognition and working memory [74], here is a potential explanation: while state anxiety attenuates MB, but not MF learning, higher working memory rescues this attenuation, mediating the positive relationship between metacognition and the MB drift rate. Additionally, our model fits show that this result is washed away if we do not control for

STAI_S scores. This STAI_S regressor is to control for the effect of state anxiety on participants' performances, to understand the independent effect of excessive self-reflection during decision process, especially when facing a difficult decision that involves cognitive conflicts. Considering the limited sample size in this dataset to draw any conclusion about individual traits, further research is needed to look into the effect.

Appendix A: Conflict Model Parameter Fits

Table A.1: Kool,2016/RK Group Parameters

Parameter	Median and CI
Drift rate, model-based, mean	3.30 [2.68, 3.92]
Drift rate, model-based, sd	1.71 [1.43, 2.03]
Drift rate, model-free, mean	1.01 [0.33, 1.69]
Drift rate, model-free, mean	1.74 [1.36, 2.15]
Boundary separation, mean	1.56 [1.52, 1.60]
Boundary separation, sd	0.25 [0.22, 0.28]
Q conflict change, model-based, mean	-4.12 [-5.92, -2.31]
Q conflict change, model-based, sd	1.02 [0.05, 2.79]
Action conflict change, model-based, mean	-0.76 [-1.99, 0.47]
Action conflict change, model-based, sd	1.83 [0.69, 2.55]

TABLE A.1, cont'd

Conflict interaction change, model-based, mean	1.34 [-2.90, 5.62]
Conflict interaction change, model-based, sd	2.60 [0.14, 6.57]
First-stage switch regressor, model-based, mean	-2.08 [-5.99, 1.81]
First-stage switch regressor, model-based, sd	2.44 [0.12, 7.34]
Q conflict change, model-free, mean	1.58 [-1.55, 4.73]
Q conflict change, model-free, sd	4.94 [1.95, 7.32]
Action conflict change, model-free, mean	10.36 [-7.34, 28.28]
Action conflict change, model-free, sd	7.01 [0.33, 22.63]
Conflict interaction change, model-free, mean	-0.64 [-1.17, -0.11]
Conflict interaction change, model-free, sd	0.79 [0.24, 1.14]
First-stage switch regressor, model-free, mean	0.86 [-0.01, 1.72]
First-stage switch regressor, model-free, sd	0.47 [0.02, 1.34]
Q conflict change, boundary, mean	0.72 [0.57, 0.88]
Q conflict change, boundary, sd	0.38 [0.10, 0.56]
Action conflict change, boundary, mean	-0.05 [-0.10, -0.01]
Action conflict change, boundary, sd	0.04 [0.00, 0.09]
Conflict interaction change, boundary, mean	-0.11 [-0.40, 0.20]

TABLE A.1, cont'd

Conflict interaction change, boundary, sd	0.28 [0.02, 0.63]
First-stage switch regressor, boundary, mean	-0.01 [-0.03, 0.01]
First-stage switch regressor, boundary, sd	0.03 [0.00, 0.06]
Drift rate perseverance, mean	0.39 [0.35, 0.43]
Drift rate perseverance, sd	0.23 [0.20, 0.27]
Starting point perseverance, mean	-0.11 [-0.13, -0.08]
Starting point perseverance, sd	0.09 [0.06, 0.12]
Non-decision time, mean	0.38 [0.36, 0.39]
Non-decision time, sd	0.11 [0.10, 0.13]
Q conflict on model-based delta when there is action conflict	-2.77 [-6.71, 1.17]
Q conflict on model-free delta when there is action conflict	8.34 [-9.04, 25.98]
Q conflict on boundary when there is action conflict	0.62 [0.33, 0.91]

Table A.2: Solway,2019/TS Group Parameters

Parameter	Median and CI
First-stage drift rate, model-based, mean	4.41 [2.44, 6.35]
First-stage drift rate, model-based, sd	7.25 [5.82, 8.96]
First-stage drift rate, model-free, mean	3.81 [2.82, 4.81]
First-stage drift rate, model-free, sd	3.76 [3.00, 4.67]
Second-stage drift rate, mean	8.48 [7.63, 9.33]
Second-stage drift rate, sd	4.38 [3.77, 5.13]
First-stage boundary separation, mean	1.53 [1.48, 1.58]
First-stage boundary separation, sd	0.26 [0.23, 0.31]
Second-stage boundary separation, mean	1.61 [1.57, 1.66]
Second-stage boundary separation, sd	0.24 [0.21, 0.28]
Q conflict change, model-based, mean	-16.76 [-24.60, -9.11]
Q conflict change, model-based, sd	3.23 [0.16, 10.78]
Action conflict change, model-based, mean	0.19 [-2.74, 3.06]
Action conflict change, model-based, sd	4.26 [0.75, 7.14]
Conflict interaction change, model-based, mean	4.90 [-8.10, 18.19]

