Modeling Multiple Valuation Systems in Human Decision Making

Rong Guo

Modeling Multiple Valuation Systems in Human Decision Making

vorgelegt von
M.Eng.
Rong Guo
geb. in Xi'an, China

von der Fakultät IV – Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
- Dr. rer. nat. -

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. Manfred Opper
Gutachter:	Prof. Dr. Klaus Obermayer
Gutachter:	Prof. Dr. Felix Blankenburg
Gutachter:	Dr. Jan Gläscher

Tag der wissenschaftlichen Aussprache: 27. May 2015

Berlin 2015

To my husband Xu He

“Always be positive and persistent.”

- Prof. Dr. rer. nat. Klaus Obermayer, 2009 Berlin

(The simple but truly survival guide on my PhD journey)

ABSTRACT

Humans may consider various sources of information when making a decision. Traditional reinforcement-learning algorithms mainly focus on learning the expected reward and ignore other psychophysiological factors that may affect human decisions, such as perceptual interference or emotional regulation. This thesis aims to integrate these other factors into the reinforcement-learning models and addresses two questions: (1) How do conflicting salient stimuli influence reward estimation? (2) How are the counterfactual consequences integrated into economic decision-making? I hypothesize that the neurobiological mechanism of error-correction via reinforcement is commonly utilized by multiple valuation systems.

In the study of contextual modulation of prediction-error representations, I designed a value-based choice paradigm that dissociated stimulus-based and reward-based expectations. Participants traded off reward against the predictability of the stimulus location. Behavioral results were analyzed on a trial-by-trial basis using two independent Rescorla-Wagner models, which were then combined by a non-linear weighting function. Using model-based fMRI analysis, I found a co-existence of stimulus and reward prediction errors in the ventral striatum, suggesting that this brain region responded to surprising perceptual events as well as unexpected reward delivery or omission. Furthermore, the amygdala activity correlated with the weighting function, suggesting that it might be negotiating between the initial stimulus saliency based choices and the later reward-driven choices.

In the study of valuation with counterfactual learning signals, I extended the Q-learning model by incorporating both counterfactual gains and losses into fictive temporal-difference prediction errors. The model was used to investigate the potential influence of counterfactual valuation using both behavioral and fMRI data from a strategic sequential investment paradigm. The results demonstrated that counterfactual learning signals improved the Q-learning model fit, and this improved model predicted BOLD signal changes that correlated with expected value and reward prediction. Expected values derived from the model robustly modulated activity in the ventral medial prefrontal cortex and orbital frontal cortex. Furthermore, the model showed that individuals had different sensitivity to counterfactual gains and losses, which led to distinct neural correlations with fictive prediction error in the ventral striatum.

Together these two studies highlighted the neural correlates of multiple prediction errors in the ventral striatum and re-interpreted them in the form of an *information prediction error*, thus integrating the multiple valuation systems into a single coherent decision-making framework.

ZUSAMMENFASSUNG

Menschliche Entscheidungen basieren wahrscheinlich auf einer Vielzahl von Einflüssen. Traditionelle Algorithmen zum „Reinforcement Learning“, von welchen häufig angenommen wird dass sie diesen Entscheidungen zugrunde liegen, beschränken sich dagegen auf das Erlernen der mittleren zu erwartenden Belohnung und ignorieren dabei andere psychophysiologische Faktoren wie Wahrnehmung oder emotionale Kontrolle, welche menschliche Entscheidungen nachweislich ebenso beeinflussen. Diese Doktorarbeit hat das Ziel, mit Markov-Entscheidungsprozessen und „Reinforcement Learning“-Modellen experimentelle Hypothese in den Neurowissenschaften zu formulieren und zu testen. Sie integriert hierbei zwei dieser Faktoren in Modelle der Entscheidungsfindung und deren angenommenen neuronaler Korrelate: (1) Wie beeinflussen im Konflikt zur Gewinnmaximierung stehende saliente Reize die Schätzung von Belohnung, und (2) wie werden entgangene Gewinne und Verluste - die sogenannten „kontrafaktische Folgen“ - in ökonomische Entscheidungen integriert? Meine Hypothese ist, dass die neurobiologischen Mechanismen der Fehlerkorrektur, welche dem Reinforcement Learning zugrunde liegen, gleichzeitig in mehreren Bewertungssystemen involviert sind.

Ich entwarf eine Studie, welche stimulusbasierte und belohnungbasierte Erwartungen und deren Vorhersagefehler dissoziiert, indem diese in einer traditionellen belohnungbasierten Aufgabe unabhängig voneinander manipuliert wurden. Versuchsteilnehmer wogen hierbei die Vorhersagbarkeit von Stimuli gegen die zu erwartende Belohnung ab. Jede Entscheidung der Teilnehmer wurde mittels zweier unabhängiger Rescorla-Wagner-Modelle analysiert, deren Vorhersagen durch eine nichtlinearen Gewichtungsfunktion kombiniert wurden. Eine modellbasierte fMRT-Analyse fand die Vorhersagefehler sowohl für Stimuli als auch für die erwartete Belohnung im ventralen Striatum. Dies deutet darauf hin, dass diese Hirnregion sowohl auf überraschende Wahrnehmungsereignisse, als auch auf unerwartete Belohnung reagiert. Außerdem korrelierten die individuellen Gewichtungsfunktionen mit der Aktivität der Amygdala, was darauf hindeutet, dass diese Gehirnregion möglicherweise zwischen den ursprünglichen stimulusbasierten Entscheidungen, und den späteren von Belohnung getriebenen Entscheidungen abwägt.

In einer Studie zu kontrafaktischen Lernsignalen habe ich ein klassisches Q -Learning-Modell durch die Einbeziehung von entgangenen, „kontrafaktischen“ Gewinnen und Verlusten erweitert. Das Modell wurde verwendet, um anhand von Verhaltens- und fMRT-Daten den Einfluß von kontrafaktische Bewertungen auf den Entscheidungsfindungsprozess in einem sequentiellen Investitions-Paradigma zu untersuchen. Die im Modell integrierten kontrafaktischen Lernsignale konnten hierbei das Verhalten der Versuchsteilnehmer und die BOLD-Signale im fMRT, welche mit dem Erwartungswert und der Gewinnvorhersage korrelieren, deutlich besser vorhersagen als das klassische, rein „faktische" Model. Der aus dem erweiterten Modell abgeleitete Erwartungswert moduliert die Aktivität im

ventralen medialen präfrontalen Kortex und orbital-frontalen Kortex. Darüber hinaus zeigte das Modell, dass Personen unterschiedliche Empfindlichkeit gegenüber entgangenen Gewinnen und Verlusten haben, welche mit unterscheidbaren neuralen Korrelationen von fiktiven Vorhersagefehlern im ventralen Striatum einhergehen.

Zusammengenommen unterstreichen diese beiden Studien die Koexistenz von mehreren Vorhersagefehlern im ventralen Striatum und interpretiert diese als Spezialfälle eines allgemeinen Informations-Vorhersagefehlers. Diese Sichtweise integriert mehrere Bewertungssysteme in eine kohärente Interpretation von menschlichen Entscheidungsfindungsprozessen.

ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to Klaus Obermayer for all the training, support, and guidance about all aspects of my scientific career. Sometimes conducting a research project involving both theoretical and experimental work can be tough, but Klaus' persistent patience and constant mentoring made me through. I guess he must have seen some great potential when he offered me such a wonderful chance to live in Berlin. Were it not for the feeling that one has to graduate at some point, I would happily stay on for several more years working with Klaus.

I was very fortunate to work with Jan Gläscher in Hamburg during the later half of my PhD. I would have had none of the skills necessary to carry out the fMRI study in this thesis without his extensive guidance. Jan's big enthusiasm in data analysis greatly shaped my thinking about decision neuroscience. He is also a warm-hearted big brother and a great life mentor. His brilliant idea and wisdom deeply inspired me to try to become a scientist. In addition, Michael Spezio, Sam Chien and Martin Hebart in Jan's lab provided a great deal of help with my project and made my lab stay fruitful and enjoyable. I would also like to thank Michael J. Tobia and Tobias Sommer-Blöchl. Some of our collaborative work is presented in this thesis. Tobias arranged my lab stay in Hamburg and encouraged me to get into fMRI study, without his great help, I would not have been able to manage the thesis work in such a short time.

I am also very grateful to Felix Blankenburg for severing as a member on my thesis committee. I had my first series of fMRI lectures and SPM tutorials from Felix during the AAND course in the first year of my PhD at BCCN. That was the beginning of my fMRI research. I really appreciate his time reading this thesis, which made it possible for me to arrange my defense in such a short time.

I would also like to thank my lab mate Wendelin Böhmer, who was the first person read my thesis draft. His idea about reinforcement learning and research in general greatly influenced my academic agenda. There were also other people who took their time to read drafts of this thesis and gave me very useful comments: Sam Chien, Ana Sušac, Yun Shen, Johannes Mohr, Martin Hebart, Timm Lochmann, Konstantin Mergenthaler. In addition, this thesis would have not been possible without the support of many colleagues and friends who has been very kind with their time and expertise: Arno Onken, Marcel Stimberg, Sinem B. Beylergil, Josef

Ladenbauer, Moritz Augustin, Seo Sambu, Audrey Houillon, Fang Hui, Zhang Yan, Li Qiang, ShanYao, Kim Jisung, Shi Liang, Yanfang Song.

I would also like to deeply thank Camilla Bruns. Her excellent professional skills made my daily work worry free. I am in debt to Vanessa Casagrande and Robert Martin, who integrated me into the GRK PhD program and helped me to quickly adapt to the scientific and wonderful life in BCCN Berlin.

I thank all my family, my grandpa, my loving parents, and my parents-in-law whose love and support keep me going. Lastly, this thesis is dedicated to my dear Xu He, he deserves all the credits for the tremendous effort that he has spent in supporting me to finish this thesis.

During the work on this thesis, I was funded by Chinese Scholarship Council through a PhD scholarship and by German Federal Ministry of Education and Research through the grants GRK 1589 and Bersteinfokus Lernen TP1, 01GQ0911.

CONTENTS

Chapter 1: Introduction.......... 1

1.1 Background and general framework.......... 2

1.1.1 Reinforcement learning in the human brain.......... 2

1.1.2 Multiple valuation systems in learning and decision-making.......... 6

1.2 A hypothesis of multiple prediction errors.......... 9

1.3 The structure of this thesis.......... 10

Chapter 2: Using computational models to understand human decision-making.......... 13

2.1 Reinforcement learning.......... 14

2.2 A formal framework for learning from reinforcement.......... 15

2.3 Dopamine and the temporal-difference hypothesis.......... 18

2.3.1 Rescorla-Wagner learning.......... 18

2.3.2 Temporal-difference learning.......... 20

2.4 Beyond the temporal-difference hypothesis.......... 25

2.4.1 Model-based reinforcement learning as cognitive search.......... 25

2.4.2 Partially observable MDP incorporating perceptual uncertainty.......... 27

2.5 Thesis work relating to computational modeling.......... 30

Chapter 3: Model-based fMRI analysis.......... 31

3.1 What are we measuring with fMRI?.......... 32

3.1.1 fMRI physics.......... 32

3.1.2 Neural activity and BOLD fMRI.......... 34

3.2 fMRI preprocessing and basic analysis principles.......... 37

3.2.1 Preprocessing of fMRI data.......... 37

3.2.2 Basic statistical analysis.......... 39

3.3 Combining reinforcement-learning theory with fMRI data.......... 41

3.4 Parameter estimation and model comparison.......... 43

3.4.1 Maximum likelihood estimation.......... 45

3.4.2 Hierarchical Bayesian modeling.......... 47

3.5 Thesis work relating to fMRI experiment.......... 50

Chapter 4: Estimating prediction errors in paradigms of multiple valuation systems.......... 51

4.1 The encoding of higher-order reward prediction errors.......... 52

4.1.1 Introduction.......... 52

4.1.2 Methods 52
4.1.3 Results 58
4.1.4 Discussion 63
4.2 The interaction between stimulus likelihoods and rewards 66
4.2.1 Introduction 66
4.2.2 Methods 66
4.2.3 Results 73
4.2.4 Discussion 74
4.3 Fictive and factual prediction errors in strategic learning 75
4.3.1 Introduction 75
4.3.2 Methods 76
4.3.3 Results 80
4.3.4 Discussion 83
4.3.5 Supplementary material: examining the property and convergence of the FPE-Q model 84
4.4 Summary 85

Chapter 5: Contextual modulation of prediction error representations 87

5.1 Introduction 88
5.2 Methods 89
5.2.1 Participants 89
5.2.2 Experimental design 89
5.2.3 Experimental task and procedure 90
5.2.4 Imaging data 91
5.3 Results 93
5.3.1 Behavioral analysis 93
5.3.2 Computational model-based analysis 95
5.3.3 fMRI results 101
5.4 Discussion 102

Chapter 6: Neural system for valuation with fictive prediction error 107

6.1 Introduction 108
6.2 Methods 109
6.2.1 Experimental paradigm 109
6.2.2 Computational modeling 113
6.3 Results 114
6.3.1 fMRI data protocol and processing 114

6.3.2 fMRI results ... 115

6.4 Discussion ... 115

Chapter 7: Conclusions and general discussion ... 119

7.1 Multiple prediction errors in human basal ganglia ... 120

7.1.1 Summary of behavioral results ... 120

7.1.2 Summary of modeling results ... 121

7.1.3 Summary of fMRI results ... 122

7.2 Future directions: expectation, value, and attention ... 123

Bibliography ... 125

Appendices ... 143

Appendix 1 Prepublication ... 145

Appendix 2 Eidesstattliche Erklärung ... 147

Chapter 1: INTRODUCTION

"Follow your heart" or "use your head"—whichever strategy we use, we usually have more than one alternative to choose from, and yet we often make irrational decisions. Evidence to support this has been found in a variety of psychology and economics studies. This thesis aims at using computational neuroscience to address how the human brain generates paradoxical decisions. Assuming the human brain is a computing device, we can then use computational models to study the neural mechanism of the decision-making processes in a more precise way. Indeed, reinforcement-learning models that account for neural activity underlying different valuation systems in the human brain have been successfully adopted.

In this chapter, I will mainly review three valuation systems in human decision-making, detailing their putative neuroanatomical and computational underpinnings. I will also describe situations under which different valuation systems might interact with each other in a manner that influences behavior in either adaptive or maladaptive ways. Understanding their interaction may provide key insight into such pathological disorders of decision-making as schizophrenia, addiction, depression, and anxiety. I propose that error correction via reinforcement is a common neural mechanism underlying different valuation systems.

1.1 Background and general framework

1.1.1 Reinforcement learning in the human brain

First, we define *decision-making* as choosing among actions based on their relative values of potential consequences. Accordingly, optimal decision-making means to maximize *reward* and minimize punishment. Reward can be food, juice, money or anything that is attractive to the decision-maker and so may serve such roles as an incentive. In machine learning, rewards are simply numerical scalars (either positive or negative) that indicate the consequences of actions. It is not easy to take the optimal action, especially when reward or punishment may depend on a series of actions. Therefore, it is important to learn from experience and errors. This type of learning through *trial and error* is exactly the core of reinforcement learning: the learner makes a prediction, observes actual events and, if the prediction was wrong, updates the knowledge base so that future predictions are more accurate. In this thesis, reinforcement-learning models are used to study the human decision-making process. Unless otherwise stated, I will use the term *reward-based learning* interchangeably with *value-based decision-making* and *reinforcement-learning valuation*.

Reinforcement learning has a rich history in psychology, traceable to the early twentieth century, when Thorndike proposed the theory of stimulus-response associative learning (Thorndike, 1933). Slightly later on, Hebb proposed a neurobiological theory of learning (Hebb, 1949), claiming that the synaptic connections between neurons are strengthened after repeated simultaneous activation. In the late twentieth century, reinforcement learning (Sutton and Barto, 1998) became the subject of active research in the field of machine learning, meanwhile, neural validity of reinforcement-learning models has been demonstrated in dopamine neurons that are located in the midbrain nuclei of the substantia nigra and the adjacent ventral tegmental area (Montague et al., 1996; Schultz, 1998). Schultz and colleagues showed that the phasic response of dopamine neurons recorded from primates resembles the *temporal-difference prediction error* used in reinforcement-learning models (Schultz et al., 1997). Notably, the effort devoted to testing the involvement of the dopamine system in reward learning was initially motivated by research in Parkinson's disease (Fearnley and Lees, 1991; Graybiel et al., 1994). The brain regions damaged in Parkinson patients overlap largely with the dopamine-related regions identified by self-stimulation (Olds, 1958) and pharmacological studies (Wise and Rompre, 1989).

The dopamine neurons primarily project to such regions in the brain as the striatum, amygdala and frontal cortex. These widespread projections make dopamine neurons an ideal broadcast center for learning signals. The striatum, pallidum, subthalamic nucleus and substantia nigra together form the basal ganglia. The striatum also receives projections from cortical areas, as well as from the amygdala and hippocampus. This cortical-basal ganglia circuit is at the heart of decision-making valuation systems. The functional anatomy and connectivity of dopamine-related regions are shown in Figure 1.1. Nevertheless, a complete picture of the dopamine-dependent learning system may be more complex and diverse.

Human neuroimaging studies have revealed activities in both ventral and dorsal striatum consistent with prediction-error signals during a variety of decision-making tasks (Cooper et al., 2012; Kim et al., 2006; McClure et al., 2003; O'Doherty et al., 2004; Pessiglione et al., 2008). The anterior cingulate cortex is commonly involved in conflict monitoring (Botvinick et al., 1999, 2004). More recently, learning signals have also been found to be computationally characterized in the amygdala (Li et al., 2011; Prévost et al., 2011). Value representations have been suggested in the ventromedial prefrontal cortex (vmPFC), orbital frontal cortex (OFC) and intraparietal sulcus (Daw et al., 2006; Gläscher et al., 2009; Hare et al., 2008; Valentin et al., 2007). Levy and Glimcher conducted a meta analysis using data from thirteen different human fMRI studies published in recent years (Levy and Glimcher, 2012). The results suggest that a subregion of the vmPFC/OFC represents subjective values of different types of rewards on a neural common scale for guiding choice behavior. They proposed one possible schema for understanding the decision-making networks of the human brain, as shown in Figure 1.2. The schema suggests that sensory information from cortical and subcortical structures converges toward a single common value representation before passing on to the choice-related motor control circuitry. However, it is not clear whether striatum encode different types of prediction errors in a common scale as well.

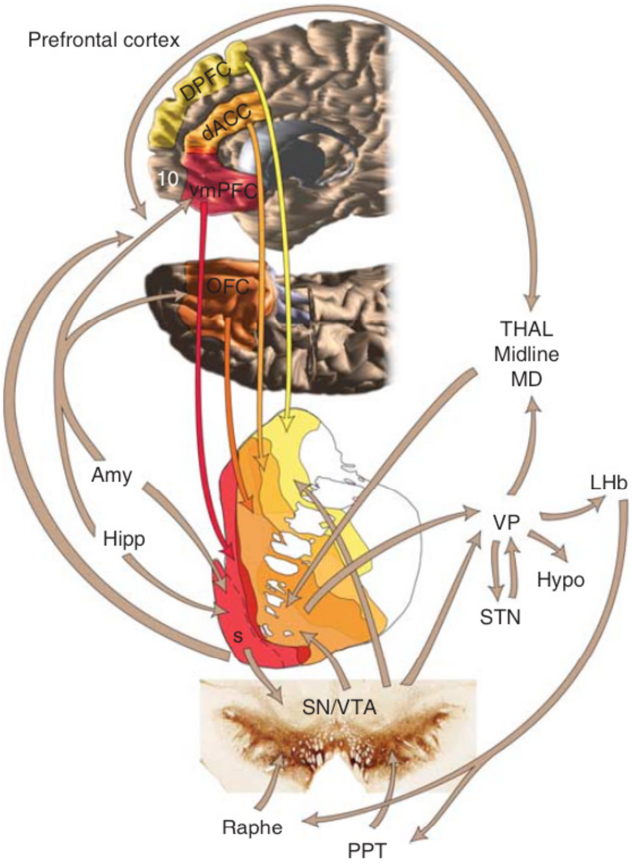

Figure 1.1 Schematic illustration of the anatomy and connectivity of reward circuit in the human brain. Dopamine neurons in substantia nigra (SN) and the adjacent ventral tegmental area (VTA) project to ventral and dorsal striatum, orbital frontal cortex (OFC), ventral medial prefrontal cortex (vmPFC), dorsal anterior cingulate cortex (dACC), dorsal prefrontal cortex (DPFC). Other abbreviations: Amy = amygdala; Hipp = hippocampus; THAL = thalamus; MD = medial dorsal; LHb = lateral habenula; VP = ventral pallidum; Hypo = hypothalamus; STN = subthalamic nucleurs; Raphe = raphe nuclei; PPT = pedunculopontine nucleus. Taken from (Haber and Knutson, 2010).