TABLE A.2, cont'd

Conflict interaction change, model-based, sd	6.98 [0.35, 20.09]
Q conflict change, model-free, mean	0.50 [-2.10, 3.09]
Q conflict change, model-free, sd	1.16 [0.05, 3.70]
Action conflict change, model-free, mean	-1.36 [-2.52, -0.19]
Action conflict change, model-free, sd	0.83 [0.06, 1.74]
Conflict interaction change, model-free, mean	0.52 [-3.20, 4.35]
Conflict interaction change, model-free, sd	1.43 [0.07, 4.06]
Q conflict change, boundary, mean	0.69 [0.46, 0.94]
Q conflict change, boundary, sd	0.71 [0.42, 1.05]
Action conflict change, boundary, mean	-0.03 [-0.06, 0.00]
Action conflict change, boundary, sd	0.05 [0.01, 0.09]
Conflict interaction change, boundary, mean	-0.27 [-0.49, -0.05]
Conflict interaction change, boundary, sd	0.18 [0.01, 0.42]
Drift rate perseverance, mean	1.05 [0.89, 1.20]
Drift rate perseverance, sd	0.84 [0.73, 0.97]
Starting point perseverance, mean	-0.31 [-0.37, -0.26]
Starting point perseverance, sd	0.27 [0.23, 0.32]

TABLE A.2, cont'd

Non-decision time, mean	0.34 [0.32, 0.35]
Non-decision time, sd	0.09 [0.08, 0.10]
Q conflict on model-based delta when there is action conflict	-11.83 [-22.94, -0.53]
Q conflict on model-free delta when there is action conflict	1.01 [-2.14, 4.27]
Q conflict on boundary when there is action conflict	0.42 [0.16, 0.70]

Table A.3: Unpublished/TS Group Parameters

Parameter	Median and CI
First-stage drift rate, model-based, mean	6.10 [4.18, 8.05]
First-stage drift rate, model-based, sd	5.40 [4.12, 7.03]
First-stage drift rate, model-free, mean	2.72 [1.87, 3.57]
First-stage drift rate, model-free, sd	0.76 [0.13, 1.32]
Second-stage drift rate, mean	8.68 [7.31, 10.06]
Second-stage drift rate, sd	6.22 [5.28, 7.44]
First-stage boundary separation, mean	1.42 [1.38, 1.46]
First-stage boundary separation, sd	0.17 [0.14, 0.20]
Second-stage boundary separation, mean	1.63 [1.58, 1.67]
Second-stage boundary separation, sd	0.21 [0.18, 0.25]
Q conflict change, model-based, mean	-17.93 [-30.04, -6.02]
Q conflict change, model-based, sd	3.67 [0.16, 13.01]
Action conflict change, model-based, mean	0.99 [-1.87, 3.95]
Action conflict change, model-based, sd	1.25 [0.06, 3.95]
Conflict interaction change, model-based, mean	-9.43 [-28.92, 8.59]

TABLE A.3, cont'd

Conflict interaction change, model-based, sd	8.09 [0.41, 23.10]
Q conflict change, model-free, mean	0.14 [-3.28, 3.68]
Q conflict change, model-free, sd	1.01 [0.05, 3.65]
Action conflict change, model-free, mean	-1.58 [-3.00, -0.22]
Action conflict change, model-free, sd	0.44 [0.02, 1.41]
Conflict interaction change, model-free, mean	0.20 [-4.60, 5.39]
Conflict interaction change, model-free, sd	1.17 [0.05, 3.96]
Q conflict change, boundary, mean	0.17 [-0.18, 0.51]
Q conflict change, boundary, sd	0.99 [0.60, 1.45]
Action conflict change, boundary, mean	-0.03 [-0.05, 0.00]
Action conflict change, boundary, sd	0.04 [0.00, 0.07]
Conflict interaction change, boundary, mean	-0.21 [-0.53, 0.08]
Conflict interaction change, boundary, sd	0.24 [0.01, 0.66]
Drift rate perseverance, mean	0.69 [0.57, 0.80]
Drift rate perseverance, sd	0.52 [0.44, 0.62]
Starting point perseverance, mean	-0.20 [-0.24, -0.15]
Starting point perseverance, sd	0.18 [0.15, 0.23]