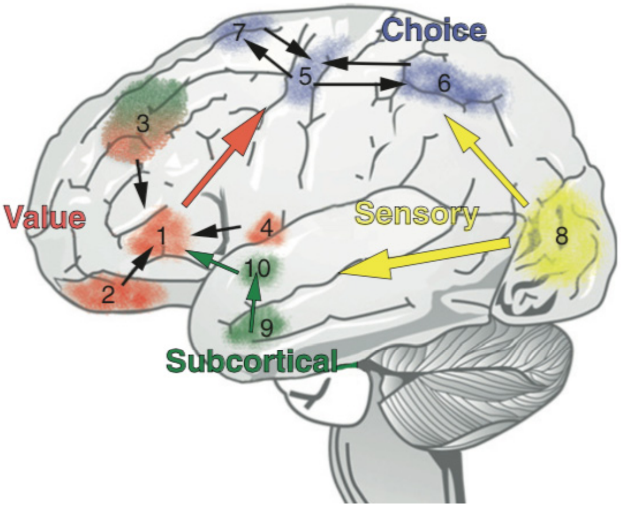

Figure 1.2 Schematic illustration of putative decision-making networks of the human brain. Information from cortical (3 dorsal lateral prefrontal cortex, 8 visual cortex) and subcortical (9 amygdala, 10 striatum, 4 insula) structures converges toward a single common value representation (1 ventral medial prefrontal cortex, 2 orbital frontal cortex) before passing on to the choice-related motor control circuitry (5 primary motor cortex, 6 posterior parietal cortex, 7 frontal eye fields). Sensory signals from visual areas, shown in yellow, stand for information from all sensory modalities. Taken from (Levy and Glimcher, 2012).

In summary, the functional neural anatomy of decision-making typically involves both cortical and striatal regions, including orbitofrontal, anterior cingulate and posterior parietal cortices, as well as the striatum, amygdala, and hippocampus. Reinforcement learning can capture a variety of human decision-making and reward-based learning behavior. The dopamine response and related neural activity in the cortical-basal ganglia circuit seem well accounted for by temporal-difference learning. Brain areas and functional mechanisms other than those described in this section may influence learning and decision-making processes as well. The human brain is a much more complex system than any state-of-art machinery. Characterization of the neurobiological computations that underlie sophisticated behaviors requires better integration of computer science, psychology, neuroscience, and economics than has been available heretofore; in other words, a multilevel and multidisciplinary research approach is needed. Recent emerging disciplines such as *neuroeconomics* (Glimcher, 2010; Glimcher and Fehr, 2014) and *computational psychiatry* (Montague et al., 2012; Wang and Krystal, 2014) have fostered a growing sense of interdisciplinary collaboration.

1.1.2 Multiple valuation systems in learning and decision-making

The hypothesis that human decisions are controlled by multiple valuation systems has been prevalent in psychology, neuroscience and behavioral economics (Balleine et al., 2009; Daw et al., 2005; Dayan and Balleine, 2002; Dickinson and Balleine, 2002; Dolan and Dayan, 2013; Tversky and Kahneman, 1981). The same decision can arise from distinct psychological and neurobiological pathways. The brain may employ the strategy that involves the least effort in a specific situation. In this section, I describe three types of valuation systems, which all ultimately concern reward-based learning and can be framed with reinforcement-learning models. These three systems are (1) an instrumental system that forms associative learning and goal-directed actions, (2) a Pavlovian system that triggers reflexive classical conditioning responses, and (3) a counterfactual learning system that evaluates actions while incorporating fictive outcomes.

1.1.2.1 Instrumental valuation system

Since the early days of psychology, theorists and experimentalists have struggled with the question of which associative structure controls human learning behavior. On the one hand, Thorndike proposed that actions are strengthened by positive reinforcement and weakened by negative reinforcement (Thorndike, 1933). This sort of instrumental valuation involves the formation of stimulus-response associations, which are initially strengthened by outcomes but eventually lead to outcome-independent habitual formation. On the other hand, Tolman used latent learning tasks to demonstrate the insufficiency of stimulus-response valuation and further proposed that animals can instead learn to plan goal-directed actions using an internal representation of environmental contingencies (Tolman, 1948). He called this internal representation a “cognitive map.” Abundant evidence collected over decades of behavioral and neuroscientific research indicates that habitual and goal-directed instrumental valuations not only cooperate but also compete with each other for control over decision-making (Dayan, 2008; Dolan and Dayan, 2013).

Instrumental valuation refers to behavior that aims at achieving a specific goal such as maximizing benefit or minimizing cost. In the machine learning literature, instrumental valuation means determining optimal policy in each state of the environment so as to maximize the value function. Reinforcement-learning methods can be broadly divided into two classes of algorithms: model based and model free. These two methods differ in their computational costs and adaptability to changes in the environment. The model-free algorithm is computationally inexpensive, but it fails to show rapid adaptation when the environmental contingencies change. In contrast, the model-based algorithm keeps track of an abstract model of the task structure and has the power to account for such choice behavior as post-training manipulation of reinforcer devaluation (Holland and Gallagher, 2004) or immediate changes in the reward contingency and outcome utilities (Hampton et al., 2006). The theoretical difference between model-based and model-free reinforcement-learning methods echoes the distinction between habitual learning and instrumental learning. In fact, recent work has begun to use model-based and model-free reinforcement learning as theoretical counterparts for studying the computational mechanisms underlying habitual and goal-directed learning, respectively.

If we suppose that the human brain employs both habitual and goal-directed instrumental valuations in parallel, an interesting question would be how the brain arbitrates different valuation systems when they disagree. It has been proposed that the two decision systems are arbitrated according to their respective reliability of estimation (Daw et al., 2005). Consistent with animal physiology studies, accumulating evidence from model-based reinforcement-learning computation has indicated that the human striatum and ventromedial prefrontal cortex are involved in goal-directed learning (Balleine and O'Doherty, 2010). Furthermore, the connectivity between the ventral medial prefrontal cortex and the dorsomedial striatum has been shown to correlate with individual control of the trade-off between goal-directed and habitual actions (Wit et al., 2012). Other studies have implicated the hippocampus (Bornstein and Daw, 2012) and parietal cortices in the encoding of learning signatures that are related to model-based reinforcement learning (Gläscher et al., 2010; Simon and Daw, 2011).

1.1.2.2 Pavlovian valuation system

Pavlovian valuation, which involves learning a cue-outcome association, is also called *classical conditioning* (Rescorla, 1987). By contrast with instrumental conditioning, Pavlovian responses cannot actually influence the outcome in the environment. Pavlovian valuation traditionally concerns learning predictive relationships between a neutral stimulus and a biologically relevant outcome. One example is the observation of Pavlov's dog salivating reliably to the ring of a bell that precedes the delivery of food (Pavlov, 1927). Nonetheless, Pavlovian valuation is not only a passive process that forms associations between co-occurring stimuli. Such associations often require sophisticated perceptual representations and the learning of contingent relations among events (Clark et al., 2012; Dayan and Berridge, 2014). Moreover, a learned Pavlovian association can in fact influence instrumental valuation.

The influence of Pavlovian valuation on instrumental learning has recently been studied in humans with the Pavlovian-to-instrumental transfer (PIT) paradigm (Bray et al., 2008; Crockett et al., 2012; Geurts et al., 2013; Hebart and Gläscher, 2014; Huys et al., 2011; Leanne et al., 2011; Prévost et al., 2012; Talmi et al., 2008). In the PIT paradigm, subjects are firstly trained passively by Pavlovian conditioning to associate a stimulus with reward and then by separately instrumental conditioning to learn actions for either the same or different reward. Afterwards, subjects are tested when different actions are available with the presentation of the Pavlovian stimulus but without the delivery of any outcome. The stimulus either augments performance of previously learned response in general (Dickinson and Balleine, 2002; Estes, 1948) or biases a specific action which leads to the same outcome that the stimulus was originally paired with (Balleine, 1992).

Early studies mainly suggested the cortical-basal ganglia circuit in different types of Pavlovian valuations (Clark et al., 2012). In addition, the role of amygdala is convincingly identified in stimulus-reward learning (Baxter and Murray, 2002; Roesch et al., 2010; Seymour and Dolan, 2008; Whalen and Phelps, 2009) and recent human fMRI studies with the PIT paradigm (Prévost et al., 2012; Talmi et al., 2008) further emphasize the functional role of the amygdala in learning and decision-making.

1.1.2.3 Counterfactual learning

In many situations individuals learn not only about the outcome of the chosen but also of the unchosen options. Via counterfactual thinking, the factual and the fictive outcome are compared and this may lead to a psychological emotion of regret when a better option was missed. Regret theory proposes that individuals are regret averse and therefore try to minimize potential regret, which can result in suboptimal, or in other words, irrational choices (Bell, 1981; Loomes and Sugden, 1982a; Zeelenberg et al., 1996). Along with the theory of regret, the impact of counterfactual thinking has attracted increasing attention in decision-making and game theory (Epstude and Roese, 2008; Hart and Mas-Colell, 2003; Marchiori and Warglien, 2008). The difference between factual and counterfactual outcome may serve as a learning signal analogue to the factual reward prediction error in reinforcement-learning models. Empirical studies lead to a more descriptive term of a *fictive prediction error* (Lohrenz et al., 2007) in computational modeling. The behavioral relevance of this fictive error signal is highlighted by its consistent impact on subsequent decisions and its impaired action guidance in chronic smokers and schizophrenics(Chiu et al., 2008; Roese et al., 2008).

Lohrenz and colleagues (Lohrenz et al., 2007) demonstrated that a fictive error signal contributes to changes in choice behavior and further suggested that situating a 'fictive error signal' within the theoretical framework of reinforcement-learning models may provide additional insight on decision-making process. Moreover, Li and colleagues modeled the counterfactual prediction error with a reinforcement-learning model in a two-alternative forced choice task, where the outcomes of chosen and unchosen options were both explicitly shown to subjects (Li and Daw, 2011). Although their results provide evidence in the human brain for a policy-specific update signal, they did not rule out the possibility that fictive error signals may directly influence instrumental valuation via modifying the expected long-term reward. Together, these studies show that choice behavior is responsive to counterfactual consequences, and variations of reinforcement-learning models can be applied to study the corresponding neural correlates. In particular, the orbitofrontal cortex(Camille et al., 2004; Coricelli et al., 2005; Liu et al., 2007) and the striatum (Chiu et al., 2008; Lohrenz et al., 2007) have been identified as potential sources of the fictive error signals.

1.2 A hypothesis of multiple prediction errors

As mentioned above, human neuroimaging studies have emphasized a central role of the striatal system in learning and value representations. In particular, reinforcement-learning models have been successfully integrated into the design and analysis of a variety of experiments. However, the main purpose of a reinforcement-learning agent is to maximize the expected reward. Thus, reinforcement-learning models cannot readily explain many decision-making and learning processes: one example of this concerns the behavioral preference towards non-rewarding perceptual stimuli. Another concern is the psychological influence on economic choices, such as the emotional regulation of regret.

The idea that decisions are driven by multiple valuation systems is supported by converging behavioral, neural and computational evidence, as discussed in the previous section. When multiple valuation systems influence decisions at the same time, behavior may become paradoxical because of the conflicts among these systems. By means of formulating reinforcement-learning models according to specific experimental context, we can examine the circumstances under which the decisions become non-optimal. For instance, we can modify reinforcement-learning models by including additional computational terms to account for experimental factors that may influence the estimation of the expected reward.

Most of the work in this thesis focuses on inferring the computational processes performed by distinct brain regions during reward-based learning. I hypothesize that the neurobiological mechanism of error-correction via reinforcement is commonly engaged in multiple valuation systems, although respective valuation may require a different *prediction error* signal. In particular, I will use human behavioral choice and neuroimaging data to test the following two hypotheses.

First, the same neural populations that encode the reward prediction error are recruited for encoding learning signals of value-nonspecific stimuli. Therefore, reward context can be extended to the unrewarding stimuli and induce sub-optimal choice behavior.

Second, counterfactual consequences can be incorporated into the temporal-difference prediction error term of reinforcement-learning models. The encoding of factual prediction errors and fictive prediction errors share a common computational mechanism to optimize the learning process.

1.3 The structure of this thesis

This thesis comprises five main chapters of theoretical framework and experimental results, which altogether test hypotheses regarding multiple valuation systems via reinforcement-learning models.

The current chapter is meant to provide relevant context of human decision-making, under which this thesis should be understood. I distinguished three putative valuation systems with respect to their functional neural anatomy. Most notably, the cortical and striatal brain regions are commonly involved in different valuation system. This leads to the question of whether each valuation system shares a common neural encoding of the prediction error signal.

Chapter 2 introduces the theoretical framework of Markov decision process and reinforcement learning with a focus on their putative neurophysiological implications. In particular, models implemented in this thesis are adapted to account for the expected reward, non-rewarding stimuli, and counterfactual outcomes. Each model is used to infer computational processes performed by the brain. This inference is made possible by incorporating model variables into the fMRI analysis.

Chapter 3 describes the experimental methodology of fMRI, especially in terms of model-based fMRI analysis. Since data acquisition and basic statistical analysis are relatively mature in fMRI, I mainly address some practical controversy in interpreting fMRI results. In addition, I also discuss some modeling techniques that are used in this thesis, including maximum likelihood estimation and hierarchical Bayesian modeling.

Chapter 4 presents behavioral findings from three experimental paradigms. Each paradigm involves choosing among options for monetary reward. In the first paradigm, the models that estimate the higher-order probabilities explain the choice behavior best, suggesting that multiple prediction errors are computed during sequential learning. The second paradigm dissociates the learning processes of simultaneous stimulus-response and action-outcome associations, which is used in the fMRI study presented in Chapter 5. The third paradigm studies the influence of the fictive prediction error in the complex strategic decision-making, which is used in the fMRI study presented in Chapter 6.

Chapter 5 presents results suggesting that the neural activity in the ventral striatum and the amygdala correlate with the computational characterization of a dual-learning process. I scanned participants with fMRI while they performed the learning task that is designed to dissociate the neural correlates of stimulus-based and reward-based expectations. The validity of a hybrid reinforcement-learning model is tested with both behavioral and neural data.

Chapter 6 examines the neural system involved in the valuation with counterfactual learning. I incorporated a 'fictive prediction error' into the Q-learning model for explaining human economic behavior in financial markets with information from both factual and counterfactual outcomes. The model-derived expected value and fictive reward prediction error respectively correlated with

BOLD signal changes in the ventral medial prefrontal cortex/orbital frontal cortex and the ventral striatum.

Chapter 7 details the conclusions and contributions of the thesis as a whole, as well as directions for future research. Different streams of research encompassed in this thesis provide evidence suggesting that multiple valuation systems share a common neurobiological mechanism of error-correction via reinforcement. Lastly, the work of this thesis paves the way for future theoretical and experimental investigations of both *perceptual* and *economic* aspects of decision-making via the reinforcement-learning framework.

Chapter 2: USING COMPUTATIONAL MODELS TO UNDERSTAND HUMAN DECISION-MAKING

Our ultimate goal is to pinpoint the neural activity underlying the highly adaptive cognitive processes of decision-making. The big challenge is that these processes are not stationary, as decisions largely depend on subjective preferences and might change during learning. To deal with this difficulty, computational models are used to draw a link between individual cognitive process and the responsible neural activity.

In this chapter, I describe the Markov decision process and reinforcement-learning models as theoretical frameworks for understanding behavioral and neural mechanisms underlying sophisticated human learning. Furthermore, I discuss some classic and recent studies that have applied reinforcement-learning models to analyze neural activity in reward-based learning. Recent attention has turned to uncover neural correlates of goal-directed instrumental learning with mode-based reinforcement-learning algorithms. I will highlight some on-going studies in this direction. After all, the study of human decision-making is developing at an overwhelming speed. I can only at my best review a sample of the most relevant studies. More efforts are required to integrate evidence from electrophysiological studies of animal learning and human neural imaging studies.

2.1 Reinforcement learning

We start with taking a backgammon player as an example for a reinforcement-learning agent. In a backgammon game, each move is informed both by anticipating a reply (i.e., reward) and by immediate judgments of the desirability of a particular board position (i.e., state). The essential feature of this example is learning via interacting with the environment. The decision-making agent, in this case the game player, seeks to win the play (i.e., achieve a goal) despite that the effect of every move cannot be fully predicted. The uncertainty about the environment is everything that the agent cannot control, such as a dice roll. Actions are selected based on both the *exploitation* of the agent's past experience and the *exploration* of certain unobserved part of the state-action space. This essentially defines the *trial and error* learning.

There are three basic elements in a formal reinforcement-learning framework: (1) Transition model: the environment where the subjects learn. The agent can learn the transition model to plan actions by considering possible future states before those are actually experienced. (2) Reward function: an immediate, possibly stochastic, payoff that results from performing an action in a state. The learning task for the agent is to optimize a sum or average of future rewards. (3) Policy and value function. Policy is how the learning agent maps perceived states of the environment to actions to be taken in those states. Value function is an estimate of the total, possibly discounted, reward expected in the future and it is computed to improve the policy.

Each of these elements will be discussed in more detail in the next section. Note that the reward in the formal reinforcement algorithm is different from the psychological notion of reward that often instead implies pleasure or hedonic impact. The reward in the reinforcement models can take both positive and negative numbers, while the negative can also be interpreted as a punishment. The goal of a reinforcement-learning agent is to take actions that maximize the expected future reward when traveling through different states by taking actions following a policy.

Reinforcement-learning theories mainly involve two disciplines. One is machine learning and optimal control, which have been largely addressed with dynamic programming. The other is animal learning, especially classical and instrumental conditioning rooted in psychology. This thesis aims at combining both machine learning and psychology aspects to further address neuroscience questions about where and how learning takes place in the human brain.

I will briefly summarize some of the basic reinforcement-learning models and try to distinguish them from both theoretical and experimental perspectives. The machine learning perspective deals with policy, value, reward, and state in order to find an optimal solution, whereas the neuroscience perspective tries to find the neural encoding of such learning signals as the reward prediction error or the expected value. Rather than attempt to provide an extensive review of state-of-art reinforcement-learning models, I will mainly focus on some remarkable similarities between the computations used by reinforcement-learning agents and the brain mechanism thought to be responsible for animal and human learning.

2.2 A formal framework for learning from reinforcement

Reinforcement-learning tasks can be formalized with the framework of *Markov decision process* (MDP) (Howard, 1960; Puterman, 1994), which is a class of discrete-time stochastic control processes in decision-making problems. An MDP consists of states, actions, transitions between states and a reward function. A set of environment states S is defined as a finite set $\{s^1, \dots s^N\}$, where the size of the state space is N. The set of actions that can be applied in a particular state $s \in S$ is denoted as $A(s)$ and $A(s) \subset A$. A is defined as a finite set $\{a^1, \dots a^M\}$, where the size of action space is M. By taking an action $a \in A$ in a state $s \in S$, the system is transitioned from state s to a new state s' ($s' \in S$) according to a probabilistic transition function $T(s, a, s')$. T is the probability of ending up in state s' by taking action a in state s, defined as $T: S \times A \times S \rightarrow [0,1]$. It is required that $0 \leq T(s, a, s') \leq 1$ and $\sum_{s' \in S} T(s, a, s') = 1$, for all states s and actions a.

Importantly, the state transition depends only on the current action and state, in other words, the new coming state does not depend on any previous actions or the history of visited states:

$$p(s_{t+1} | s_t, a_t, s_{t-1}, a_{t-1} \dots) = p(s_{t+1} | s_t, a_t), \tag{2.1}$$

where t is defined as a discrete time point in which decisions occur, s_t denotes the state at time t, and a_t denotes the action at time t. Equation (2.1) is the *Markov property*, which simplifies the solution of an MDP by allowing a formulation of the so-called *Bellman equation* (Bellman, 1957).

The learning agent takes actions to achieve rewards, which is defined by a reward function $R: S \times A \times S \rightarrow \mathbb{R}$. This reward function sets the goal of learning in an MDP. Given an MDP $\langle S, A, T, R \rangle$, the goal of gathering rewards can be achieved by means of computing optimal policies. The policy is a controlling element of the learning agent, which is defined as a function $\pi(s)$ that maps each state $s \in S$ to an action $a \in A(s)$. In this thesis, we only consider deterministic polices, i.e., $\pi: S \rightarrow A$. We set the goal of an MDP as maximizing the expected discounted sum of future rewards, i.e., maximizing $E[\sum_{t=0}^{\infty} \gamma^t r_t]$. The rewards r_t to be received in the future are discounted exponentially in their delay by the discount factor γ, with $0 \leq \gamma < 1$.

The value of taking an action a in a state s, following some policy π thereafter, can be written as a state-action value function $Q: S \times A \rightarrow \mathbb{R}$, which is:

$$Q^{\pi}(s,a) = E_{\pi}\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k} | s_t = s, a_t = a\right\}. \tag{2.2}$$

The expression in Equation (2.2) can be rewritten in a recursive form in terms of Bellman Equation:

$$Q^{\pi}(s,a) = \sum_{s'} T(s,a,s')(R(s,a,s') + \gamma Q^{\pi}(s', \pi(s'))), \tag{2.3}$$

and Q^{π} is the value given that the agent has taken an action a in the state s, and from there on follows the policy π. This value function can be used to improve the policy as following:

$$\pi'(s') = \underset{a'}{argmax} Q^{\pi}(s', a'), \tag{2.4}$$

and π' is the next improved policy. Such policy evaluation and improvement can find the best policy π^* in the end, that is, the policy that maximizes its value function. Similarly, we denote the value function under an optimal policy as Q^*. According to the Bellman optimality, the value of a state under the optimal policy equals the expected value for taking the best action in that state. Therefore, if we substitute Equation (2.4) into Equation (2.3), we have the following possible calculation for the Q^*:

$$Q^*(s,a) = \sum_{s'} T(s,a,s')(R(s,a,s') + \gamma \underset{a'}{max} Q^*(s',a')). \qquad (2.5)$$

Now that we have defined the MDP with an optimality criterion, the next step is to approach the question of computing the optimal policy. There are generally two families of reinforcement-learning models: one is called *model-based* algorithms and the other is called *model-free* algorithms. The *model* here means a model of the MDP, which is essentially determined by both the transition function T and the reward function R. Although these two names can be confusing, their distinction is simple: the model-based algorithms estimate the transition function and reward function as a complete description of the MDP; whereas the model-free algorithms estimate the value function directly from sampling. Both methods use Bellman equation and dynamic programming.

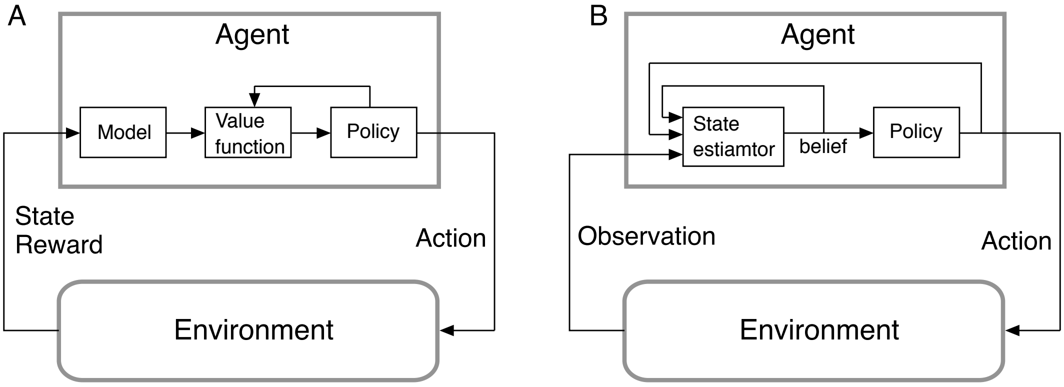

Figure 2.1 Illustrations of learning through MDP and POMDP frameworks. (A) Model-based reinforcement-learning agents firstly learn a model of the environment, and then use this model to compute a policy for selecting actions. Model-free reinforcement-learning agents skip the model procedure and directly estimate the value function from experience about the rewards of visited states. (B) POMDP agents cannot determine the state where they currently are. The agents make observations and estimate the belief state, which is updated by a state estimator based on previous beliefs, current observation, and the last action. Therefore, a policy is mapped from a sequence of observations to a probability distribution of actions.

A model-based reinforcement-learning system is outlined in Figure 2.1 A. The agent firstly learns a model of the environment and then uses this model to compute its value function. The computation can take forms of running a model-free algorithm to estimate the model, such as Monte Carlo Tree Search. Worth noting, the MDP framework assumes that the agent knows about the states of the environment with full certainty at all times. However, this might not be true in the real world, especially when the agent's perceptual abilities are imperfect and the agent is no longer able to observe the current state with complete reliability. This kind of problems, i.e., choosing optimal actions in partially observable stochastic environments, can be modeled by *partially observable Markov decision process* (POMDP) (Kaelbling et al., 1998). For comparison, the task of learning a POMDP is illustrated in Figure 2.1 B. Although the POMDP framework is more realistic in explaining real-world decision processes, most of our experiments have been designed according to the MDP framework for simplicity. Therefore, I focus on talking about MDP in the following and briefly discuss POMDP in the end of this chapter.

In this section, I have introduced the main components of reinforcement learning and the necessary background of Markov decision processes. The MDP is a straightforward way to describe how a learning process changes dynamically depending on the agent's reward experience. In fact, reinforcement learning as a research topic in the field of machine learning provides abundant algorithms to solve an MDP. Nonetheless, the major aim of this thesis is to infer human decision-making process with reinforcement-learning models rather than to implement efficient algorithms from a pure machine learning point of view. In the next section, I will illustrate some connections between basic reinforcement-learning concepts and the electrophysiological experiments.

2.3 Dopamine and the temporal-difference hypothesis

To show the links between reinforcement learning and the function of neural activity, we will start with evidence from classical conditioning experiments showing that learning is driven by the discrepancy between what was predicted and what actually happened. Such discrepancy is called *prediction error*. The idea of prediction errors is exactly the central tenet of reinforcement learning.

2.3.1 Rescorla-Wagner learning

In classical conditioning, animals (e.g., Pavlov's dog) learn to predict how outcomes (e.g., meat) are contingent on certain events (e.g., the sound of a bell) (Pavlov, 1927). For instance, Pavlov observed that his dog salivates to the sound of a bell after having been repeatedly exposed to a pairing of the ring and meat. The ring here is called a conditional stimulus and the meat is called an unconditional stimulus or a primary reward. In another experiment, Pavlov flashed a light to his dog whenever he ringed the bell. This time, the dog is trained with two simultaneous conditional stimuli (i.e., the ring and the flash) before the delivery of meat.

After having been firstly trained with the conditioning of a ring on the meat and then trained on the simultaneous conditioning of both a ring and a flash on the meat, the dog is then tested with a pairing of only the flash and the meat. Surprisingly, the dog does not salivate to the flash in this case. This suggests that the ring of the bell already explains the learning experience during training and therefore blocks out the learning of any relationship between the flash and the meat. This interesting behavior is referred to as the *blocking effect*.

To explain this puzzling learning behavior, Rescorla and Wagner proposed a formal theory of Pavlovian conditioning as following (Rescorla and Wagner, 1972):

$$V_{k+1}(\mathrm{s}_i) = V_k(\mathrm{s}_i) + \alpha \left[r_k - \sum_j V_k(\mathrm{s}_j) \right]. \tag{2.6}$$

On each trial $k + 1$, the value of every conditional stimulus s_i is denoted as $V(s_i)$. $V(s_i)$ is updated with the difference between the actual outcome r_k and the sum of all the predictions from different stimuli $\sum_j V_k(s_j)$. j indicates each of the conditional stimuli. α is a learning rate that determines the size of the update steps and $0 < \alpha < 1$. Taken from Equation (2.6), the prediction error that drives learning is defined as:

$$\delta_k(\mathrm{s}_i) := r_k - \sum\nolimits_j V_k(\mathrm{s}_j). \tag{2.7}$$

The blocking effect in the light and bell example can be explained simply by Equation (2.6). We denote s_1 as the ring, s_2 as the light and r_k is the reward on each trial, i.e., the meat. The blocking effect can be explained by $\delta_k(s_2) = 0$, because the ring already fully predicts the meat. Therefore, the basic idea behind Rescorla-Wagner model (Equation (2.6)) is that leaning should occur only when observed events violate expectations. Note that here the notation of s indicates a stimulus, which is different from a state in the MDP described in the previous section. The model here primarily focuses on learning an association between a reward and the stimulus rather than learning an explicit state value.

The Rescorla-Wagner model has successfully explained a variety of behavioral phenomena. The basic learning unit of a Rescorla-Wagner model is the discrete experimental trial, in other words, sequential trails are treated as independent and identical. However, many practical decision tasks have sequential structures and a long-term goal of maximizing expected values. In these tasks, a Rescorla-Wagner model that only considers the immediate reward is no longer valid. Therefore, the Rescorla-Wagner model needs to be extended to account for temporal relations within a learning trial in sequential decision-making tasks.

2.3.2 Temporal-difference learning

Motivated by the approach applied in classical conditioning, Sutton and Barto (Sutton and Barto, 1998) proposed a temporal-difference learning rule based on the MDP framework described in the previous section. The temporal-difference learning rule divides each experimental trial into smaller time points. At each time point t, the reinforcement-learning agent experiences a state s_t, which produces a reward r_t. The goal of the agent is to estimate the value of a state $V(s_t)$ in terms of its cumulative future rewards. The prediction error in Rescorla-Wagner model (Equation (2.7)) is replaced with:

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t), \tag{2.8}$$

which is called a *temporal-difference prediction error* and the learning rule is:

$$V_{k+1}(s_t) = V_k(s_t) + \alpha\delta_t. \tag{2.9}$$

Unlike the Rescorla-Wagner learning rule, this learning rule considers not only the immediate reward r_t, but also accounts for the sum over all the rewards in the subsequent states, which is approximated by $\gamma V(s_{t+1})$. The term *temporal difference* comes exactly from the term $r_t + \gamma V(s_{t+1}) - V(s_t)$, which is a discrepancy between the state values at consecutive time points within a trial. Again, this learning rule is a model-free stochastic approximation of Q^* in Equation (2.5) when there is only one action, that is, no choice involved.

The temporal-difference model extends the discrete trial-level Rescorla-Wagner model onto a continuous-time learning. This crucial difference enables the temporal-difference model to explain some behavioral phenomena that the Rescorla-Wagner model is not able to capture, such as within-trial temporal relationships and second-order conditioning. For instance, the temporal-difference model predicts that a reward prediction error will be induced at the omission of a reward when the reward has already been indicated beforehand. This leads to the influential hypothesis that phasic dopaminergic firing patterns encode a temporal-difference reward prediction error (Montague et al., 1996; Schultz, 1998; Schultz et al., 1997). While several other studies and interpretations about if and how dopamine neurons affect behavior have also been proposed (Berridge, 2012; Smith et al., 2011; Tindell et al., 2009), of most interest to this thesis is the compelling evidence that dopamine neuron activity reflects a reward prediction error.

Dopamine neurons of the ventral tegmental area and substantia nigra report a reward prediction error, shown in Figure 2.2 A. A monkey is trained to touch a lever after the appearance of a small light in order to get a primary reward of fruit juice. In the initial phase of training, dopamine neurons respond with a burst of firing to the unexpected delivery of juice, shown in the top panel. After several days of training, the phasic burst of dopamine neurons is shifted to the presentation of the light as the monkey has learned to reach the lever as soon as the light is on, shown in the middle panel. After learning, if the monkey accidently fails to touch the lever and no reward is delivered, the activity of dopamine neuron is depressed below the background baseline of firing rate exactly at the time when the reward should have occurred, shown in the bottom panel. This phasic activity of dopamine neurons that changes with the prediction of reward exactly resembles a scalar prediction error signal in the temporal-difference learning (Schultz et al., 1997).

Fiorillo and colleagues further verified the idea that dopamine neurons encode reward prediction error by systematically manipulating the probability of reward in a classical conditioning experiment (Fiorillo et al., 2003). Monkeys are conditioned in a Pavlovian procedure with five distinct visual stimuli that predicte the delivery of a liquid reward with different probabilities of 0%, 25%, 50%, 75% and 100% as shown in each panel from top down of Figure 2.2 B. The phasic dopaminergic responses at the presentation of the reward decrease monotonically as the probability of reward increase, in other words, they decrease in response to the decline of the reward prediction error. For instance, dopamine neurons show no response to a fully predicted reward while the prediction error is zero, displayed in the bottom panel. On the one hand, this pattern of activity can be well explained by the Rescorla-Wagner model of Equation (2.6). On the other hand, dopaminergic responses also change at the presentation of the conditional stimulus, which reflects the property of the temporal-difference learning of Equation (2.8).

The hypothesis of reward prediction error tested in primate physiological experiments has provided a quantitative basis for the design and analysis of human fMRI experiments. Early studies mainly suggested that fMRI BOLD responses of human ventral striatum (i.e., nucleus accumbens, ventral and medial portions of putamen and caudate) represent what has been observed in the dopamine neurons of non-human primates. For instance, Abler and colleagues replicated the non-human primates study conducted in (Fiorillo et al., 2003) with a human fMRI experiment (Abler et al., 2006). Their results suggested that BOLD activity in human nucleus accumbens scales with the size of the reward prediction error (Figure 2.3) in the same manner as what have been observed in monkey's dopaminergic neurons, presented in Figure 2.2 B.

In another study, O'Doherty and colleagues have shown that the BOLD signals in human ventral striatum changes in accordance with a temporal-difference learning model during a reward learning task (O'Doherty et al., 2003). Furthermore, McClure and colleagues devised a paradigm in which a reward prediction error is induced by varying the timing of the reward delivery across trials. They found that the BOLD signal changes in putamen correlate with this temporal-specific prediction error (McClure et al., 2003). Again, sensitivity to timing is a key feature of temporal-difference learning. The temporal-difference learning is driven by the difference between temporally successive predictions.

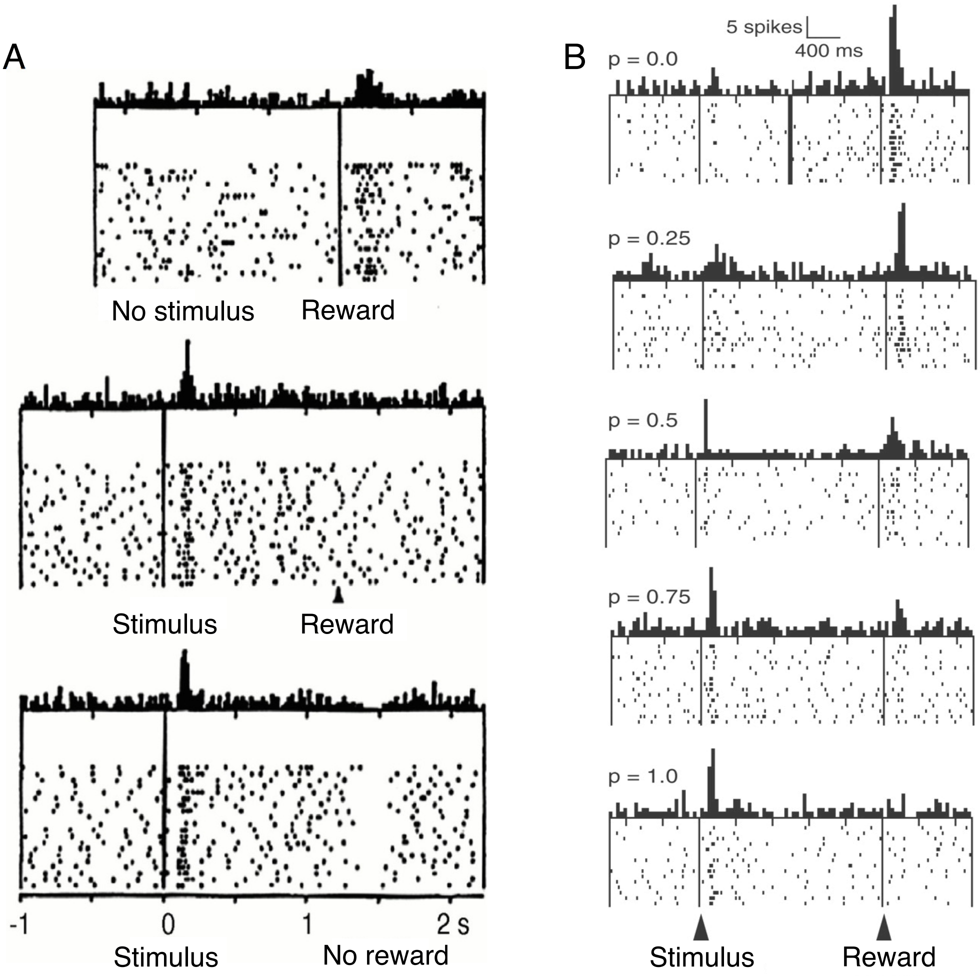

Figure 2.2 Firing patterns of dopamine neurons report temporal-difference prediction errors. Each panel shows raster and histogram of activity in a single cell, with each row of dots as one trial. (A) Neural activity is aligned at the delivery of a reward (top panel) or the onset of the stimulus (middle and bottom panels). In the initial phase of training, dopamine neurons respond with a burst of firing to the unexpected delivery of reward (top panel). After learning, the phasic burst is shifted to the presentation of the predictive stimulus. The reward occurs as expected, and hence no prediction error at the delivery of reward (middle panel). When the reward fails to occur, the activity of dopamine neuron is depressed at the time when the reward is expected. Adapted from (Schultz et al., 1997). (B) Dopamine neurons respond to the conditional stimuli and reward in accordance to a variety of reward probabilities p, increasing with 0.25 from the top to bottom panels. The top panel is spliced together from two situations, where reward is given in the absence of stimulus. The phasic dopamine responses decrease monotonically at the presentation of reward as the probability of reward increases from the top penal down; in other words, they decrease as the prediction error decline. Adapted from (Fiorillo et al., 2003).

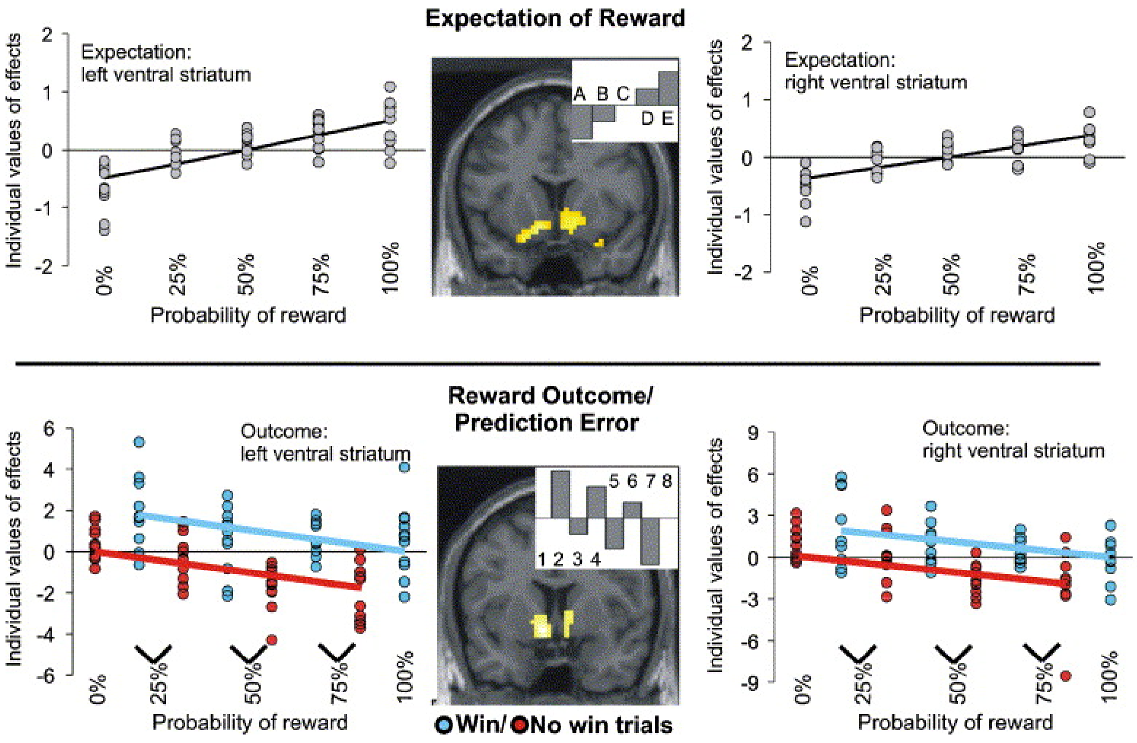

Figure 2.3 fMRI activity in the nucleus accumbens modulated by reward prediction errors in the same manner as what have been observed from dopamine neurons in non-human primate shown in Figure 2.2 B. Upper part: Five experimental conditions A-E with probability of reward from 0 to 1, increasing with a step of 0.25. fMRI activity in the nucleus accumbens linearly increases at the presentation of reward-predictive stimulus as the probability of reward increase. Lower part: fMRI activity in the nucleus accumbens linearly declines as the reward prediction error decreases (1, 3, 5, 7, red: reward expected at probability 0–75%, reward omission inducing negative prediction error; 2, 4, 6, 8, blue: reward expected at probability 25–100%, reward delivery inducing positive prediction error). Taken from (Abler et al., 2006)

The studies reviewed above all used classical conditioning tasks that do not involve action selection. However, in many situations, the ultimate goal of learning is not only to make predictions from observations, but also to achieve more rewards as well. This leads to another type of interaction with the environment called *instrumental learning*. The temporal-difference prediction error in Equation (2.9) is used in early work of Actor-Critic learning (Barto et al., 1983; Konda and Tsitsiklis, 2003) to strengthen or weaken the selection of particular actions. Pioneering studies linked Actor-Critic learning model to the function of basal ganglia (Joel et al., 2002) in instrumental action selection. Later on, converging evidence has indicated the correlates between BOLD fMRI response of human striatum and variant versions of temporal-difference prediction error during instrumental learning tasks (Abler et al., 2006; Delgado et al., 2005; Diuk et al., 2013; Gershman et al., 2009; Li and Daw, 2011; Niv et al., 2012; Schönberg et al., 2007; Seymour et al., 2004).