TABLE A.3, cont'd

Non-decision time, mean	0.22 [0.21, 0.24]
Non-decision time, sd	0.06 [0.05, 0.07]
Q conflict on model-based delta when there is action conflict	-27.46 [-43.91, -12.92]
Q conflict on model-free delta when there is action conflict	0.39 [-3.52, 4.55]
Q conflict on boundary when there is action conflict	-0.04 [-0.43, 0.32]

Table A.4: Kool, 2016/TS Group Parameters

Parameter	Median and CI
First-stage drift rate, model-based, mean	7.18 [5.11, 9.24]
First-stage drift rate, model-based, sd	7.07 [5.70, 8.58]
First-stage drift rate, model-free, mean	2.35 [1.55, 3.14]
First-stage drift rate, model-free, sd	1.26 [0.81, 1.69]
Second-stage drift rate, mean	7.54 [6.70, 8.39]
Second-stage drift rate, sd	5.37 [4.74, 6.11]
First-stage boundary separation, mean	1.27 [1.23, 1.30]
First-stage boundary separation, sd	0.23 [0.20, 0.26]
Second-stage boundary separation, mean	1.60 [1.56, 1.64]
Second-stage boundary separation, sd	0.27 [0.25, 0.31]
Q conflict change, model-based, mean	-32.10 [-43.61, -20.85]
Q conflict change, model-based, sd	3.68 [0.16, 12.17]
Action conflict change, model-based, mean	-0.41 [-3.95, 3.18]
Action conflict change, model-based, sd	5.42 [1.19, 8.38]
Conflict interaction change, model-based, mean	12.58 [-6.49, 31.82]

TABLE A.4, cont'd

Conflict interaction change, model-based, sd	7.86 [0.40, 25.71]
Q conflict change, model-free, mean	2.47 [-0.82, 5.73]
Q conflict change, model-free, sd	1.01 [0.04, 3.21]
Action conflict change, model-free, mean	-1.74 [-3.06, -0.43]
Action conflict change, model-free, sd	1.05 [0.09, 2.01]
Conflict interaction change, model-free, mean	0.48 [-4.51, 5.55]
Conflict interaction change, model-free, sd	1.33 [0.07, 4.37]
Q conflict change, boundary, mean	0.32 [0.17, 0.49]
Q conflict change, boundary, sd	0.46 [0.28, 0.64]
Action conflict change, boundary, mean	-0.04 [-0.07, -0.02]
Action conflict change, boundary, sd	0.06 [0.02, 0.09]
Conflict interaction change, boundary, mean	-0.10 [-0.29, 0.09]
Conflict interaction change, boundary, sd	0.17 [0.01, 0.42]
Drift rate perseverance, mean	0.98 [0.85, 1.11]
Drift rate perseverance, sd	0.82 [0.73, 0.93]
Starting point perseverance, mean	-0.22 [-0.26, -0.18]
Starting point perseverance, sd	0.23 [0.20, 0.27]

TABLE A.4, cont'd

Non-decision time, mean	0.30 [0.29, 0.31]
Non-decision time, sd	0.09 [0.08, 0.10]
Q conflict on model-based delta when there is action conflict	-19.63 [-35.86, -3.41]
Q conflict on model-free delta when there is action conflict	2.95 [-1.12, 7.19]
Q conflict on boundary when there is action conflict	0.23 [0.05, 0.41]

Appendix B: Individual Differences Model Parameter Fits

Table B.1: Unpublished/TS Group Parameters with MCQCSC Regressors, Controlling for State Anxiety

Parameter	Median and CI
First Stage Drift Rate, model-based, mean	6.23 [4.30, 8.18]
First Stage Drift Rate, model-based, sd	5.12 [3.87, 6.69]
First Stage Drift Rate, model-free, mean	2.68 [1.85, 3.51]
First Stage Drift Rate, model-free, sd	0.54 [0.04, 1.13]
Second Stage Drift Rate, mean	8.69 [7.24, 10.08]
Second Stage Drift Rate, sd	6.21 [5.24, 7.44]
First Stage Boundary Separation, mean	0.34 [0.32, 0.37]
First Stage Boundary Separation, sd	0.12 [0.10, 0.14]
Second Stage Boundary Separation, mean	0.48 [0.45, 0.51]
Second Stage Boundary Separation, sd	0.13 [0.11, 0.16]