A commonly used model-free reinforcement-learning model is the Q-learning (Watkins and Dayan, 1992) model, which explicitly learns the state-action value. We can denote the state-action value as $Q(s, a)$. The learning rule is a variation on the theme of temporal-difference learning as following:

$$Q_{k+1}(s_t, a_t) = Q_k(s_t, a_t) + \alpha\delta_t, \tag{2.10}$$

where δ_t is the temporal-difference reward prediction error computed by:

$$\delta_t = r_t + \gamma \underset{a}{max} Q_k(s_{t+1}, a_t) - Q_k(s_t, a_t). \tag{2.11}$$

The learning takes place when the agent selects an action a_t and receives a reward r_t by moving from state s_t to s_{t+1}. In general, Q-learning algorithm will converge to an optimal policy if the learning rate α decreases properly and all the stat-action pairs are visited infinitely often.

2.4 Beyond the temporal-difference hypothesis

The function of dopaminergic neurons during reward-based learning seems to be well explained by the temporal-difference model. Nonetheless, formal theoretical suggestions from other reinforcement-learning models might also help to examine how the brain deals with complex decision problems. In this section, I review some of the current research that tries to map advanced reinforcement-learning concepts onto neurobiological underpinnings of human decisions. I will mainly discuss model-based reinforcement learning and partially observable Markov decision process (POMDP).

2.4.1 Model-based reinforcement learning as cognitive search

Psychology studies have distinguished two types of instrumental learning: *habitual* and *goal-directed*. Habitual learning dates back to the associative learning pioneered by Edward Thorndike and Ivan Pavlov (Pavlov, 1927; Thorndike, 1933). In Thorndike's study about law of effect, he trained his hungry cat to escape a puzzle box faster, and hence he argued that the reward has reinforced the association between the puzzle box and the cat's action of pressing a lever to escape. Goal-directed learning dates back to the latent learning proposed by Tolman (Tolman, 1948). Tolman firstly exposed his lab rats to a maze without reward. Next, these trained rats showed faster learning than naïve rats in finding the route to a reward in the maze. Tolman argued that the pre-trained rats have planed actions by using an internal representation of a "field map of the maze". These two learning strategies seem to echo the theories of model-free and model-based reinforcement learning algorithms described in Section 2.2 (Dayan and Niv, 2008; Doll et al., 2012; Ito and Doya, 2011; Schultz, 2013).

The habit learning has been successfully studied from the perspective of model-free reinforcement learning, especially the temporal-difference learning hypothesis that is supported by the observations about the activity of dopaminergic neurons. Empirical evidence for unifying the goal directed learning and model-based reinforcement learning has emerged in recent years. In particular, there have been studies showing distinct neural correlates related to model-based reinforcement-learning algorithms (Daw et al., 2011; Gläscher et al., 2010; Simon and Daw, 2011). In these studies, the transition function in Equation (2.5) is directly adapted to explain human choice behavior and to examine the supporting brain mechanism. Gläscher and colleagues used a latent learning paradigm (Gläscher et al., 2010) in an fMRI study, where subjects are firstly exposed to the state space of a probabilistic sequential Markov decision task and then are tested to make choices for monetary rewards. In their computational model-based analysis, they derived a *state prediction error* to estimate the transition function as:

$$T_{k+1}(s,a,s') = T_k(s,a,s') + \alpha\delta_{SPE}, \tag{2.12}$$

where δ_{SPE} is the state prediction error calculated as:

$$\delta_{SPE} = \sigma_{s,s'} - T(s,a,s'). \tag{2.13}$$

$\sigma_{s,s'} = 1$ for the observed transition and $\sigma_{s,s'} = 0$ for the unobserved transition.

This state prediction error δ_{SPE} well explains the fMRI BOLD response in human intraparietal sulcus and lateral prefrontal cortex during both the pre-training and test sessions of the experiment. At the same time, they also found correlates of the temporal-difference reward prediction error in the ventral striatum. This study demonstrates the possibility that both model-based and model-free reinforcement-learning strategies co-exist in the human brain.

Similarly, Daw and colleagues designed a two-staged Markov decision task, where human choice behavior can be distinguished according to whether or not their strategies have taken into account the transition structure of the task (Daw et al., 2011). The task design allowed them to distinguish behavior and neural substrates of different learning strategies. Interestingly, the results suggest that striatal BOLD activity reflect not only model-free reward prediction errors but also model-based reward prediction errors in terms of their respective contributions to choice behavior. A follow-up study by Wunderlich and colleagues (Wunderlich et al., 2012) further implicated dopamine involvement in the arbitration between model-free and model-based behavioral control via reinforcement learning. These results are consistent with electrophysiological studies in animals, showing that the function of dopaminergic neurons can be far more sophisticated than a basic temporal-difference learning rule (Bromberg-martin and Hikosaka, 2009; Kobayashi and Schultz, 2014; Schultz, 2013).

Given the hypothesis that both model-based and model-free valuations control choice behavior, the next question is how these systems compete or/and cooperate. Theoretical work (Daw et al., 2005) has proposed an arbitration process, suggesting that different valuation systems are arbitrated according to their respective reliability of estimation. Recent work by Lee and colleagues (Lee et al., 2014) has started to address the putative neural instantiation of this arbitration process. Their study suggests that the inferior lateral prefrontal and frontopolar cortex encode a Bayesian reliability estimation of the model-based state prediction error and the model-free reward prediction error, as well as a comparison between their corresponding reliability. This Bayesian reliability estimation can be interpreted as an arbitrator between model-based and model-free valuations. Their effective connectivity analysis further shows that the arbitrator appears to work by selectively gating the model-free system. This result is consistent with the idea that model-based valuation may require more computational power.

2.4.2 Partially observable MDP incorporating perceptual uncertainty

Studies about how the brain converts physical stimuli into perception and sensation have their roots in the field of psychophysics and Bayesian decision theory (Dayan and Daw, 2008). The ability of human observers to perceive physical properties is often quantitatively described by a so-called psychometric function (Klein, 2001). This function quantifies certain task performance to stimulus strength in a perceptual task, for example, correct response rate as a function of levels of image contrast. Subjects usually have to respond to or discriminate between some perceptual stimuli, such as whether a blurry image is a house or a face. These behavioral data are repeatedly measured across different level of the stimulus strength. The psychometric function typically has a form of a sigmoid curve, which is fitted onto the behavioral data, showing that the task performance improves monotonically across increasing level of stimulus strength. A particular level of stimulus strength can be defined as a threshold at which subjects' task performance switches between pure guessing and near optimal. Decisions made around the threshold are generally under strong uncertainty.

As mentioned in Section 2.2, most of the experimental studies inspired by reinforcement-learning framework largely rely on the assumption that the learning agent has a complete knowledge of states. These studies usually provide subjects with explicit instructions about the state information and use unique background images or cues to indicate each state. Subjects are instructed to estimate the state value either with or without learning the state transition probability. Accordingly, most of the neuroimaging studies mainly provide evidence addressing the neural signatures of value estimation rather than the state uncertainty. This leads to an interesting yet unanswered question: what is the neural representation of the *state* per se? It is unclear whether the state itself is somehow encoded before any value estimation. To answer this question, a most likely candidate of the abstract mathematical notation *state* is probably the physical stimuli that we observe and experience with perception and sensation everyday.

When cues are ambiguous or identical among different states, Bayesian inference is a straightforward way to relate states to observations. The posterior distribution over a state can be termed as a belief state. We denote a belief state by $b(s)$, which is a probability distribution over all the true states S. Thus, $0 \leq b(s) \leq 1$ for all the $s \in S$ and $\sum_{s \in S} b(s) = 1$. After executing an action a and observing the outcome o, the belief state b at current state s can be updated according to Bayesian rule:

$$b(s) \Leftarrow p(s|o,a,b) = \frac{p(o|s,a,b)p(s|a,b)}{p(o|a,b)}. \tag{2.14}$$

If we substitute the true states with such belief states, the optimal policy has to be calculated as a function of the belief states accordingly. A POMDP framework provides exactly a systematic method for such *belief estimation*.

The learning agent cannot directly observe the underlying state in a decision problem of POMDP, so it has to make an observation, estimate the belief state based on the current action and the previous belief states, and generate a new action afterwards, as presented in Figure 2.1 B. The agent's current belief state includes all the information about its past actions and observations. Thus, the transition and reward functions over belief states can be simplified so as to satisfy the Markov property. The remaining problem is to solve an MDP on the belief states. In practice, the belief states can become computationally intractable. The optimal solution of a POMDP with a large state space is still a vital ongoing research topic in machine learning, whereas far fewer empirical experiments have been conducted to test the neural validity of POMDP models.

It is not immediately clear how POMDP models can be translated into neural computations. As in the perceptual domain, the neural encoding of *belief states* concerns not only reward learning but also perceptual inference. The perceptual inference (Ding and Gold, 2013; Gold and Shadlen, 2002, 2007) has been largely studied in disjoint from reward-based decision-making, because the focus is primarily on perceptual uncertainty rather than on optimizing reward. One recent study (Rao, 2010) implemented a POMDP model to identify neural probabilistic representations for choosing actions that maximize expected reward. The model's predictions are consistent with the dopaminergic responses recorded when monkeys were performing a task with manipulations of both sensory properties and reward associations (Nomoto et al., 2010). The results postulate a mapping between POMDP inference and brain anatomy of decision-making networks. This mapping illustrates the neural plausibility of the POMDP model for unifying both perceptual and reward-based decisions. However, this mapping has so far not been tested in a direct manner with human fMRI experiment. Furthermore, it remains unclear whether perceptual and reward-based decisions are encoded in the same neural circuit and how they are compared and converted into motor control. Further experimental work is needed to characterize biological underpinnings of the computations that combine Bayesian inference and reward learning.

2.5 Thesis work relating to computational modeling

This thesis uses algorithmic ideas from machine learning to study decision-making processes in the human brain. The theoretical framework and classic models described in this chapter are foundations for the work in this thesis. These models offer a variety of possibilities to explain seemingly irrational choice behavior. In the next chapter, I introduce some experimental methodology for investigating the neural correlates of these models in the human brain.

In Chapter 4 of this thesis, I present three experimental paradigms and results of behavioral modeling. In the first paradigm, I adapted Rescorla-Wagner models to predict learning of higher-order temporal dependencies. In the second paradigm, I constructed a hybrid model by combining two Rescorla-Wagner models in parallel. This model allowed for testing dynamic interactions between two learning processes. In the third paradigm, I extended the standard Q-learning model to include a fictive prediction error during counterfactual learning. This model made predictions for choice behavior in a complex strategic sequential decision-making task.

Chapter 3: MODEL-BASED FMRI ANALYSIS

This thesis uses functional magnetic resonance imaging (fMRI) data to test the neural validity of computational models in explaining multiple decision-making processes. In the previous chapter, I have shown how the integration between neuroimaging data and computational models can identify neural mechanism of complex learning process in the human brain. In this chapter, I will describe in detail the fMRI data acquisition and analysis methods. Although it is not the purpose of this thesis to extensively understand fMRI physics or to explore how noise level may influence fMRI statistics, it is important to be aware of the actual capacities and limitations of this neuroimaging technology in interpreting the results from fMRI studies.

The basic idea of model-based fMRI analysis is to estimate the brain's cognitive process with a dynamic learning model and to seek correlations with the model's internal variables in the fMRI data. A critical part in this model-based analysis is to accurately estimate the model's internal variables from the experimental stimuli, subjects' choices and obtained rewards. Therefore, I will also discuss different parameter estimation and model fitting techniques in the end of this chapter.

3.1 What are we measuring with fMRI?

3.1.1 fMRI physics

fMRI relies on a set of elegant physical principles, including the proton's nuclei magnetic resonance (NMR) property and the processing of the MR signals by Fourier transform. The human brain contains abundant water molecules, which are composed of hydrogen and oxygen. A single proton of hydrogen nuclei possesses the NMR property (i.e., both a magnetic moment and an angular momentum) and is often referred as a *spin*. Imaging the human brain in an fMRI scanner, the net magnetization of the collection of spins in such a strong magnetic field is the basis for generating an MR signal. These spins precess around an axis in either parallel or antiparallel to the main static magnetic field. Here, the parallel and antiparallel axes are termed as *longitudinal plan* and *transverse plane*, which can be considered as low- and high-energy states, respectively. Usually, the net magnetization is a vector in the longitudinal plane, in other words, in the low-energy state. However, when the head coils send an electromagnetic pulse that oscillates at the resonant frequency of the spins (i.e., the Larmor frequency), the net magnetization vector is tipped from the longitudinal to the transverse plane. This process is called *excitation*. When such excitation pulse ceases, the spins release the additional energy and restore their longitudinal magnetization. In the meanwhile, the time-course of the energy release known as the free induction decay provides an MR signal that goes into the MR images.

On one side, the time constant that describes the recovery of the longitudinal component of net magnetization is called *T1 recovery time*. On the other side, the time constant that describes the decay of the transverse component of net magnetization due to the accumulated phase differences caused by the interactions among spins is called *T2 decay time*. These two time parameters are very important, because specifying a pulse sequence that targets one of these parameters can collect MR images that are sensitive to specific properties of the brain tissue. In addition, local magnetic field inhomogeneity also affect the decay of the transverse component. Hence, a combination of T2 effect and the additive effect caused by field inhomogeneity is defined as another time constant T2*, which is critical for BOLD fMRI.

Three spatial gradient fields superimposed on the main magnetic field are used in a sequence to change the strength of the magnetic field along a specific direction.

There are three directions: one longitudinal and the other two perpendicular in the transverse plane. Simultaneous application of a longitudinal gradient and a excitation pulse allows the selection of a specific slice within the imaging volume and the use of two perpendicular transverse gradients within the slice allows for an unique encoding of information about spatial locations. Then, the three-dimensional fMRI image is constructed from a set of these two-dimensional slices.

There are two important factors that govern the time at which MR images are collected: (1) the repetition time (TR), which is the time interval between successive excitation pulses, usually expressed in seconds; (2) the echo time (TE), which is the time interval between excitation and data acquisition, usually expressed in milliseconds. By controlling TR and TE, the MR signal from different brain tissue types can be manipulated. fMRI data can be considered as consisting of a three dimensional matrix of volume elements (i.e., voxels) that is repeatedly sampled over time. In summary, there are three types of MR images involved in this thesis:

1. T1-weighted images, representing the relative signal intensity of voxels depending on the T1 value of the tissue. For any two tissues that differ in T1, there is an optimal TR value that maximizes the difference. At the same time, the TE value has to be optimized to minimize the T2 contrast so that the T1 contrast image can be exclusively achieved. This image is used for providing high-resolution anatomical details of the brain.

2. T2*-weighted echo-planer image (EPI) images, which is the BOLD fMRI signal. EPI is a technique that allows the collection of an entire two-dimensional image by changing spatial gradients rapidly following a single excitation. Importantly, the EPI pulse sequence allows us to collect functional images approximately at the same rate as the physiological changes of interest. While the spatial encoding is achieved through gradient fields, the magnetic field inhomogeneities at boundaries between air and tissues can shift voxels in space. This may lead to some degree of geometric distortions in the image. One practical method for correcting such spatial shifts involves the use of a gradient echo field map.

3. Gradient echo field map, which is derived from two images acquired at slightly different TEs, and are used subsequently to remove spatial distortions from the EPI images.

3.1.2 Neural activity and BOLD fMRI

When thinking about neural activity, the first thing that comes to mind is probably the action potentials, which typically form spike trains. Critically, a group of neurons can generate spike trains in oscillatory symphony through local interactions between excitatory and inhibitory neurons. In addition to direct synaptic interactions between neurons forming a network, oscillatory activity can also be modulated by neurotransmitters on a much slower time scale. However, as I search from the firing rate of a single neuron to the oscillations of neural ensembles, the meaning of a plain term neural activity becomes a big dictionary of neural signatures. One thing in common among these neural signatures at different levels is that they all contribute to the energy consumption after all. Specifically, their metabolic demand evokes changes of cerebral blood flow in neighboring vessels for the delivery of nutrients, such as oxygen or glucose. Furthermore, oxygen is supplied though hemoglobin within red blood cells. When the hemoglobin molecule is bound to oxygen, it is diamagnetic. In contrast, deoxygenated hemoglobin is paramagnetic. While the changes in the total amount of deoxygenated hemoglobin distort the surrounding magnetic field, the nearby protons will experience different field strengths and thus precess at different frequencies, resulting in a more rapid decay of the transverse component of net magnetization, that is, a shorter T2*. Therefore, images that provide information about the relative T2* values of brain tissue are sensitive to the amount of deoxygenated hemoglobin present. This is in fact the blood oxygenation level–dependent (BOLD) fMRI that we use to localize different functions in human brain.

BOLD-fMRI is measured as the change in the total amount of deoxygenated hemoglobin in a voxel over time and this change presumably triggered by neural activity is defined as the BOLD hemodynamic response. When we relate BOLD hemodynamic response to certain aspects of perception and cognition, there is again an implicit assumption: neural activity and blood flow are tightly coupled both in time and in space via energy demand and oxidative metabolism. However, this coupling between quantitative neurophysiology and changes in BOLD signal is a much more complex one. It is getting even more complicated when the coupling varies across individual brains and different experimental tasks. The complex relationships between neural activity, BOLD signal and behavior are vividly summarized in Figure 3.1, which suggests multiple neural activities (pink boxes) sitting between behavior (blue boxes) and BOLD-fMRI signals (orange boxes). Nevertheless, a complete description of all the relevant issues is beyond the scope of this chapter. We have to keep in mind the limitations of fMRI methodology while interpreting the experimental results. Although fMRI only conveys limited information, we can compare BOLD-fMRI with animal neurophysiological work (Heeger et al., 2000; Logothetis, 2008) or simultaneously measure invasive electrode recordings and BOLD-fMRI in animals (Boorman et al., 2010; Logothetis et al., 2001; Magri et al., 2012; Maier et al., 2008; Sirotin and Das, 2009) and in patients with brain lesions (Mukamel et al., 2005), so as to interpret the BOLD response with more confidence in terms of neural activity.

Most of the fMRI experiments in neuroscience result in a sparse pattern of activation reflecting regions strongly correlated with task-specific cognitive processes. At the same time, other evidence suggesting that the sparseness of activations in fMRI statistical maps can result from elevated noise levels or overly strict predictive BOLD response models. Gonzalez-Castillo and colleague used simple flickering checkerboard, letter, and number discrimination tasks with very low-noise fMRI time-series generated by combining unconventionally large amounts of data (i.e., 100 runs per subject) to challenge the localization view of brain function. They showed that fMRI activations extend well beyond areas of primary relationship to the task. BOLD signal changes correlated with task-timing appear practically everywhere (i.e., 96%) in the brain with each region having a uniquely identifiable time course (Gonzalez-Castillo et al., 2012). These results suggest that although a lot of evidence shows brain regions functionally labeled with behavioral task demand, the results from task-based fMRI research should be interpreted with caution. Finally, the increased use of computational modeling in describing cognitive processes might help us seek ways for better inferring the causal relationships between BOLD signals and the information being processed in the brain. The studies in this thesis highlighted the integration between computational modeling and the analysis of neuroimaging data.

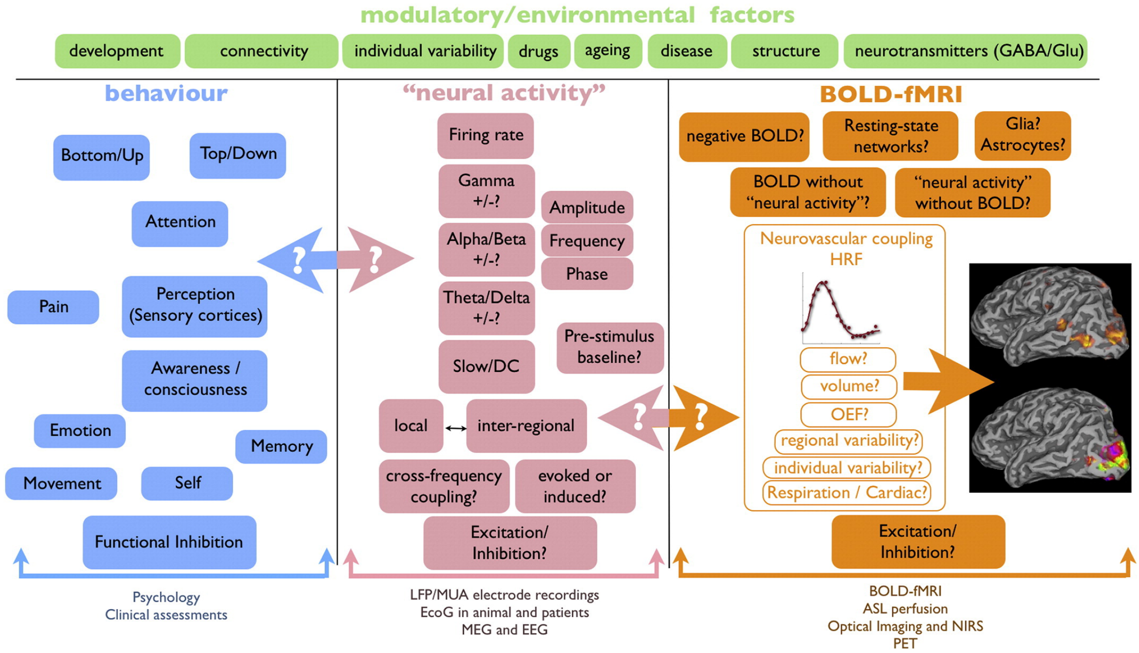

Figure 3.1 An illustration of the complex situation in which behavior and BOLD signals are linked. "It's a long way from behavior to BOLD" taken from (Singh, 2012)

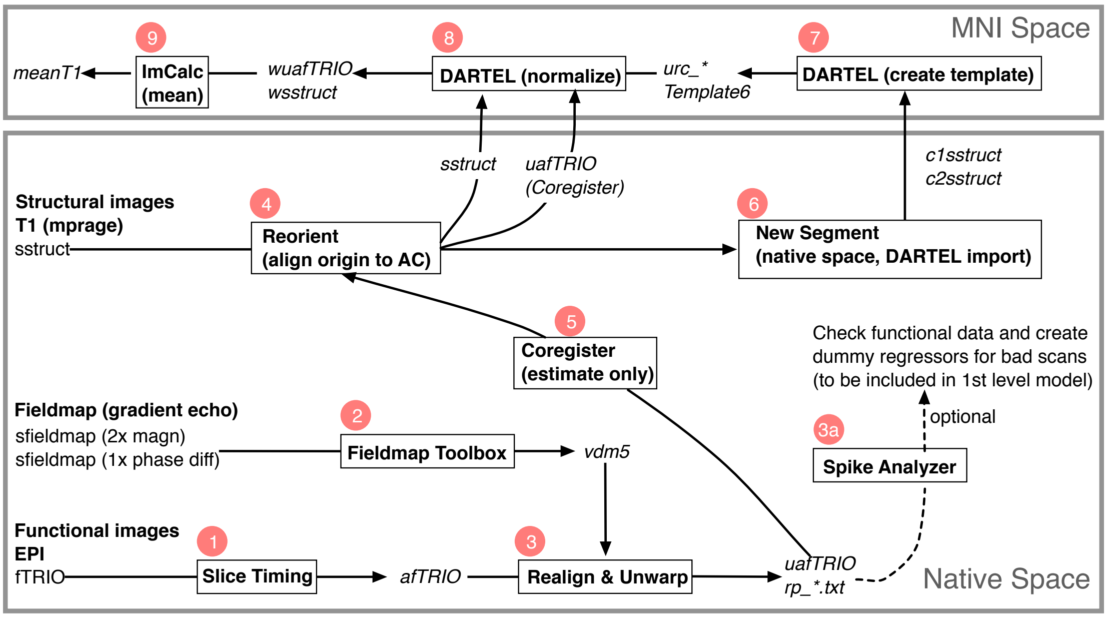

Figure 3.2 An fMRI preprocessing pipeline used in this thesis. The preprocessing is based on algorithms implemented by toolboxes in SPM8. Numbers indicate the computational procedures used in a sequence. Italic texts indicate the name of the fMRI data being processed in each step. Step 3a is optimal for motion correction. Designed by Dr. Jan Gläscher.

3.2 fMRI preprocessing and basic analysis principles

3.2.1 Preprocessing of fMRI data

Preprocessing is known as a series of computational procedures that operate on fMRI data following the data collection but prior to the statistical analysis. The purpose of preprocessing is to remove uninteresting variability from the data and to improve the signal to noise ratio, for example, variance from subjects' movement and scanner artifacts. All the fMRI data analysis in this thesis is performed with SPM (Wellcome Department of Imaging, Neuroscience, Institute of Neurology, London, UK). First, EPI images are spatially and temporally corrected within each subject. Such corrections are important to ensure that each voxel contains data from a single brain region as sampled at regular time intervals throughout the whole experimental run. Second, human brains have wide variation in shape, orientation, and gyral anatomy. To compensate for these differences and make the activation comparable between individuals, images of each brain have to be mapped onto a common space.

The preprocessing pipeline used in this thesis is outlined in Figure 3.4 and explained in details as following:

1. *Slice timing correction*. The fMRI data in this thesis are acquired using pulse sequences as mentioned in the previous section. The exact EPI protocol is: 40 slices are acquired within a TR of 2.26s in a descending order. The correction is necessary to make sure that data on each slice correspond to the same point across time. Imagine a hemodynamic response that happens continuously on two adjacent slices. Because the adjacent slices are actually collected at different times within the TR, without slice timing correction, the hemodynamic response time courses of each slice would differ, even though the underlying activity is identical.

2. *Motion correction and unwarp with field map*. Subjects' head movement is corrected by realigning all the functional images to one specific image through a least squares approach and a 6 parameter spatial transformation, so that all the functional images are in the same orientation and position. A successful realignment ensures that the source of the signal in one voxel

originates from the same location within each scan. However, some movement artifacts still remain after realignment (Poldrack et al., 2011).

Particularly, EPI images often exhibit severe geometric distortions in regions where there is an air-tissue interface, for instance, the orbitofrontal cortex and the anterior medial temporal lobes. In this case, the field map can be used to unwarp the image distortion. The field map is created by acquiring two images of the signal phase with slightly different echo times, i.e., short TE 5ms and long TE 7.47 ms in this thesis. If the magnetic field is completely uniform, the phase difference induced by different echo times will be the same in all voxels otherwise some voxels will be displaced. In line with this idea, a voxel displacement map (VDM) is created and used with unwarp for a combined static and dynamic distortion correction (Andersson et al., 2001; Hutton et al., 2002; Jezzard and Balaban, 1995).

3. *Functional-structural coregistration*. EPI images are typically of rather low resolution and thus have little anatomical contrast. To overcome this limitation, EPI images have to be mapped onto high-resolution structural image via coregistration algorithms within each subject.

4. *Normalization and smoothing*. Spatial normalization is achieved through ‘Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra’ (Dartel) and New Segmentation algorithms. The segmentaion is a process of partitioning the image into such constituents of brain tissues as grey matter, white matter, cerebrospinal fluid, skull, and soft tissue. The basic idea of Dartel is to create an *averaged brain*. Mathematically, it is a high dimensional warping process that increases the registration between individuals, which results in improved localization and increased sensitivity in statistical analyses. This involves taking the parameters of white matter and grey matter produced by the segmentation to create an individual flow field, such that a template can be created in as close alignment as

possible with the tissue probability maps. A Dartel group template is created and this template is normalized to the standard MNI space. Finally, a Gaussian kernel of 8mm full-width at half-maximum is used to smooth the EPI images (Ashburner, 2007).

3.2.2 Basic statistical analysis

3.2.2.1 The general linear model and the analysis of variance

After preprocessing, a general linear model (GLM) for multiple regression is used to identify brain regions that show significant signal change in response to the experimental manipulation. Based on the timing and duration of events in the experiment, we can predict the hemodynamic response evoked as a linear combination of several regressors plus noise. The GLM can be represented by a set of matrices:

$$y = X\beta + \varepsilon, \tag{3.1}$$

where y is the observed BOLD signal as a two-dimensional matrix consisting of time points by voxels. The time points are spaced with TR. X is the design matrix, which consists of multiple regressors, each the same time points in length as the BOLD data. The design matrix is the core of fMRI analysis, because it represents experimental hypotheses. The general linear model attempts to find the parameters β, that is, a regressors-by-voxels matrix for a specific design matrix X. β best accounts for the BOLD data by minimizing the unexplained error ε. Least-squares error is commonly used as a cost function in solving the GLM.

Regressors in the design matrix are typically built by convolving a stick function with an ideal noiseless hemodynamic response function (HRF). For instance, a canonical double-gamma HRF implemented in SPM has a delay of response for 6 seconds and a delay of undershoot for 16 seconds. By adopting an event-related design, the stimulus-onset time series consist of stick functions of equal height. The convolution is guided by linear time invariant properties between the neural response and BOLD signal. Notably, the only free parameter of the HRF actually estimated in the general linear model is the height of the response function. In addition, parametric regressors can also be created to model the strength of BOLD response with respect to the parametrically varied stimuli. In this case, each stimulus stick function has a height reflecting the modulation value in a specific trial.

As mentioned in the previous section, head motion during the scan can cause artifacts in the data and such residual motion artifacts can remain even after motion correction. Thus, the six time courses of the translation and rotation parameters can be included as nuisance regressors in the GLM model to account for residual motion-related variance. Altogether, we have three types of regressors in the GLM design matrix: unmodulated onset regressors, modulated parametric regressors and nuisance regressors. Importantly, SPM subtracts the mean value from each regressor so that the variance associated with the mean signal intensity is not assigned to any experimental condition. Ideally all the regressors in the model should be independent of, in other words, orthogonal to, the other regressors so as to improve the chances of identifying meaningful statistical effects. However, in terms of different experimental hypotheses, orthogonalization should be considered as a method for clarifying the unique effects attributable to a specific regressor.

After this single-subject level analysis, the GLM parameters are raised to a succeeding group level of analysis. Combining data from multiple subjects as random-effects increases the experimental power. Furthermore, it is important to go beyond the identification of significant activations to understand differences between groups of subjects. For example, group comparison is especially useful for better understanding brain dysfunctions in clinical studies.

3.2.2.2 Corrections for multiple comparisons

Statistic results are generally evaluated by applying a threshold and observing the spatial distribution of statistics that survive the threshold. However, there is a massive issue with multiple comparisons in practice. Consider testing 100,000 voxels at a threshold of p<0.05. This means that on average 5000 will be significant by chance, which is termed as *false positives*, or *type I errors*. Thus, the statistic results always need to be somehow corrected for multiple comparisons. A straightforward correction method would be Bonferroni correction:

$$p_{corrected} = \frac{p_{uncorrected}}{N}, \tag{3.2}$$

where N is the number of voxels in the whole brain. This method is appropriate if the N voxels are independent, whereas adjacent voxels in fMRI data are correlated due to inherent limitations in data collection and preprocessing. As a result, Bonferroni correction might become overly conservative. The family-wise error correction accounts for this problem using Gaussian random field theory and calculates the effective correction factor, which is less conservative. However, such a correction method requires smoothing the data.

One alternative to the analysis of individual voxels is to use a cluster-wise threshold, where clusters of a large number of voxels are counted as surviving statistical scrutiny. The likelihood of a false-positive result decreases with increasing cluster size. This cluster-size threshold can also be estimated through Gaussian random field theory based on the spatial correction and smoothness of the data. The thesis work related to corrections for multiple comparison is implemented through 3dClustSim in AFNI (Cox, 1996).

3.3 Combining reinforcement-learning theory with fMRI data

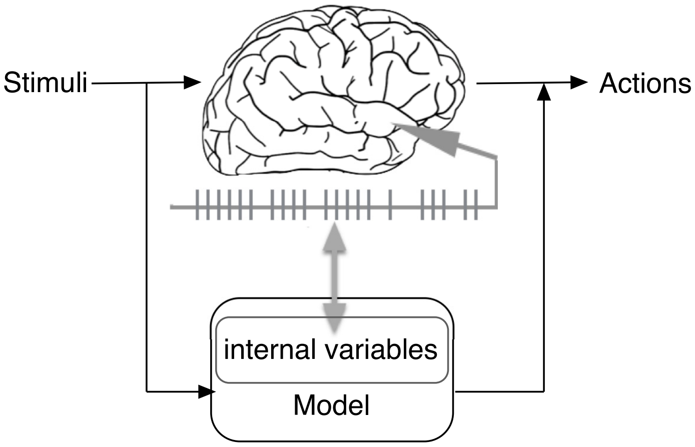

Figure 3.3 Illustration of model-based analysis of decision variables. The brain experiences sensory stimuli and generates decisions and motor outputs. To search for the neural correlates of the brain's internal decision dynamics, we assume a computational learning model (e.g., a reinforcement learning model), estimate the model's internal variables and parameters, and seek for any correlation with them in the neural signal (e.g., fMRI data).

A number of general computational frameworks have been developed for describing cognitive processes. These frameworks are used as bridges between hypotheses of mental operations and the corresponding neural activities. The learning model-based analysis of neural signal is illustrated in Figure 3.3. We assume the brain's dynamic cognitive process can be captured as internal decision variables and parameters when the subjects experience some sensory stimuli and perform motor actions. Similarly, we define a computational model that takes the same input and generate outputs that are similar to the real actions. To seek for neural correlates of the brain's internal decision variables, we can estimate the variables and parameters of the model and look for correlations with them in the measured neural signal.

This thesis applies reinforcement-learning models to account for human choice behavior during decision-making tasks. Internal variables derived from these learning models are integrated into the analysis of fMRI data. This method is termed as *model-based fMRI analysis* (Doya et al., 2011; Gläscher and O'Doherty, 2010; O'Doherty et al., 2003), which provides us a tool for identifying how a particular brain circuitry might carry out certain neural computations. In general, this computational approach is not limited to reinforcement-learning models. Instead, various other computational models can be used as long as the models can explain cognitive operations on a trial-by-trial basis or rather in a continuous time domain.

The model-based fMRI analysis consists of three steps:

1. Defining candidate computational models according to specific experimental hypotheses. The computational framework as mentioned in Chapter 2 represents experimental-driven quantitative hypothesis about how the brain might approach a specific decision-making problem. In particular, the trial-by-trial value update scheme of reinforcement-learning models provides a dynamic learning process, which enables us to search for neural correlates of the value estimates during learning. Therefore, competing hypotheses are generally formed in terms of model selection.

2. Fitting each model to the empirical behavioral data and select the best fitting model. The free parameters of each model are

determined by fitting models to the behavioral data. This model-fitting procedure confers that the model used in the fMRI analysis is behaviorally relevant and psychologically valid. Details about the model fitting in terms of parameter estimation and model comparison are summarized in the next section. In fact, reliable parameters not only reflect the underlying psychological traits but also affect the subsequent fMRI analysis. However, due to the limited information from behavioral data, this method might not be sufficient after all. Another potential but computationally much more complicated method is to combine both behavioral data and BOLD signal in a full Bayesian framework for the model fitting and comparison.

3. Deriving the time series of a model's internal variables and using them as parametric modulators in the GLM of fMRI data analysis. After having fitted a model to the behavioral data, regressors can be derived on a trial-by-trial basis. The model is taken as a puppet experimental participant, which generates exactly the same choice behavior as a real human subject. At the same time, the model gives out a record of computational values as if the real subject has used these values for generating decisions on each trial, e.g., expected value, reward prediction error etc. There are two ways to put these values into the GLM design matrix. One is to use the time series at each trial as a parametric modulator for the main event onset regressor. The other way is to convolve the time series with a standard hemodynamic function and build a main regressor by hand.

3.4 Parameter estimation and model comparison

Computational models typically have a number of free parameters, measuring the experimental manipulations to be estimated from the experimental data. Different computational models constitute different experimental hypotheses about the cognitive process that give rise to the data. These hypotheses may be tested against one another on the basis of their fit to the data. In this section, we investigate methods for fitting models into the trial-by-trial behavioral data and consider issues regarding the comparison of different models.

Models are mathematical formulas that have parameters and we sometimes have prior beliefs about these parameters. After observing empirical data, estimation about those parameters can be updated with posterior beliefs. Given the empirical data D, to seek a set of parameters θ in the context of a particular model M can be formalized by a Bayesian rule (Gelman et al., 2003):

$$P(\theta|D,M)=\frac{P(D|\theta,M)P(\theta|M)}{P(D|M)}. \tag{3.3}$$

Parameter estimation means determining the posterior probability over the parameters, i.e., $P(\theta|D,M)$. We can rewrite Equation (3.3) without explicitly annotating the model parameters:

$$P(M|D)=\frac{P(D|M)P(M)}{P(D)}. \tag{3.4}$$

Based on the posterior probability of a model given data, i.e., $P(M|D)$, we want to determine the relative fit of one model over another. For instance, we can suppose two models M_1 and M_2. By taking a ratio of their respective evidence according to Equation (3.4), we have

$$\frac{P(M_1|D)}{P(M_2|D)}=\frac{P(D|M_1\ )P(M_1\ )}{P(D|M_2\ )P(M_2\ )}. \tag{3.5}$$

Here, the marginal likelihood $P(D|M)$ is the key *model evidence*, which does not make reference to any particular model parameters. In fact, the model parameters must be averaged out by:

$$P(D|M)=\int d\theta P(D|\theta,M)P(\theta|M). \tag{3.6}$$

The ratio of the evidence $P(D|M_1\ )/P(D|M_2\ )$ is called the Bayesian factor, which is a standardized measure for comparing the relative fit of each model. Notably, Bayesian model comparison automatically takes the model complexity into account. A more complex model is punished for being vulnerable to over fitting. Nevertheless, it is important to bear in mind that a model comparison process can merely tell us which model is less bad than the other. Sometimes, it might be helpful to use the knowledge of a posterior distribution to simulate further data.

An additional comparison between simulated and the original data might facilitate intuitive diagnostics of the model.

The Bayesian inference described above looks simple and neat, but in reality, it can be very difficult. The computation of the integral in Equation (3.6) can be intractable even thought the parameter space is only moderately large. In the following section, we summarize some practical methods for approximating the calculation of this Bayesian inference. In particular, the maximum likelihood estimation and hierarchical Bayesian modeling.

3.4.1 Maximum likelihood estimation

Equation (3.3) says that the posterior probability of parameters is proportional to the product of the likelihood of data given the model parameters and the prior probability of the parameters. A widely used estimating method is *maximum likelihood estimation (ML)*. ML estimation seeks a single point estimate rather than a full posterior distribution of the parameters over all possible values. It uses a uniform prior of $P(\theta|M)$ and estimate the parameter θ by maximizing a likelihood function of $P(D|\theta, M)$.

First, the likelihood function $\mathcal{L}(\theta)$ of a reinforcement-learning model, given the empirical data denoted by $e_t \coloneqq \{s_1, a_1, r_1, \dots s_t, a_t, r_t\}$ with states s_t, action a_t and rewards r_t, can be written as:

$$\mathcal{L}(\theta) = \prod_t P(a_t|s_t, e_{t-1}, \theta). \tag{3.7}$$

Next, the estimation for ML parameter θ_{ML} can be reduced into the following nonlinear optimization problem:

$$\theta_{ML} = \underset{\theta}{argmax} \mathcal{L}(\theta). \tag{3.8}$$

In practice, Equation (3.8) is computed after a transformation by a logarithm function to reduce the computational burden. The data log-likelihood $log\ (\mathcal{L}(\theta_{ML}))$ or BIC-corrected likelihood is often reported as indicators of the goodness of fit. The BIC score (Schwarz, 1978) is an approximation of Equation (3.6) defined as:

$$log\big(P(D|M)\big) \approx log\big(P(D|M,\theta_{ML})\big) - \frac{n}{2} log\ (m), \tag{3.9}$$

where n is the number of parameters in the model and m is the number of data points, for instance the number of trials in a MDP task. BIC score punishes the model complexity with the term $\frac{n}{2} log\ (m)$ for preventing over fitting.

Another typical way of model comparison in statistics is *likelihood ratio test* (Neyman and Pearson, 1933), which is especially useful for comparing nested models. Suppose that a complex model M_1 nests a simpler model M_2, the test statistic can be calculated as:

$$d = 2\big[log(P(D|M_1,\theta_{ML}^{M_1}) - log\ (P(D|M_2,\theta_{ML}^{M_2}))\big]. \tag{3.10}$$

The probability distribution of d follows approximately a chi-square distribution (Huelsenbeck and Crandall, 1997; Wilks, 1938) with a degree of freedom equals the difference of the number of parameters in each model. Therefore, a p-value can be computed to test if model M_1 fits the data better than model M_2 simply by chance. Moreover, if we assume a random agent as an empty model that has zero parameter, a likelihood ratio test can be used to check whether any model performs better than the purely random null model.

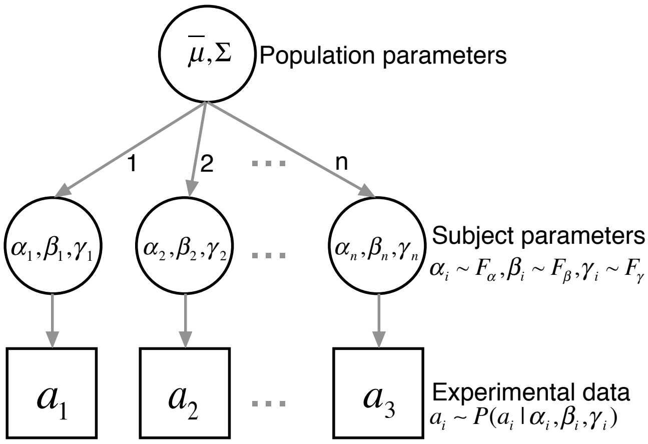
$\bar{\mu},\Sigma$
Population parameters
1
2
...
n
$\alpha_1,\beta_1,\gamma_1$
$\alpha_2,\beta_2,\gamma_2$
...
$\alpha_n,\beta_n,\gamma_n$
Subject parameters
$\alpha_i \sim F_\alpha, \beta_i \sim F_\beta, \gamma_i \sim F_\gamma$
a_1
a_2
...
a_3
Experimental data
$a_i \sim P(a_i | \alpha_i, \beta_i, \gamma_i)$

Figure 3.4 Illustration of hierarchical Bayesian analysis for the parameter estimation. There is a group of n subjects and each subject has a set of parameters $\{\alpha_i, \beta_i, \gamma_i\}$. The subject parameters are drawn from prior distributions with certain mean $\bar{\mu}$ and variance $\sum$, which are the population parameters that characterize the underlying psychological traits of this group of subjects.