TABLE B.1, cont'd

Q Conflict Change, model-based, mean	-19.30 [-31.47, -7.14]
Q conflict change, model-based, sd	3.95 [0.18, 14.05]
Action conflict change, model-based, mean	0.92 [-2.05, 3.93]
Action conflict change, model-based, sd	1.36 [0.06, 4.10]
Conflict interaction change, model-based, mean	-8.84 [-29.32, 9.78]
Conflict interaction change, model-based, sd	10.15 [0.57, 26.04]
Q conflict change, model-free, mean	0.16 [-3.21, 3.70]
Q conflict change, model-free, sd	1.01 [0.05, 3.36]
Action conflict change, model-free, mean	-1.49 [-2.89, -0.10]
Action conflict change, model-free, sd	0.49 [0.02, 1.52]
Conflict interaction change, model-free, mean	0.00 [-4.98, 5.25]
Conflict interaction change, model-free, sd	1.38 [0.06, 4.58]
Q conflict change, boundary, mean	0.16 [-0.22, 0.51]
Q conflict change, boundary, sd	1.08 [0.69, 1.55]
Action conflict change, boundary, mean	-0.02 [-0.05, 0.00]
Action conflict change, boundary, sd	0.04 [0.00, 0.07]
Conflict interaction change, boundary, mean	-0.20 [-0.52, 0.09]

TABLE B.1, cont'd

Conflict interaction change, boundary, sd	0.26 [0.01, 0.68]
Drift rate perseverance, mean	0.69 [0.57, 0.80]
Drift rate perseverance, sd	0.52 [0.44, 0.62]
Starting point perseverance, mean	-0.20 [-0.25, -0.15]
Starting point perseverance, sd	0.18 [0.15, 0.22]
Non-decision time, mean	0.22 [0.21, 0.24]
Non-decision time, sd	0.06 [0.05, 0.07]
MCQCSC coefficient of model-based drift rate	2.21 [0.26, 4.21]
MCQCSC coefficient of model-free drift rate	-0.35 [-1.19, 0.47]
MCQCSC coefficient of boundary separation	0.01 [-0.02, 0.04]
STAI_S coefficient of model-based drift rate	-1.86 [-3.82, 0.13]
STAI_S coefficient of model-free drift rate	-0.58 [-1.37, 0.22]
STAI_S coefficient of boundary separation	0.02 [-0.01, 0.05]
MCQCSC coefficient of Q conflict on model-based drift rate when there is no action conflict	-10.50 [-22.99, 2.00]
MCQCSC coefficient of Q conflict on model-based drift rate when there is action conflict	2.92 [-13.03, 19.81]

TABLE B.1, cont'd

MCQCSC coefficient of A conflict on model-based drift rate	-0.60 [-3.90, 2.61]
MCQCSC coefficient of Q conflict on model-free drift rate, when there is no action conflict	3.40 [-0.05, 6.87]
MCQCSC coefficient of Q conflict on model-free drift rate, when there is action conflict	1.29 [-3.02, 5.61]
MCQCSC coefficient of Action conflict on model-free drift rate	0.01 [-1.42, 1.43]
MCQCSC coefficient of Q conflict on boundary separation, when there is no action conflict	0.12 [-0.24, 0.49]
MCQCSC coefficient of Q conflict on boundary separation, when there is action conflict	0.07 [-0.33, 0.47]
MCQCSC coefficient of Action conflict on boundary separation	-0.01 [-0.04, 0.02]
STAI_S coefficient of Q conflict on model-based drift rate when there is no action conflict	6.56 [-5.40, 18.40]
STAI_S coefficient of Q conflict on model-based drift rate when there is action conflict	-0.70 [-15.47, 13.92]
STAI_S coefficient of A conflict on model-based drift rate	0.12 [-2.76, 2.94]
STAI_S coefficient of Q conflict on model-free drift rate, when there is no action conflict	-0.98 [-4.20, 2.19]

TABLE B.1, cont'd

STAI_S coefficient of Q conflict on model-free drift rate, when there is action conflict	-1.63 [-5.49, 2.26]
STAI_S coefficient of Action conflict on model-free drift rate	0.81 [-0.53, 2.17]
STAI_S coefficient of Q conflict on boundary separation, when there is no action conflict	-0.13 [-0.50, 0.22]
STAI_S coefficient of Q conflict onboundary separation, when there is action conflict	0.03 [-0.35, 0.44]
STAI_S coefficient of Action conflict on boundary separation	0.01 [-0.02, 0.04]