3.4.2 Hierarchical Bayesian modeling

So far we have focused on estimating model parameters with experimental data from a single subject. Many experimental studies involve groups of subjects and the research of interest often lies in the comparison between groups. The question is how to estimate the population parameters from a set of experimental data. A straightforward method is to firstly estimate the ML parameters for each subject and then to simply summarize the mean and variance of all the parameters from the same group of subjects. However, this method ignores the inherent noise in the single-subject parameter estimation, which may badly inflates the variance of the population estimation (Daw, 2011).

We are interested in situations where parameters vary across populations as shown in Figure 3.4. Suppose we have N groups of subjects and n subjects in each group. We denote a set of parameters for subject i as φ_i, for instance $\varphi_i = \{\alpha_i, \beta_i, \gamma_i\}, i \in \{1, \dots, n\}$. We assume each of the parameters in φ_i is drawn from a certain prior probability distribution $\mathcal{F}$, i.e. $\varphi_i = \{\alpha_i \sim \mathcal{F}_\alpha, \beta_i \sim \mathcal{F}_\beta, \gamma_i \sim \mathcal{F}_\gamma\}$. $\mathcal{F}$ is conventionally a Gaussian, Gamma or Beta distribution. The shape parameters charactering such prior distributions are the parameters that we would like to compare between different populations. We can term these shape parameters as *population parameters*, such as the mean μ and variance σ. We denote a set of population parameters from the jth population by ϑ_j, where $j \in \{1, \dots, N\}$. We can again use the Bayesian rule to recover the population parameters in terms of the model and data:

$$P\left(\vartheta_j \middle| D, M\right) = \frac{P(D|\vartheta_j, M)P(\vartheta_j|M)}{P(D|M)}. \tag{3.11}$$

Then the likelihood $P\left(D \middle| \vartheta_j, M\right)$ is averaged over all possible sets of individual subject's parameters according to their prior probability distribution:

$$P\left(D \middle| \vartheta_j, M\right) = \prod_{i=1}^{n} \int d\, \alpha_i d\beta_i d\gamma_i \mathcal{F}_\alpha \mathcal{F}_\beta \mathcal{F}_\gamma P(D|\alpha_i, \beta_i, \gamma_i, M). \tag{3.12}$$

If we assume $\mathcal{F}$ as a Gaussian distribution for each prior probability, that is $\vartheta_j = (\bar{\mu}, \Sigma)$, $\bar{\mu} \in \mathbb{R}^k$, $\Sigma \in \mathbb{R}^{k \times k}$, where k is the number of parameters in the model for each individual subject. If we take $\varphi_i = \{\alpha_i, \beta_i, \gamma_i\}$ as an example, then we have $\bar{\mu} = \{\mu_\alpha, \mu_\beta, \mu_\gamma\}, \Sigma = diag(\sigma_\alpha, \sigma_\beta, \sigma_\gamma)$, and $k = 3$. The prior distribution for φ_i is thus:

$$P(\varphi_i|\vartheta_j) = \frac{1}{2\pi^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}} e^{(-\frac{1}{2}(\varphi_i-\bar{\mu})\Sigma^{-1}(\varphi_i-\bar{\mu}))}. \tag{3.13}$$

Then ϑ_j can be computed as A Maximum Posterior estimation by Expectation Maximization algorithm as following:

$$\varphi_i = \underset{\varphi_i}{argmax} P(D|\varphi_i)P(\varphi_i|\vartheta_j), \tag{3.14}$$

$$\mu = \frac{1}{n}\sum_i \varphi_i, \tag{3.15}$$

$$\Sigma = \frac{1}{n-1}\sum_i (\varphi_i - \mu)(\varphi_i - \mu)^T. \tag{3.16}$$

The steps of Equation (3.15)(3.16) are repeated until the prior parameters $\vartheta_j = (\mu, \Sigma)$ converge. However, the Expectation Maximization algorithm in this case only supports Gaussian priors and the nonlinear optimization requires calculation of analytical derivatives accordingly. The computation gets more difficult, since reinforcement-learning models in general involve a lot of nonlinear transformations.

Another technique is to assess the properties of a posterior distribution in terms of directly sampling it. Samples of representative random values can be generated by means of Monte Carlo simulation. For example, the Metropolis algorithm and Gibbs sampling are specific types of a Markov chain Monte Carlo (MCMC) process. A hierarchical modeling approach treats the single-subject parameters as being generated from a group-wise distribution. In particular, a prior belief about the group-wise distribution can constrain the estimation for parameters of a particular individual subject. As a result, single-subject parameters are pulled toward the group mean, which leads to more robust estimates. Thus, hierarchical models simultaneously account for both differences and similarities among subjects within a same population (Ahn et al., 2011; Lee, 2011; Nilsson et al., 2011; Perfors et al., 2011; Shiffrin et al., 2008).

Generally, Bayesian MCMC fitting software, such as winBUGs/openBUGS software (Lunn et al., 2012), Stan (Team, 2014), and Jags (Plummer, 2003), are employed in practice for Bayesian analysis of statistical models. In this thesis, I used the JAGS software and its runjags R interface for the computations of model estimation. Full details about the underlying MCMC algorithms are beyond the scope of this chapter (Kruschke, 2010). Briefly, taking a random walk through the parameter space generates sample values of a target probability distribution. Summary statistics are inferred from these random samples. In the meanwhile, the MCMC simulation needs to be diagnosed for whether the samples have reliably and effectively covered the target probability distribution (Cowles and Carlin, 1996). Firstly, multiple chains of MCMC simulations have to be executed in parallel with a certain burn in period (i.e. get rid of initial samples) and specific thinning interval (i.e. take only the every 3rd or 5th samples). Secondly, there are different methods to check whether an MCMC chain has converged. One method is to see how well the chains are mixing and moving around the parameter space. Trace and the density of samples can be plotted for visual inspection. Another way to assess convergence is to examine the autocorrelations between the neighboring samples of a Markov chain. The autocorrelation are expected to decrease as the number of lags increase. Lastly, the variance of samples within a single chain or between multiple chains can be compared with Gelman-Rubin test to statistically diagnose the convergence (Brooks and Gelman, 1997).

The deviance information criteria (DIC) can be calculated for quantitatively comparing the model fit from the Bayesian MCMC analysis (Gelman et al., 2003). DIC is the hierarchical modeling generalization of the BIC criterion, which is an approximation of the model evidence of Equation (3.6). The DIC is calculated as:

$$DIC = D(\hat{\theta}) + 2p_D, \tag{3.17}$$

where $\hat{\theta}$ is the average of model parameters, $D(\hat{\theta})$ is proportional to a log likelihood function of the data, and p_D is the effective number of parameters. $D(\hat{\theta})$ measures how well the model fits the data, while p_D is a penalty on the model complexity. The lower a DIC score, the better a model fit. As a rule of thumb, value difference of DICs between 5 to 10 is considered substantial (Spiegelhalter et al., 2002).

3.5 Thesis work relating to fMRI experiment

The experimental paradigms and behavioral findings presented in Chapter 4 are the basis for fMRI studies presented in Chapter 5 and 6. The modeling analysis in Section 4.1 and 4.3 were conducted through maximum likelihood estimation. The analysis in Section 4.2 was conducted through Bayesian hierarchical modeling.

Chapter 5 presents contextual modulation of both stimulus and reward prediction errors in the BOLD signals from the ventral striatum during the behavioral paradigm introduced in Section 4.2. Chapter 6 investigates the neural structures that support counterfactual learning during the behavioral paradigm introduced in Section 4.3.

Chapter 4: ESTIMATING PREDICTION ERRORS IN PARADIGMS OF MULTIPLE VALUATION SYSTEMS

The processes of successful decision-making depend on the precise monitoring of action-outcome associations and the flexible integration of different aspects of reward information. A wide range of factors can modulate these complex processes, such as temporal properties of the rewards, the novelty or emotional significance of choices, and the amount of risk and variability associated with the outcomes. Furthermore, the ability to predict reward at different time scales allows us to pursue larger future rewards and longer-term goals instead of focusing on immediately available but less rewarding outcomes.

In this chapter, I introduce three new paradigms of reward-based learning appropriate for investigating multiple valuation systems in human decision-making. Each paradigm is designed and analyzed based on the hypothesis formulated by reinforcement-learning models. These hypotheses are tested with subject's choice behavior. The model-based behavioral findings in this chapter suggest that error-correction via reinforcement is commonly engaged in multiple valuation systems.

4.1 The encoding of higher-order reward prediction errors

4.1.1 Introduction

Humans can identify patterns in temporal sequences in order to predict future event. For instance, we observe whether on a particular day it rains or not, then we wish to predict whether it will rain or not on the next day. If we treat such observations as an independent and identical distribution, the only information we can glean from the data is the relative frequency of rainy days. However, we know in practice that the weather often exhibits trends that may last for several days. Observing whether or not it rains today is therefore of significant help in predicting if it will rain tomorrow (Bishop, 2006). For many of such sequential observations, we anticipate that the statistical relationships among several successive observations provide important information in predicting the next value.

This kind of sequential learning requires not only tracking the frequency of encountered stimuli, but also the higher-order statistical relationships among them. Standard reinforcement learning models learn the mean reward frequency and bias action selections towards the most rewarding action. However, if average rewards are equal for all the available options, but the stimuli occur in a specific temporal pattern, how does an observer learn such temporal dependencies? Specifically, can we use the prediction errors to learn the conditional correlations among sequential stimuli? To address this question, we designed a two-armed bandit task where a visual stimulus alternates between the left and right sides of the screen according to a specific Markov chain.

4.1.2 Methods

4.1.2.1 Experimental paradigm

We designed a two-armed bandit task where subjects had to predict whether a stimulus (i.e., money) would appear on the left or the right side of the screen. Subjects had to correctly predict the forthcoming stimulus location indicated by two treasure chests to gain a monetary reward (Figure 4.3). The stimulus alternated between the left and right locations according to specific probabilities. We aimed to study how subjects learned the sequential aspects of the stimuli, especially the correlations between two or three stimuli that were close in a sequence. We anticipated that subjects used the information about the temporal correlations of the stimuli to predict the forthcoming stimulus. One way to allow earlier observations to have an influence is to move to higher-order Markov chains.

We define three types of probabilities: (1) *zero-order probabilities*, the frequency of a stimulus appearing on either the left (L) or the right (R) location, e.g., $p(L)$. (2) *first-order* probabilities, the conditional probability of a stimulus appearing on one location given its previous one, e.g., $p(L|L)$. If we allow the current stimulus to depend only on its previous one, we obtain a first-order Markov chain. (3) *second-order probabilities*, the conditional probability of a stimulus appearing on one location given its previous two stimuli, e.g., $p(L|L,L)$. If we allow the current stimulus to depend only on its previous two stimuli, we obtain a second-order Markov chain.

We set up two experiments by keeping the lower-order probabilities fixed at 0.5 but systematically changing the higher-order conditional probabilities using Markov properties. There are four first-order conditional probabilities of the neighboring stimuli: $p(L|L)$, $p(L|R)$, $p(R|L)$ and $p(R|R)$. If we fix the zero-order probability at 0.5, we can reduce the four first-order probabilities to one degree of freedom. Similarly, there are eight second-order conditional probabilities $p(L|L,L)$, $p(L|L,R)$, $p(L|R,R)$, $p(L|R,L)$, $p(R|L,L)$, $p(R|L,R)$, $p(R|R,L)$ and $p(R|R,R)$. If we keep both the zero-order and first-order probabilities fixed at 0.5, we can reduce the eight second-order probabilities to one degree of freedom as well. As described in detail in the following subsections, we set $p(L|L)$ as the only free parameter in the experiment of first-order Markov task and $p(L|L,L)$ as the only free parameter in the experiment of second-order Markov task.

4.1.2.1.1 Experiment 1: first-order Markov task

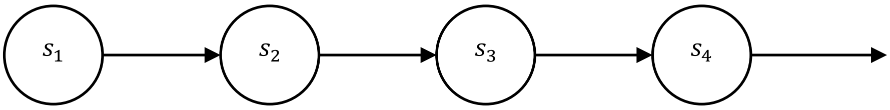

Figure 4.1 A first-order Markov chain of stimuli. The distribution $p(s_t|s_{t-1})$ of a particular observation s_t is conditioned on its previous observation s_{t-1}.

We denote the stimulus location as $s_t \in \{L,R\}$ on each trial t. In the first-order Markov task, we generated the sequence of stimuli by controlling the first-order probability $p(s_t|s_{t-1})$ as shown in Figure 4.1. Thus, a whole sequence of N stimuli has the property:

$$p(s_1, \dots s_N) = p(s_1) \prod_{t=2}^{N} p(s_t|s_{t-1}). \tag{4.1}$$

We set the probabilities of the stimulus appearing on either the left or the right location to be the same:

$$p(L) = p(R) = 0.5. \tag{4.2}$$

Then, we define a first-order parameter a and $p(L|L) = a$. We have:

$$\begin{bmatrix} p(L|L) & p(L|R) \\ p(R|L) & p(R|R) \end{bmatrix} = \begin{bmatrix} a & 1-a \\ 1-a & a \end{bmatrix}. \tag{4.3}$$

We set up seven experimental conditions. Each condition had a different conditional probability of $a \in \{0.125, 0.375, 0.5, 0.625, 0.75, 0.875\}$ and was run with a block of 150 trials. There zero-order probabilities were always 0.5. Therefore, nothing could be learned by the subjects on the basis of the frequency information alone. However, the subjects might improve their performance if they learned the transition probabilities, e.g., the fact that the stimulus was more likely to alternate or to repeat.

4.1.2.1.2 Experiment 2: second-order Markov task

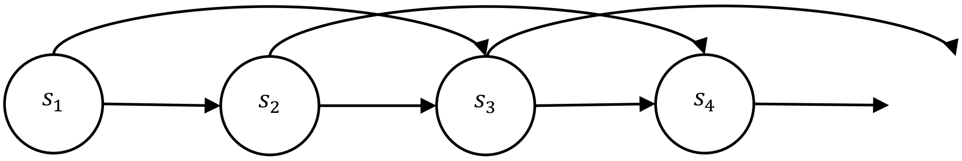

Figure 4.2 A second-order Markov chain of stimuli. The distribution $p(s_t|s_{t-1}, s_{t-2})$ of a particular observation s_t depends on the values of its previous two observations s_{t-1} and s_{t-2}.

In the second-order Markov experiment, we generated the sequence of stimuli by controlling the second-order probability $p(s_t|s_{t-1}, s_{t-2})$. Thus, a whole sequence of N stimuli has the property:

$$p(s_1, \dots s_N) = p(s_1)p(s_1|s_2)\prod_{t=2}^{N} p(s_t|s_{t-1}, s_{t-2}). \tag{4.4}$$

We set the zero-order and first-order conditional probabilities fixed at 0.5:

$$p(L) = p(R) = 0.5, \tag{4.5}$$

$$p(L|L) = p(L|R) = p(R|L) = p(R|R) = 0.5. \tag{4.6}$$

We define a second-order parameter b and $p(L|L, L) = b$. We have:

$$\begin{bmatrix} p(L|L,L) & p(R|L,L) \\ p(L|L,R) & p(R|L,R) \\ p(L|R,L) & p(R|R,L) \\ p(L|R,R) & p(R|R,R) \end{bmatrix} = \begin{bmatrix} b & 1-b \\ 1-b & b \\ b & 1-b \\ 1-b & b \end{bmatrix}. \tag{4.7}$$

We set up seven experimental conditions. Each condition had a different conditional probability of $b \in \{0.125, 0.375, 0.5, 0.625, 0.75, 0.875\}$ and was run with a block of 150 trials. Both the zero-order and the first-order probabilities were not informative in this experiment. Only if the subjects had learned the second-order probabilities, would they make more correct predictions.

Note that we can mathematically prove Equation (4.3) holds true given Equation (4.2), so does Equation (4.7) given Equation (4.5) and (4.6) (Bishop, 2006).

4.1.2.2 Experimental task and procedure

The ordering of the blocks was fully counterbalanced across subjects. Every participant completed 7 blocks of 150 trials each, with self-paced breaks in between blocks. On each trial, there were two treasure chests, on the left and the right of the screen. The subjects were instructed to predict which treasure chest contains money by making a button press on the response trigger pad with the right index or middle finger. Then both treasure chests were open. If there were money in the chosen treasure chest, the subject won 1 point. If there was no money in the chosen treasure chest, the subject received no reward (Figure 4.3). Subjects were instructed to maximize the cumulative number of points scored during the experiment. They were paid according to the total number of treasures they collected during the experiment. Each trial started with a 2s inter trial interval when a fixation cross was presented at the center of the screen. The two treasure chests were then displayed and the subject had to make a choice. If no choice was made within 2s, a message "Too slow!" was displayed for a time-out of 4s and that particular trial was abandoned. The chosen treasure chest was indicated with a key inserted in the treasure chest, after which a stack of coins or an empty treasure chest was shown for 1.5s.

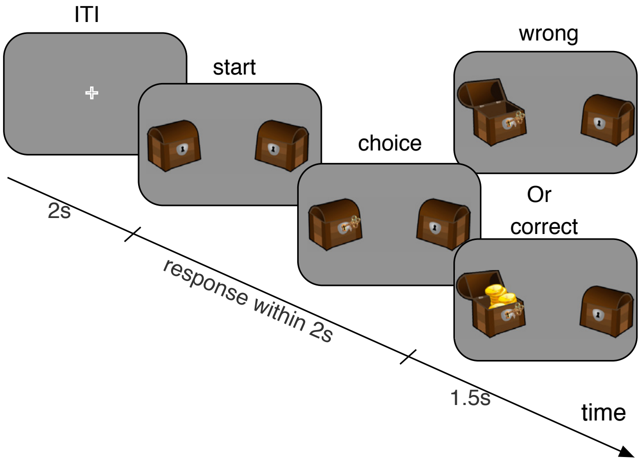

Figure 4.3 Sample trial: subject had to make a choice when the two treasure chests were displayed. Here the left was chosen and indicated by a key inserting in the left treasure chest, after which subjects would either win a stack of coins (indicating a reward of 1 point) or get an empty treasure chest (indicating no reward).

Twenty participants (mean age, 26 years; age range, 20-38 years; 12 male and 8 female) with normal or corrected-to-normal vision were recruited from the student population at Technische Universität Berlin. Each participant was paid a base rate of 7€ for participating in the experiment and a bonus depending on the amount of money they won during the experiment (mean 6.1€ ± SD 0.4€). Ten subjects participated in Experiment 1 and the other ten subjects participated in Experiment 2. This study was conducted in accordance with the principles of the Declaration of Helsinki for subject participation in scientific studies and was approved by the local ethics committee.

4.1.2.3 Computational modeling

We hypothesized that subjects learned the first-order probabilities in Experiment 1 and learned the second-order probabilities in Experiment 2 for making correct predictions. To test whether different prediction errors could be adapted to estimate such higher-order probabilities, we formulated three reinforcement-learning models.

4.1.2.3.1 Zero-order learner

The standard Rescorla-Wagner model estimates the expected value of each option. The expected value is proportional to the frequency of a stimulus appearing on either location, i.e., $p(L)$ and $p(R)$. We define the zero-order learner as a Rescorla-Wagner model that estimates the expected value of taking action $a_t \in \{L, R\}$ on each trial:

$$V^{t+1}(a_t) = V^t(a_t) + \alpha(r_t - V^t(a_t)), \tag{4.8}$$

Then, the action is selected stochastically through a softmax function:

$$P(a_t) = \frac{exp\ (\beta V(a_t))}{exp(\beta V(L)) + exp(\beta V(R))}. \tag{4.9}$$

$r_t \in \{0,1\}$ is the reward that depends on whether subjects made a correct prediction $(r_t = 1, if\ a_t = s_t)$ or not $(r_t = 0, if\ a_t = s_t)$. α is the learning rate and β is the inverse temperature parameter that controls the exploration against exploitation during learning. We use α and β for same notations in the following two models.

4.1.2.3.2 First-order learner

We adapt the Rescorla-Wagner model to directly estimate the first-order conditional probability $p(s_t|s_{t-1})$:

$$p^{t+1}(s_t|s_{t-1}) = p^t(s_t|s_{t-1}) + \alpha(\sigma_t - p^t(s_t|s_{t-1})), \tag{4.10}$$

where $p(s_t|s_{t-1}) = \{\, p(L|L), p(L|R), p(R|L), p(R|R)\}$. $\sigma_t \in \{0,1\}$ is a binary indicator that $\sigma_t = 1$ for the observed pair of successive stimuli (s_t, s_{t-1}) and $\sigma_t = 0$ for all the other three unobserved pairs at trial t. Since the second trial, the model tracks the four conditional probabilities in parallel. Each of these probabilities is initialized to be 0.5. Then, these conditional probabilities are used in the softmax function for action selection.

$$P(a_t = L|s_{t-1}) = \frac{exp\ (\beta p(L|s_{t-1}))}{exp(\beta p(L|s_{t-1})) + exp\ (\beta p(R|s_{t-1}))}. \tag{4.11}$$

$$P(a_t = R|s_{t-1}) = 1 - P(a_t = L|s_{t-1}). \tag{4.12}$$

Similarly as in the zero-order model, we assume a softmax action selection procedure for this first-order model, and for the second-order model described in the next subsection. The softmax function is to make the learning models more flexible in characterizing subjects' learning process with regards to exploration and exploitation trade-off.