Table B.2: Unpublished/TS Group Parameters with MCQCSC Regressors, Without Controlling for State Anxiety

Parameter	Median and CI
First Stage Drift rate, model-based, mean	6.14 [4.20, 8.14]
First Stage Drift rate, model-based, sd	5.42 [4.15, 7.02]
First Stage Drift rate, model-free, mean	2.71 [1.85, 3.55]
First Stage Drift rate, model-free, sd	0.78 [0.13, 1.35]
Second Stage Drift rate, mean	8.66 [7.26, 10.05]
Second Stage Drift rate, sd	6.22 [5.26, 7.43]
First Stage Boundary Separation, mean	0.34 [0.31, 0.37]
First Stage Boundary Separation, sd	0.12 [0.10, 0.14]
Second Stage Boundary Separation, mean	0.48 [0.45, 0.51]
Second Stage Boundary Separation, sd	0.13 [0.11, 0.16]
Q conflict change, model-based, mean	-18.43 [-30.63, -6.29]
Q conflict change, model-based, sd	3.89 [0.19, 13.73]
Action conflict change, model-based, mean	0.93 [-1.99, 3.95]
Action conflict change, model-based, sd	1.24 [0.05, 3.96]

TABLE B.2, cont'd

Conflict interaction change, model-based, mean	-9.07 [-28.74, 9.09]
Conflict interaction change, model-based, sd	8.37 [0.42, 23.80]
Q conflict change, model-free, mean	0.25 [-3.18, 3.90]
Q conflict change, model-free, sd	1.16 [0.05, 3.99]
Action conflict change, model-free, mean	-1.57 [-2.97, -0.16]
Action conflict change, model-free, sd	0.45 [0.02, 1.46]
Conflict interaction change, model-free, mean	0.25 [-4.72, 5.40]
Conflict interaction change, model-free, sd	1.29 [0.06, 4.33]
Q conflict change, boundary, mean	0.17 [-0.20, 0.52]
Q conflict change, boundary, sd	1.03 [0.63, 1.50]
Action conflict change, boundary, mean	-0.03 [-0.05, 0.00]
Action conflict change, boundary, sd	0.04 [0.00, 0.07]
Conflict interaction change, boundary, mean	-0.22 [-0.54, 0.08]
Conflict interaction change, boundary, sd	0.26 [0.01, 0.69]
Drift rate perseverance, mean	0.69 [0.57, 0.80]
Drift rate perseverance, sd	0.52 [0.44, 0.63]
Starting point perseverance, mean	-0.20 [-0.25, -0.15]

TABLE B.2, cont'd

Starting point perseverance, sd	0.18 [0.15, 0.22]
Non-decision time, mean	0.22 [0.21, 0.24]
Non-decision time, sd	0.06 [0.05, 0.07]
MCQCSC coefficient of model-based drift rate	1.72 [-0.24, 3.70]
MCQCSC coefficient of model-free drift rate	-0.35 [-1.19, 0.48]
MCQCSC coefficient of boundary separation	0.02 [-0.01, 0.05]
MCQCSC coefficient of Q conflict on model-based drift rate when there is no action conflict	-7.93 [-19.88, 4.09]
MCQCSC coefficient of Q conflict on model-based drift rate when there is action conflict	2.55 [-12.83, 18.50]
MCQCSC coefficient of A conflict on model-based drift rate	-0.39 [-3.57, 2.68]
MCQCSC coefficient of Q conflict on model-free drift rate, when there is no action conflict	2.57 [-0.83, 6.00]
MCQCSC coefficient of Q conflict on model-free drift rate, when there is action conflict	0.88 [-3.34, 5.21]
MCQCSC coefficient of Action conflict on model-free drift rate	0.05 [-1.34, 1.44]
MCQCSC coefficient of Q conflict on boundary separation, when there is no action conflict	0.08 [-0.27, 0.43]

TABLE B.2, cont'd

MCQCSC coefficient of Q conflict on boundary separation, when there is action conflict	0.08 [-0.31, 0.47]
MCQCSC coefficient of Action conflict on boundary separation	0.00 [-0.03, 0.03]

Bibliography

[1] Nathaniel D Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, 8(12):1704–1711, 2005.

[2] Sang Wan Lee, Shinsuke Shimojo, and John P O'Doherty. Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, 81(3):687–699, 2014.

[3] Nathaniel D Daw and John P O'Doherty. *Neuroeconomics: Chapter 21. Multiple Systems for Value Learning*. Elsevier Inc. Chapters, 2013.

[4] Catherine A Thorn, Hisham Atallah, Mark Howe, and Ann M Graybiel. Differential dynamics of activity changes in dorsolateral and dorsomedial striatal loops during learning. *Neuron*, 66(5):781–795, 2010.

[5] John P O'Doherty, Sang Wan Lee, and Daniel McNamee. The structure of reinforcement-learning mechanisms in the human brain. *Current Opinion in Behavioral Sciences*, 1:94–100, 2015.

[6] Alec Solway and Matthew M Botvinick. Evidence integration in model-based tree search. *Proceedings of the National Academy of Sciences*, 112(37):11708–11713, 2015.

[7] Sang Wan Lee, John P O'Doherty, and Shinsuke Shimojo. Neural computations mediating one-shot learning in the human brain. *PLoS biology*, 13(4), 2015.

[8] Amir Dezfouli and Bernard W Balleine. Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized. *PLoS computational biology*, 9(12), 2013.

[9] Fiery Cushman and Adam Morris. Habitual control of goal selection in humans. *Proceedings of the National Academy of Sciences*, 112(45):13817–13822, 2015.

[10] Matthew M Botvinick, Todd S Braver, Deanna M Barch, Cameron S Carter, and Jonathan D Cohen. Conflict monitoring and cognitive control. *Psychological review*, 108(3):624, 2001.

[11] Nathaniel D Daw, Samuel J Gershman, Ben Seymour, Peter Dayan, and Raymond J Dolan. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215, 2011.

[12] Roger Ratcliff. A theory of memory retrieval. *Psychological review*, 85(2):59, 1978.

[13] Roger Ratcliff and Gail McKoon. The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, 20(4):873–922, 2008.

[14] Roger Ratcliff, Philip L Smith, Scott D Brown, and Gail McKoon. Diffusion decision model: Current issues and history. *Trends in cognitive sciences*, 20(4):260–281, 2016.

[15] Roger Ratcliff and Jeffrey N Rouder. A diffusion model account of masking in two-choice letter identification. *Journal of Experimental Psychology: Human perception and performance*, 26(1):127, 2000.

[16] Joshua I Gold and Michael N Shadlen. Neural computations that underlie decisions about sensory stimuli. *Trends in cognitive sciences*, 5(1):10–16, 2001.

[17] Pablo Gomez, Roger Ratcliff, and Manuel Perea. A model of the go/no-go task. *Journal of Experimental Psychology: General*, 136(3):389, 2007.

[18] Ian Krajbich and Antonio Rangel. Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, 108(33):13852–13857, 2011.

[19] Jennifer Batchelor, Allison G Harvey, and Richard A Bryant. Stroop colour word test as a measure of attentional deficit following mild head injury. *The Clinical Neuropsychologist*, 9(2):180–186, 1995.

[20] Jonathan D Cohen, Kevin Dunbar, and James L McClelland. On the control of automatic processes: a parallel distributed processing account of the stroop effect. *Psychological review*, 97(3):332, 1990.

[21] James F Cavanagh, Joseph L Sanguinetti, John JB Allen, Scott J Sherman, and Michael J Frank. The subthalamic nucleus contributes to post-error slowing. *Journal of cognitive neuroscience*, 26(11):2637–2644, 2014.

[22] Daniel Kahneman and Shane Frederick. Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment*, 49:81, 2002.

[23] Donald Norman and Tim Shallice. *Attention to action: Willed and automatic control of behavior*. 01 1986.

[24] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[25] Gavin A Rummery and Mahesan Niranjan. *On-line Q-learning using connectionist systems*, volume 37. University of Cambridge, Department of Engineering Cambridge, UK, 1994.

[26] Fredrik Edin, Torkel Klingberg, Pär Johansson, Fiona McNab, Jesper Tegnér, and Albert Compte. Mechanism for top-down control of working memory capacity. *Proceedings of the National Academy of Sciences*, 106(16):6802–6807, 2009.

[27] Quentin JM Huys and Daniel Renz. A formal valuation framework for emotions and their control. *Biological psychiatry*, 82(6):413–420, 2017.

[28] Jan Gläscher, Nathaniel Daw, Peter Dayan, and John P O'Doherty. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4):585–595, 2010.

[29] Nachshon Meiran, Gary M Diamond, Doron Toder, and Boris Nemets. Cognitive rigidity in unipolar depression and obsessive compulsive disorder: Examination of task switching, stroop, working memory updating and post-conflict adaptation. *Psychiatry research*, 185(1-2):149–156, 2011.