4.1.2.3.3 Second-order learner

We adapt the Rescorla-Wagner model to directly estimate the second-order conditional probability $p(s_t|s_{t-1}, s_{t-2})$:

$$p^{t+1}(s_t|s_{t-1}, s_{t-2}) = p^t(s_t|s_{t-1}, s_{t-2}) + \alpha(\sigma_t - p^t(s_t|s_{t-1}, s_{t-2})), \tag{4.13}$$

where $p(s_t|s_{t-1}, s_{t-2}) = \{p(L|L,L)\}$, $p(L|L,R)$, $p(L|R,L)$, $p(L|R,R)$, $p(R|L,L)$, $p(R|L,R), p(R|R,L), p(R|R,R)\}$. $\sigma_t = 1$ for the observed triple of successive stimuli (s_t, s_{t-1}, s_{t-2}), and $\sigma_t = 0$ for all the other unobserved seven triples at trial t. Since the third trial, this model tracks the eight conditional probabilities in parallel. Each of these probabilities is initialized to be 0.5. Then, these conditional probabilities are used in the softmax function for action selection:

$$P(a_t = L|s_{t-1}, s_{t-2}) = \frac{exp\ (\beta p(L|s_{t-1}, s_{t-2}))}{\sum_{A=\{L,R\}} \beta p(A|s_{t-1}, s_{t-2})}. \tag{4.14}$$

$$P(a_t = R|s_{t-1}, s_{t-2}) = 1 - P(a_t = L|s_{t-1}, s_{t-2}). \tag{4.15}$$

4.1.3 Results

In both experiments, subjects' performance reached at least 70 percent correct when the stimulus appeared in a highly predictive manner, that is, when $p(L|L)$ or $p(L|L,L)$ approached 0.125 or 0.875, shown in blue in Figure 4.4 A for Experiment 1 and in Figure 4.4 B for Experiment 2. The ideal prediction for the first-order experiment was to always choose the location opposite to the previous stimulus if $p(L|L) \leq 0.5$, whereas to always choose the same location as the previous stimulus if $p(L|L) > 0.5$, presented as green and cyan solid lines respectively in Figure 4.4 A. This strategy made best use of the first-order conditional probabilities. Therefore, it was optimal for decisions during Experiment 1. However, such a strategy was no longer optimal for Experiment 2, where the first-order conditional probability was not informative at all. Indeed, the first-order learner only performed at chance of 50% correct prediction rate during Experiment 2 (green and cyan solid line in Figure 4.4 B).

The ideal predication for the second-order experiment had to make best use of the second-order conditional probabilities. Therefore, the optimal strategy was to choose the location opposite to the previous-but-one stimulus if $p(L|L, L) \leq 0.5$, whereas to choose the same location as the previous-but-one stimulus if $p(L|L, L) > 0.5$, presented as green and cyan dashed line respectively in Figure 4.4B.

On average, subjects chose both left and right with 50% chance in Experiment 1 and Experiment 2, shown in red in Figure 4.5 A and B. This could be attributed to the fact that $p(L) = p(R) = 0.5$ in both experiments. In Experiment 1, the first-order probability $p(s_t|s_{t-1})$ was systematically changed across seven experimental conditions by setting $p(L|L)$ as the sole variable with the other three first-order conditional probabilities dependent on $p(L|L)$. The probabilities of the subjects' choice given the previous stimulus $p(a_t|s_{t-1})$ strongly correlate with $p(s_t|s_{t-1})$ (correlation coefficient = 0.9871, p=3.22e-22, t-test). As illustrated in Figure 4.5 A, the probability of $p(a_t = L|s_{t-1} = L)$ (blue) matched the probability of $p(s_t = L|s_{t-1} = L)$ (dashed black).

In the Experiment 2, the first-order probabilities are equal $p(L|L) = p(L|R) = p(R|L) = p(R|R) = 0.5$. Accordingly, the conditional probabilities of the subjects' choice given the previous stimulus were around 0.5. For example, $p(a_t = L|s_{t-1} = L)$ was always around 0.5, shown in blue in Figure 4.5 B. The probabilities of the subjects' choice given the previous two stimuli $p(a_t|s_{t-1}, s_{t-2})$ strongly correlate with the second-order probability $p(s_t|s_{t-1}, s_{t-2})$ (correlation coefficient = 0.8955, p=1.24e-10, t-test). As illustrated in Figure 4.5 B, the probability on the left $p(a_t = L|s_{t-1} = L, s_{t-2} = L)$ (purple) matched the probability of $p(s_t = L|s_{t-1} = L, s_{t-2} = L)$ (dashed black). Altogether, these results suggested that the conditional probabilities influenced subjects' decisions in respective experiment.

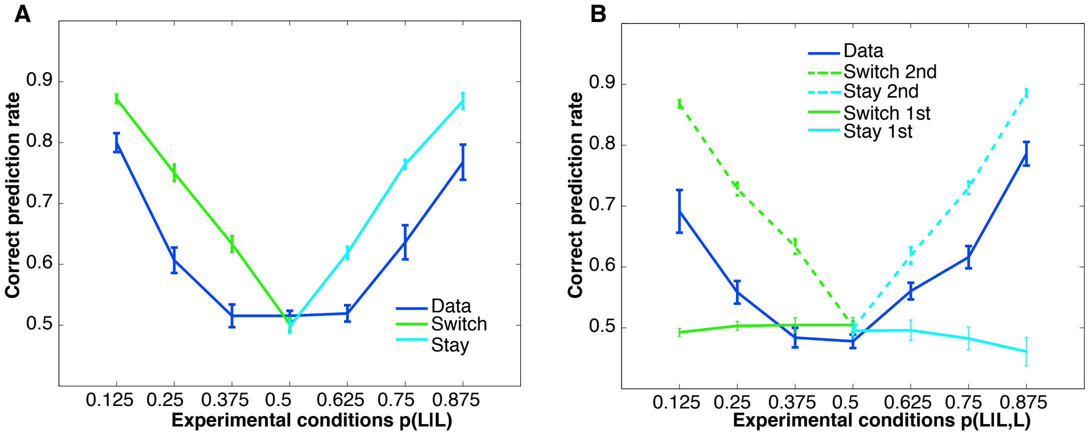

Figure 4.4 Correct prediction rates across each condition, mean over all trials shown in blue. (A) Experiment 1, green line represents ideal predictions that always chose the location opposite to the previous stimulus. Cyan line represents ideal predictions that always chose the same location as the previous stimulus. (B) Experiment 2, green and cyan solid lines represent the ideal predictions as in (A), which failed in the second-order task. The dashed green line represents ideal predictions that always chose the location opposite to the previous-but-one stimulus. The dashed cyan line represents ideal predictions that always chose the same location as the previous-but-one stimulus. Error bars indicate SEM over 10 subjects or 10 simulated agents of ideal predictions.

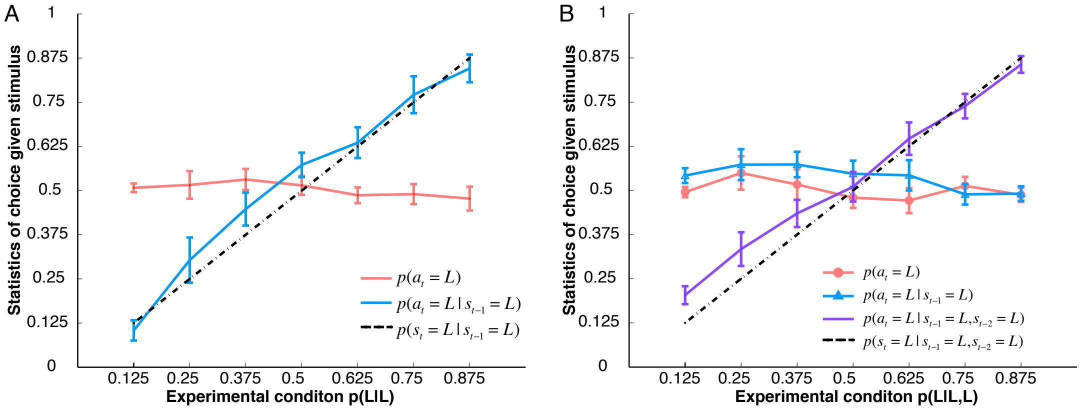

Figure 4.5 Correlations between subjects' choices and the stimuli, suggesting that subjects made decisions according to the conditional probabilities when such information is critical for the task performance. (A) Experiment 1. The probability that the subjects chose left during each experimental condition is around 50%, shown in red. The probability that the subjects chose left given a left stimulus (blue) correlates with the first-order conditional probability $p(L|L)$ (dashed black). (B) Experiment 2. Both the probability that the subjects chose left (red) and the probability that the subjects chose left given a left stimulus (blue) are around 50%. The probability that the subjects chose left given two left stimuli (purple) correlates with the second-order conditional probability $p(L|L,L)$ (dashed black). Error bars indicate SEM.

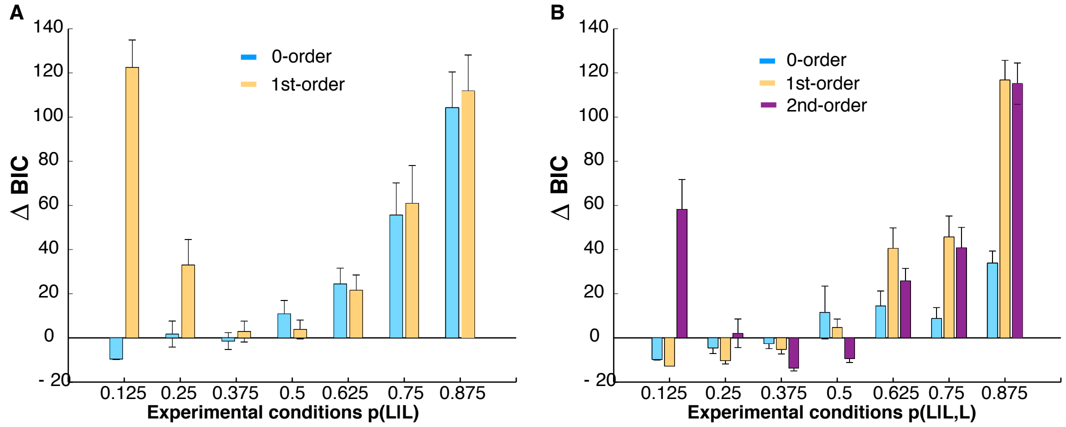

Figure 4.6 Models compared by relative BIC scores ΔBIC. The bigger ΔBIC indicates a better model fit. (A) ΔBIC of zero-order (blue) and first-order (yellow) models fitted into the experimental data in each condition of Experiment 1. The first-order model fitted the data significantly better than the zero-order model. (B) ΔBIC of zero-order (blue), first-order (yellow), and second-order (purple) models fitted into the experimental data in each condition of Experiment 2. The second-order model fitted the data significantly better than the zero-order and first-order models.

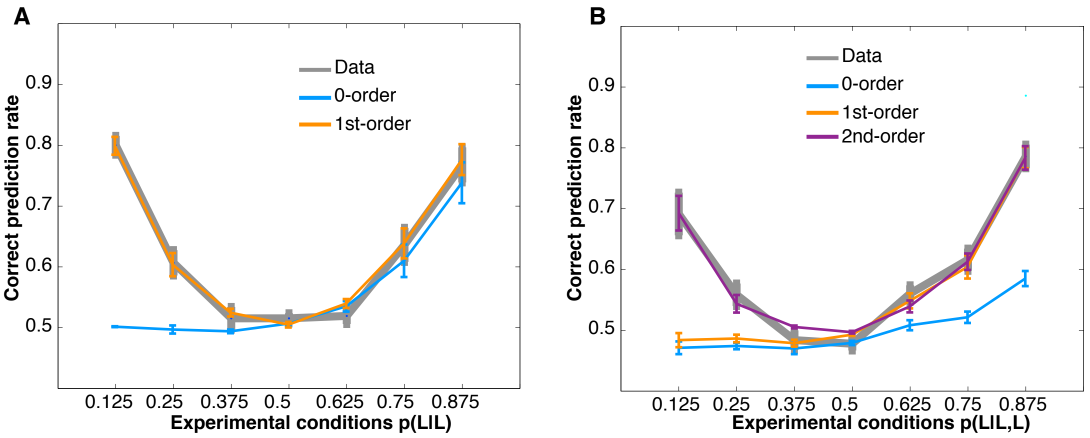

Figure 4.7 Model simulations. (A) In Experiment 1, the best-fitting zero-order model could simulate subjects' choice behavior when $p(L|L) \geq 0.5$, but failed when $p(L|L) < 0.5$. The best-fitting first-order model could precisely simulate subjects' choice behavior across each condition. (B) In Experiment 2, the best-fitting zero-order model could not simulate subjects' choice behavior at all. The best-fitting first-order model could only simulate subjects' choice behavior when $p(L|L,L) \geq 0.5$, but failed when $p(L|L,L) < 0.5$. The best-fitting second-order model could precisely simulate subjects' choice behavior across each condition.

Behavioral results were further analyzed on a trial-by-trial basis using the adapted Rescorla-Wagner models that estimate the conditional probabilities. The model updates the current prediction by sampling the interdependencies among the past two or three stimuli. We fitted the zero-order and first-order models to subjects' choice behavior in Experiment 1. We fitted zero-order, first-order, and second-order models to subjects' choice behavior in Experiment 2. Models were fitted using maximum likelihood estimation and the BIC scores were used to evaluate the goodness of fit.

The BIC scores of the models for each experiment are shown in Figure 4.6 with mean and standard deviation over ten subjects. We reported relative BIC scores, $\Delta BIC := BIC_{random} - BIC_{RW}$, where BIC_{random} is the BIC score of a random agent calculated as $-2log\ (0.5)$, and BIC_{RW} is the BIC score of each candidate model. The larger the relative BIC score is, the better the model fits the data. The BIC scores showed that the behavioral data strongly favored the first-order model for Experiment 1 and the second-order model for Experiment 2, shown in Figure 4.6. In addition, the ΔBIC of the first-order model in Experiment 1 and the ΔBIC of the second-order model in Experiment 2 increased when the respective underlying conditional probability $p(L|L)$ or $p(L|L, L)$ moved away from 0.5. The best-fitting model parameters are summarized in Table 4.1. The exploration parameter β decreased as the respective conditional probability approached 0.5, suggesting more exploration or random choices when the probabilities became less informative.

Furthermore, the ability of each model in explaining subjects' choice behavior on average were displayed in Figure 4.7, demonstrating that the best-fitting first-order model could simulate subjects' choice behavior in Experiment 1 and the best-fitting second-order model could simulate subjects' choice behavior in Experiment 2. In addition, the best-fitting zero-order model could only simulate subjects' choice behavior when $p(L|L) \geq 0.5$, but failed when $p(L|L) < 0.5$ in Experiment 1, because the model estimates the mean reward of either option which was equal across trials. Similarly, the best-fitting first-order model could only simulate subjects' choice behavior when $p(L|L, L) \geq 0.5$, but failed when $p(L|L, L) < 0.5$ in Experiment 2, because not only the frequencies of $p(L)$ and $p(R)$, but also the first-order probabilities (e.g., $p(L|L)$) were fixed at 0.5. The zero-order model could not simulate subjects' choice behavior in Experiment 2 at all. Altogether, the model-based analysis indicated that subjects might be able to learn the conditional probabilities through reinforcement-learning computations.

Table 4.1 The best-fitting parameters of the first-order model in Experiment 1 (first-order Markov task) and the best-fitting parameters of the second-order model in Experiment 2 (second-order Markov task).

Experimental conditions	First-order model in Experiment 1		Second-order model in Experiment 2	
$p(L\|L)$ or $p(L\|L,L)$	Learning rate α (mean, SE)	Temperature β (mean, SE)	Learning rate α (mean, SE)	Temperature β (mean, SE)
0.125	0.09, 0.07	12.16, 1.31	0.15, 0.09	12.21, 2.12
0.25	0.21, 0.33	6.76, 1.38	0.24, 0.11	6.10, 1.53
0.375	0.12, 0.12	3.55, 0.85	0.29, 0.12	3.38, 0.83
0.5	0.31, 0.33	4.02, 1.02	0.15, 0.09	8.76, 2.59
0.625	0.17, 0.29	6.54, 1.62	0.16, 0.09	11.22, 2.22
0.75	0.13, 0.17	7.50, 1.68	0.07, 0.03	13.53, 1.85
0.875	0.27, 0.31	8.34, 1.42	0.19, 0.04	11.67, 2.08

4.1.4 Discussion

In this section, I introduced two sequential learning tasks designed to investigate how subjects learn the temporal dependencies in a sequence of stimuli. We systematically manipulated the first-order and second-order conditional probabilities. Behavioral findings suggested that subjects might learn the conditional probabilities when such information is critical for optimizing their task performance.

We adapted Rescolar-Wagner models to track either the first-order or second-order conditional probabilities instead of estimating the standard zero-order reward frequencies. These models of higher-order probabilities accounted for the choice behavior better than the zero-order model, which failed to integrate any temporal dependencies among neighboring stimuli for guiding the decisions. Model-based behavioral analysis suggested that subjects might be able to learn the conditional dependencies using a common mechanism of error-correction via reinforcement. This leads to the hypothesis that the brain might implement the same computational mechanism in estimating higher-order conditional probabilities as in estimating the zero-order average rewards.

Through integrating this experimental paradigm into fMRI studies, we may address the following research questions: (1) is the zero-order prediction error always computed even though it is task irrelevant? (2) is there a higher-order prediction error signal encoded with respects to the task demand? (3) Are different orders of probabilities learned simultaneously? i.e., Is the lower-order probabilities learned earlier than the higher-order probabilities?

Sequential learning is the ability to learn about the temporal patterns of the environmental stimuli. Such ability is crucial for much of human decision making and learning. Decades of behavioral and neuroscience studies have shown that humans attempt to identify patterns in order to predict future events, even when these patterns are determined randomly (Huettel et al., 2002; Ivry and Knight, 2002; Squires et al., 1976). For example, a gambler may bet on black for the next spin of a roulette wheel if a run of black spots just came up. The effect of temporal patterns has been studied as changes in the behavioral and neural responses to a stimulus that violated a repeating or alternating pattern from the previous stimuli (Daltrozzo and Conway, 2014; Ranganath and Rainer, 2003).

However, in many sequential learning paradigms, it is not clear at which level the temporal patterns are detected and how this information is subsequently used in learning. For instance, the neural encoding of temporal sequences has been conventionally studied with so-called “Oddball” paradigms, where a random serial of repetitively presented audio or visual stimuli are infrequently interrupted by a target stimulus. Subjects are typically asked to react in response to the target stimuli. One plausible proposals as to what is learned during the Oddball tasks is that the conditional probability of the target occurrence is continuously evaluated (Lungu et al., 2004; Stadler et al., 2006). Such conditional probabilities can also be formally quantified with Bayesian learning algorithms of stimulus probabilities (Baldi and Itti, 2010; Ostwald et al., 2012). Nevertheless, sequential learning would require learning the associations among preceding stimuli and the responses associated with different stimuli. Although various studies have shown that humans can combine statistical evidence over time (Cleeremans and Dienes, 2008; Huettel et al., 2005; Jongsma et al., 2006; Miller et al., 2005; Turk-browne et al., 2009; Zhang and Rowe, 2015), it remains a question how such probability detection mechanism is related to associative learning, e.g., reinforcement learning.

Although previous human neuroimaging studies have demonstrated that the average reward probabilities can be estimated through reward prediction errors in the brain (Cooper et al., 2012; Kim et al., 2006; McClure et al., 2003; O'Doherty, 2004; O'Doherty et al., 2004; Pessiglione et al., 2008), there is remarkably little knowledge of whether sequential learning involves the computation of prediction errors as well. More recently, a growing body of evidence indicates the existence of more than one prediction error signals in the brain (Chiu et al., 2008; Daw et al., 2011; Diuk et al., 2013; Gläscher et al., 2010; Lee et al., 2014; Lohrenz et al., 2007; Tobia et al., 2014). In particular, model-based neuroimaging analysis have suggested the existence of prediction error signals for estimating transition probabilities rather than the averages rewards (Daw et al., 2011; Doll et al., 2012; Gläscher et al., 2010; Lee et al., 2014; Prévost et al., 2013; Simon and Daw, 2011). However, these transition probabilities are only limited to characterize the conditional contingencies between two states, e.g., two locations in a maze. It is unclear whether such prediction error signals can be generalized to the learning of conditional contingences between stimuli unfolding in time.