[30] Robin G Morris, John J Downes, Barbara J Sahakian, John L Evenden, Andrew Heald, and Trevor W Robbins. Planning and spatial working memory in parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 51(6):757–766, 1988.

[31] Amitai Shenhav, Matthew M Botvinick, and Jonathan D Cohen. The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, 79(2):217–240, 2013.

[32] William Greene. Discrete choice modeling. In *Palgrave handbook of econometrics*, pages 473–556. Springer, 2009.

[33] Amitai Shenhav, Jonathan D Cohen, and Matthew M Botvinick. Dorsal anterior cingulate cortex and the value of control. *Nature neuroscience*, 19(10):1286–1291, 2016.

[34] James F Cavanagh, Thomas V Wiecki, Michael X Cohen, Christina M Figueroa, Johan Samanta, Scott J Sherman, and Michael J Frank. Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature neuroscience*, 14(11):1462, 2011.

[35] Angus W MacDonald, Jonathan D Cohen, V Andrew Stenger, and Cameron S Carter. Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science*, 288(5472):1835–1838, 2000.

[36] Tobias Egner and Joy Hirsch. Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. *Nature neuroscience*, 8(12):1784–1790, 2005.

[37] Wouter Kool, Fiery A Cushman, and Samuel J Gershman. When does model-based control pay off? *PLoS computational biology*, 12(8), 2016.

[38] Wouter Kool, Samuel J Gershman, and Fiery A Cushman. Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological science*, 28(9):1321–1333, 2017.

[39] Daniel McFadden et al. Conditional logit analysis of qualitative choice behavior. 1973.

[40] Antonio Páez, Darren M Scott, and Erik Volz. A discrete-choice approach to modeling social influence on individual decision making. *Environment and Planning B: Planning and Design*, 35(6):1055–1069, 2008.

[41] Wim Bernasco and Richard Block. Where offenders choose to attack: A discrete choice model of robberies in chicago. *Criminology*, 47(1):93–130, 2009.

[42] Joshua I Gold and Michael N Shadlen. The neural basis of decision making. *Annual review of neuroscience*, 30, 2007.

[43] Milica Milosavljevic, Jonathan Malmaud, Alexander Huth, Christof Koch, and Antonio Rangel. The drift diffusion model can account for the accuracy and reaction time of value-based choices under high and low time pressure. *Judgment and Decision Making*, 5(6):437, 2010.

[44] Woo-Young Ahn, Nathaniel Haines, and Lei Zhang. Revealing neurocomputational mechanisms of reinforcement learning and decision-making with the hbayesdm package. *Computational Psychiatry*, 1:24–57, 2017.

[45] Aki Vehtari, Janne Ojanen, et al. A survey of bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228, 2012.

[46] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[47] Michael D Lee. How cognitive modeling can benefit from hierarchical bayesian models. *Journal of Mathematical Psychology*, 55(1):1–7, 2011.

[48] Rick L Andrews, Asim Ansari, and Imran S Currim. Hierarchical bayes versus finite mixture conjoint analysis models: A comparison of fit, prediction, and partworth recovery. *Journal of Marketing Research*, 39(1):87–98, 2002.

[49] Roger M Harbord, Penny Whiting, Jonathan AC Sterne, Matthias Egger, Jonathan J Deeks, Aijing Shang, and Lucas M Bachmann. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. *Journal of clinical epidemiology*, 61(11):1095–1103, 2008.

[50] Claire M Gillan and Trevor W Robbins. Goal-directed learning and obsessive–compulsive disorder. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1655):20130475, 2014.

[51] Paolo Cavedini, Alessandra Gorini, and Laura Bellodi. Understanding obsessive–compulsive disorder: focus on decision making. *Neuropsychology review*, 16(1):3–15, 2006.

[52] Natalia S Lawrence, Sarah Wooderson, David Mataix-Cols, Rhodri David, Anne Speckens, and Mary L Phillips. Decision making and set shifting impairments are associated with distinct symptom dimensions in obsessive-compulsive disorder. *Neuropsychology*, 20(4):409, 2006.

[53] Valerie Voon, Katherine Derbyshire, Chistian Rück, Michael A Irvine, Yulia Worbe, Jesper Enander, Liana RN Schreiber, Claire Gillan, Naomi A Fineberg, Barbara J Sahakian, et al. Disorders of compulsivity: a common bias towards learning habits. *Molecular psychiatry*, 20(3):345–352, 2015.

[54] Adrian Wells and Robert L Leahy. Cognitive therapy of anxiety disorders: A practice manual and conceptual guide, 1998.

[55] Adrian Wells and Costas Papageorgiou. Relationships between worry, obsessive–compulsive symptoms and meta-cognitive beliefs. *Behaviour research and therapy*, 36(9):899–913, 1998.

[56] Adrian Wells and Gerald Matthews. *Attention and emotion (Classic edition): A clinical perspective*. Psychology Press, 1994.

[57] Sam Cartwright-Hatton and Adrian Wells. Beliefs about worry and intrusions: The meta-cognitions questionnaire and its correlates. *Journal of anxiety disorders*, 11(3):279–296, 1997.

[58] Ben Eppinger, Maik Walter, Hauke R Heekeren, and Shu-Chen Li. Of goals and habits: age-related and individual differences in goal-directed decision-making. *Frontiers in neuroscience*, 7:253, 2013.

[59] Peter Smittenaar, Thomas HB FitzGerald, Vincenzo Romei, Nicholas D Wright, and Raymond J Dolan. Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. *Neuron*, 80(4):914–919, 2013.

[60] Elmar D Grosskurth, Dominik R Bach, Marcos Economides, Quentin JM Huys, and Lisa Holper. No substantial change in the balance between model-free and

model-based control via training on the two-step task. *PLoS computational biology*, 15(11):e1007443, 2019.

[61] Nitzan Shahar, Tobias U Hauser, Michael Moutoussis, Rani Moran, Mehdi Keramati, Raymond J Dolan, et al. Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLoS computational biology*, 15(2):e1006803, 2019.

[62] Natalie R Hall and Richard J Crisp. Anxiety-induced response perseverance and stereotyping change. *Current Research in Social Psychology*, 8(17):242–252, 2003.

[63] Asako Toyama, Kentaro Katahira, and Hideki Ohira. Biases in estimating the balance between model-free and model-based learning systems due to model misspecification. *Journal of Mathematical Psychology*, 91:88–102, 2019.

[64] Claire M Gillan, Michal Kosinski, Robert Whelan, Elizabeth A Phelps, and Nathaniel D Daw. Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *Elife*, 5:e11305, 2016.

[65] Paula Banca, Martin D Vestergaard, Vladan Rankov, Kwangyeol Baek, Simon Mitchell, Tatyana Lapa, Miguel Castelo-Branco, and Valerie Voon. Evidence accumulation in obsessive-compulsive disorder: the role of uncertainty and monetary reward on perceptual decision-making thresholds. *Neuropsychopharmacology*, 40(5):1192–1202, 2015.

[66] Alec Solway, Terry Lohrenz, and Read Montague. Loss aversion correlates with the propensity to deploy model-based control. *Frontiers in neuroscience*, 13:915, 2019.

[67] Nuria Doñamayor, Daniela Strelchuk, Kwangyeol Baek, Paula Banca, and Valerie Voon. The involuntary nature of binge drinking: goal directedness and awareness of intention. *Addiction biology*, 23(1):515–526, 2018.

[68] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[69] A Ross Otto, Samuel J Gershman, Arthur B Markman, and Nathaniel D Daw. The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological science*, 24(5):751–761, 2013.

[70] A Ross Otto, Candace M Raio, Alice Chiang, Elizabeth A Phelps, and Nathaniel D Daw. Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences*, 110(52):20941–20946, 2013.

[71] Charles D Spielberger, Sumner J Sydeman, Ashley E Owen, and Brian J Marsh. *Measuring anxiety and anger with the State-Trait Anxiety Inventory (STAI) and the State-Trait Anger Expression Inventory (STAXI).* Lawrence Erlbaum Associates Publishers, 1999.

[72] Alec Solway and Matthew M Botvinick. Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. *Psychological review*, 119(1):120, 2012.

[73] Todd A Hare, Wolfram Schultz, Colin F Camerer, John P O'Doherty, and Antonio Rangel. Transformation of stimulus value signals into motor commands during simple choice. *Proceedings of the National Academy of Sciences*, 108(44):18120–18125, 2011.

[74] Li Yan McCurdy, Brian Maniscalco, Janet Metcalfe, Ka Yuet Liu, Floris P De Lange, and Hakwan Lau. Anatomical coupling between distinct metacognitive systems for memory and visual perception. *Journal of Neuroscience*, 33(5):1897–1906, 2013.