The model-based behavioral analysis presented in this section showed that the reinforcement-learning models could account for a wide range of existing findings. Therefore, neuroimaging experiments are required to directly test and refine the potential dynamic computation of higher-order prediction errors during sequential learning. In particular, a distributed set of regions in the frontal and parietal cortices have been demonstrated to be exquisitely sensitive to the presence of patterns as well as deviations from these patterns (Huettel and McCarthy, 2004; Huettel et al., 2002, 2005; Kirino et al., 2000; Madden et al., 2004; Miller et al., 2005; Paulus et al., 2004; Zhang and Rowe, 2015). However, the exact computational mechanism that accounts for these neural activity is still absent. In addition, the anatomical connections among prefrontal and subcortical regions have yet to be functionally characterized during sequential learning.

Research on sequential stimulus effects has proceeded largely independent of the research on computational reinforcement learning, because the focus is primarily on perceiving patterns rather than optimizing performance through learning. There are very few human sequential learning studies that investigate the dynamic computational processes. Filling this gap, through assembling a unified picture of computational reinforcement-learning modeling and sequential learning, might lead to important breakthroughs in our understanding of the neural mechanisms supporting learning and decision-making.

4.2 The interaction between stimulus likelihoods and rewards

4.2.1 Introduction

Known from Pascal's wager in the 17th century, the rational decision-making procedure consists of identifying all possible outcomes by their expected values and choosing the action that yields highest total expected value. However, modern economists have summarized numerous paradoxical examples showing that our decisions often deviate from Pascal's normative theory (Glimcher and Fehr, 2014). These choice paradigms raise an interesting question: Do we always learn to make decisions that maximize expected value? Over a century of animal learning research has revealed two broad classes of learning processes: Pavlovian conditioning (Pavlov, 1927) and instrumental conditioning (Thorndike, 1933), which showed that animals can both learn to respond to the predictive value of a stimulus and take a particular action to obtain reward.

In situations such as choosing between two alternatives in order to get the most reward, predicting the potential outcomes is critical. Humans are apparently capable of learning the values associated with stimuli but on the other hand tend to violate the rational choice theory. Here we designed a visually cued two-armed bandit task, which aims to dissociate the influence of stimulus likelihood and expected reward on human decisions. The expected values of the both alternatives were defined as the product of the stimulus likelihood and the reward occurrence. We hypothesized that participants would acquire knowledge about the stimulus as well as the reward, while the stimulus driving their decisions away from optimal reward-seeking behavior. We expect to computationally quantify the interaction between stimulus and reward influences.

4.2.2 Methods

4.2.2.1 Experimental paradigm

We designed a task where subjects could acquire information about the stimulus predictability and the relative reward distribution. Participants were presented with a cover story describing the lottery prediction task diagramed in Figure 4.8 and were informed that they would receive what they earn. On each trial, there were two lottery boxes on the left and right sides of the screen. Subjects were instructed to predict the location of the lottery ticket by making a button press on the response trigger pad with the right index or middle finger. If the lottery ticket appeared in their chosen location, they had a chance to win a reward. If the lottery ticket was not in their chosen location, they would not receive a reward. They were further informed that the lottery ticket would occur on each side with a specific probability. Similarly, whether a reward would be delivered after the lottery ticket appeared in the chosen location was also determined by a specific probability. As a consequence, subjects might or might not get a reward even though the lottery location had been correctly predicted.

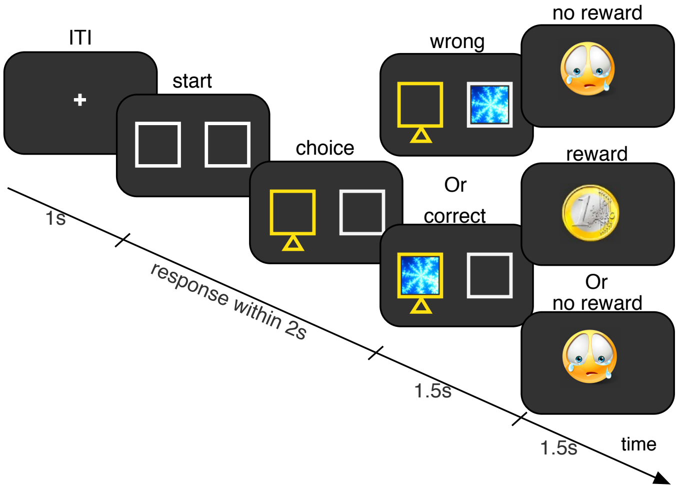

Figure 4.8 Illustration of a sample trial: subject had to make a choice when the two boxes were displayed. For instance, the left box was chosen and opened, after which a lottery ticket would either appear in the chosen box or in the unchosen box according to the stimulus likelihood. If the lottery ticket appeared in the chosen box, subjects would probably earn a reward of 1 Euro depending on the conditional reward probability. If the lottery ticket appeared in the unchosen box, subjects would not earn a reward. Importantly, the expected value of either box was the product of its respective stimulus likelihood and conditional reward probability.

4.2.2.2 Experimental task and procedure

We designed 5 experimental conditions. Each condition had a specific stimulus likelihood (i.e., the probability that a stimulus appeared on the left and on the right) and a conditional reward probability (i.e., the probability that the chosen location was rewarded if the stimulus appeared on the chosen location). The 5 experimental conditions are listed in Table 4.2. The relative outcome is a normalized product of the stimulus likelihood and the conditional reward, which is the expected value of each location. The 5 experimental conditions were organized into 10 blocks of 150 trials each. Each condition was run with two blocks with one block having the contingencies as described in one row of Table 4.2 and another reversal block having the stimulus and reward contingencies switched between the left and the right. The reversal block was designed to counterbalance potential confounds due to a specific choice location. The ordering of the blocks was fully counterbalanced across subjects. Each subject completed 10 blocks, with self-paced breaks in between blocks.

Ten participants (mean age, 22 years; age range, 19-25 years; 8 male and 2 female) with normal or corrected-to-normal vision were recruited from the student population at Technische Universität Berlin. Each participant was paid a base rate of 20€ for participating in the experiment plus a bonus depending on the amount of money won at the end of the experiment (mean 4€ ± SD 3.6€). The bonus payment was calculated as the total amount of rewards accumulated during the entire experiment with each reward was worth 1 cent.

Table 4.2. Stimulus likelihood and the conditional reward for each of the experimental conditions. Relative outcome are the product of stimulus likelihood and conditional reward.

Experimental conditions	Stimulus likelihood L, R	Conditional reward L, R	Relative outcome L, R
1	0.7, 0.3	0.3, 0.7	0.5, 0.5
2	0.7, 0.3	0.2, 0.8	0.37, 0.63
3	0.6, 0.4	0.2, 0.8	0.27, 0.73
4	0.6, 0.4	0.3, 0.7	0.39, 0.61
5	0.6, 0.4	0.4, 0.6	0.5, 0.5

4.2.2.3 Computational modeling

We attempt to computationally explain subjects' dynamic trial-by-trial learning and decision-making process using the framework of reinforcement-learning models (Sutton and Barto, 1998). We adapted 4 variants of reinforcement-learning models to fit subjects' choice behavior in the experiment. We denote a choice between left (L) and right (R) boxes on trial t as $a_t \in \{L, R\}$, the stimulus location as $stim_t \in \{L, R\}$, whether the stimulus location is correctly predicted $(a_t = stim_t)$ or not $(a_t \neq stim_t)$ as $\lambda_t \in \{1, 0\}$, and the reward as $r_t \in \{0, 1\}$.

Reward model. The first model is the standard Rescorla-Wagner model (Rescorla and Wagner, 1972) which assumes that subjects track the expected value EV, for each target location via running average over the reward. This model only updates the expected value of the chosen location with a reward prediction error δ_{RPE}, giving the difference between the received and expected rewards. Notebaly, this model does not account for any perceptual confounds.

$$EV_{a_t}^{t+1} = EV_{a_t}^{t} + \alpha\delta_{RPE}^{t}, \tag{4.16}$$

$$\delta_{RPE}^{t} = r_t - EV_{a_t}^{t}. \tag{4.17}$$

Stimulus model. The second model applies Rescorla-Wagner learning to estimate the stimulus likelihood SV, utilizing a stimulus prediction error δ_{SPE} analogous to the reward prediction error. However, this model ignores the task in which subjects were incentivized to perform for the reward.

$$SV_{a_t}^{t+1} = SV_{a_t}^{t} + \alpha\delta_{SPE}^{t}, \tag{4.18}$$

$$\delta_{SPE}^{t} = \lambda_t - SV_{a_t}^{t}. \tag{4.19}$$

Hybrid model. Following the probabilistic nature of the task, in which the reward is conditional on correct predictions of stimulus location, we extended the trial-based reinforcement-learning scheme to update both EV and SV in parallel and linearly combine them together as an action value AV to capture a trade-off between the stimulus-driven and the reward-driven decisions. This model applies the respective Rescorla-Wagner update to stimulus estimation and reward learning separately. Inspired by the behavioral results, we hypothesized a dynamic transition between the two estimates by further including a trial-by-trial exponential decay function η as a hybrid trade-off.

$$AV_{a_t}^{t} = \eta SV_{a_t}^{t} + (1-\eta)EV_{a_t}^{\mathrm{t}}, \tag{4.20}$$

$$\eta = Ie^{-Kt}. \tag{4.21}$$

Both the offset I and the slope K are free parameters. This model nests the first two models, i.e., it is reduced to the reward model when $I = 0$ and the stimulus model when $I = 1, K = 0$. This model includes a common single learning rate for both the stimulus and the reward update, assuming that subjects learn equally from both observations. To test this assumption, we also fitted a model with two distinct learning rates, one for the stimulus estimation and one for the reward estimation. The model with two learning rates yielded inferior model fit, as measured by DIC, and will therefore not be discussed further.

Forward model. Another strategy for planning actions is to build a cognitive map of the whole task structure with model-based reinforcement learning (Gläscher and O'Doherty, 2010; Doll et al., 2012). To test if subjects perceived the stimulus as a latent state and acquired all the sequential contingencies during the experiment, we implemented a model-based reinforcement-learning algorithm, which uses the experience to estimate a transition function $T(s,a,s')$.

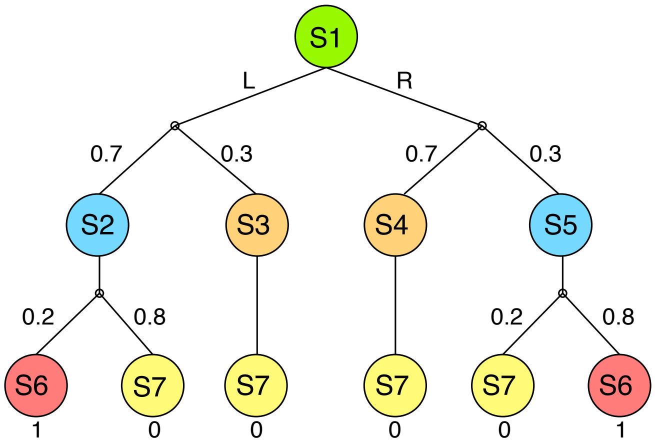

Figure 4.9 Task structure illustrated with experimental condition 2 (related to Figure 4.8). The state-action values are computed by multiplying the reward and the transition probabilities along each path through the decision tree. Subjects selected actions only at state s_1. In this example, the optimal path was from s_1 to s_5 ending at s_6. Thus, the optimal state-action value was $Q(s_1, R)$.

In each trial (Figure 4.9), subject made a choice $a_t \in \{L,R\}$ at the initial state s_1, leading to one of the four second-stage states: $s_2 \coloneqq (a_t = L \,\&\, stim_t = L)$, $s_3 \coloneqq (a_t = L \,\&\, stim_t = R)$, $s_4 \coloneqq (a_t = R \,\&\, stim_t = L)$, $s_5 \coloneqq (a_t = R \,\&\, stim_t = R)$. Each of these four second-staged states was probabilistically associated with either a final reward state $s_6 \coloneqq (r_t = 1)$ or a non-reward state $s_7 \coloneqq (r_t = 0)$. The expected value of each state is the sum over the values of all future states. We denote $s \in \{s_2, s_3, s_4, s_5\}$ and $s' \in \{s_6, s_7\}$, then the state value of each s can be calculated as following:

$$V(s) = \sum\nolimits_{s'} T\,(s, s')V(s'), \tag{4.22}$$

$$T^{t+1}(s, s') = T^{t}(s, s') + \alpha\big(\,\sigma_{s,s'} - T^{t}(s, s')\big). \tag{4.23}$$

Where $\sigma_{s,s'} = 1$ for the observed transition from s to s' and $\sigma_{s,s'} = 0$ for the unobserved transition. Equation (4.22) and (4.23) applies to all s and s'. So the action value at state s_1 is:

$$Q_{a_t} = \sum\nolimits_{s} T(s_1, a_t, s)V(s). \tag{4.24}$$

In all of the 4 models described above, α is a free learning rate parameter which weights the previous experience. For each of the models, we additionally assume a softmax action selection function:

$$P(a_t) = \frac{\exp\,(\beta\mathcal{V}_{a_t})}{\exp(\beta\mathcal{V}_L) + \exp\,(\beta\mathcal{V}_R)}. \tag{4.25}$$

where $\mathcal{V} \in \{SV, EV, AV, Q\}$ for each model respectively. β is an inverse temperature parameter, which capturers the tradeoff between exploration and exploitation.

Model fitting and parameter estimation were conducted using hierarchical Bayesian analysis. The model parameters that were estimated included the learning rate, the softmax temperature, the hybrid offset and the hybrid slope. In the Bayesian hierarchical model, individual parameters for each participant were drawn from group-wise beta distributions initialized with uniform priors. Hierarchical Bayesian analysis proceeded to estimate the actual posterior distribution over the free parameters through Bayes rule by incorporating the experimental choice data. Computation of the posterior was done through the Markov chain Monte Carlo method using the JAGS software and its runjags R interface. Three MCMC chains were run for 150,000 effective samples after 150,000 burn-in samples, which resulted in 30000 posterior samples after a thinning of 5. Each estimated parameter was checked for convergence both visually (from the trace plot) and through the Gelman-Rubin test. To quantitatively compare the model fit, we computed the deviance information criteria and a smaller DIC indicates the better fitting.

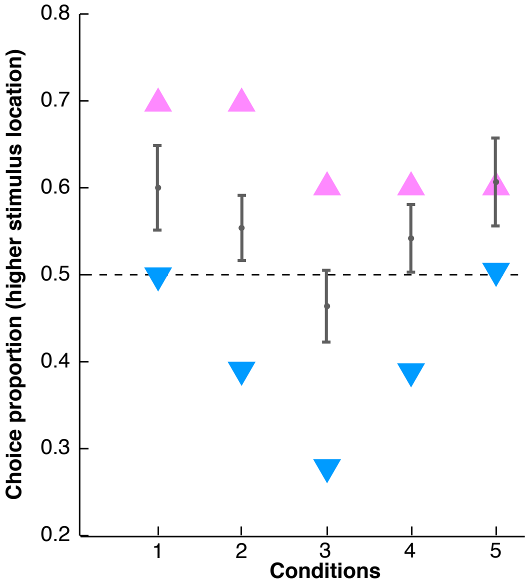

Figure 4.10 Proportion of subjects choosing the location of higher stimulus likelihood, mean indicated in gray with error bars (SEM) averaged over 10 subjects and all the trials. Triangles are theoretical predictions as listed in Table 4.2. The ideal prediction from a stimulus learner is shown in pink and follows the probabilities listed as 'Stimulus likelihood' in Table 4.2. The ideal prediction from a reward learner is shown in blue and follows the probabilities listed as 'Relative outcome' in Table 4.2. Subjects' choices were in between of the stimulus and reward learners, suggesting a dual contribution of both influences. Abscissa labels indicate the experimental conditions.

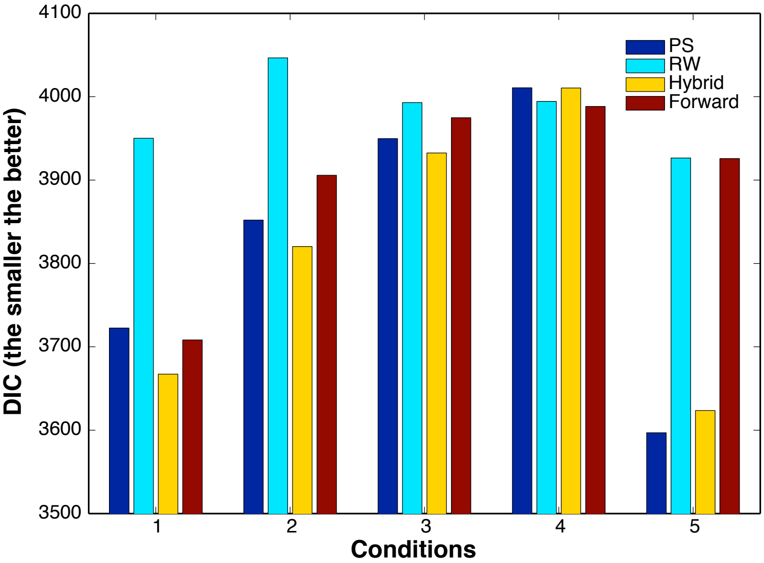

Figure 4.11 Models compared by DIC scores, showing that the hybrid model explains the behavioral data best, on average over the 5 experimental conditions. Abscissa labels indicate the experimental condition.

Table 4.3 The best fitting parameters of the hybrid model.

Experimental conditions	Learning rate α	Temperature β	Offset I	Slope K
1	0.1286	4.6612	0.8427	0.0241
2	0.1811	2.5194	0.9031	0.0175
3	0.6874	0.9893	0.9324	0.0230
4	0.3506	1.4903	0.9591	0.0184
5	0.3132	2.4690	0.9769	0.0089

4.2.3 Results

In each of the experimental conditions, the expected value was either identical for both locations or higher on the location of lower stimulus likelihood. We visualized the data by plotting the probability of subjects choosing the location that had higher stimulus likelihood across experimental conditions. Assuming choice behavior of probability-matching, the ideal reward model predicted that the choice proportion would match the relative outcome probability, resulting in blue shown in Figure 4.10, whereas a stimulus model predicted the choice according to the stimulus probability as shown in purple in Figure 4.10. When expected values were identical for both locations i.e., Condition 1 and Condition 5, subjects consistently preferred the location of higher stimulus likelihood. Subjects preferred the location of higher stimulus likelihood but lower reward probability in Condition 2 and 4, but not as significant as the preference shown in Condition 1 and 5. In contrast, subjects preferred the location of higher reward in Condition 3, when the expected value was lowest on the higher stimulus location in comparison to other experimental conditions. These data suggest that subjects showed sensitivity to both stimulus and reward and tended to trade off reward for the predictability of stimulus location. Choice behavior was systematically modulated by both stimulus likelihood and relative outcome according to the experimental conditions.

Behavioral results were analyzed on trial-by-trial basis using each of the 4 computational models. DIC measures for all the models are summarized in Figure

4.11. The DIC scores showed that the behavioral data strongly favored the hybrid model. The Maximum a posterior of the best-fitting parameters for the hybrid model are summarized in Table 4.3. Together, these results suggest that subjects were simultaneously learning the stimulus and reward contingencies based on separate reinforcement-learning computations.

4.2.4 Discussion

In this section, I introduced a learning paradigm in which subjects had to combine both stimulus likelihood and reward probability to guide their choices. We quantified the modulatory effect of each of the constituents on the overall expectation. The choice behavior was influenced by stimulus- and reward-based expectations depending on the particular experimental context. We designed the paradigm in a way that successful learning required stimulus-response and action-outcome associations being learned in parallel. Subjects were slightly more influenced by the stimulus likelihood than by the reward probability in the task. This can perhaps be attributed to the fact that the reward probability was conditioned on the stimulus likelihood. Only if subjects correctly predicted the stimulus location, they would get a chance to win a reward. Therefore, the stimulus-seeking behavior could be interpreted as seeking for information, because subjects could only access the reward distribution after they had successfully predicted the stimulus.

When we kept the stimulus likelihood the same but changed the conditional reward across different experimental conditions, subjects' choice preferences changed in accordance to the expected values. These results suggest an interaction between reward and stimulus learning, which was also reflected in the model-based behavioral analysis showing the maximum a posterior of the hybrid parameter η larger than 0. The hybrid-learning model suggests that subjects might be estimating both values in parallel. These results raised the possibility to seek for neural correlates of both learning signals. With the current paradigm, we tested hypothesis about the neural encoding of dynamic interactions between stimulus-response and action-outcome learning in Chapter 5.

4.3 Fictive and factual prediction errors in strategic learning [I]

4.3.1 Introduction

Instrumental conditioning involves learning to select actions that will maximize reward and minimize punishment in a complex world. Computational models of reward-based learning assert that learning occurs when expectations are violated. Expectations are derived from experience and variety of information can influence the decisions. Prediction errors occur when expectations are inconsistent with the factual consequences. Counterfactual consequences, the gains and losses associated with alternative actions that were not selected, can affect subsequent choices as well, presumably by influencing the computation and representation of expectations.

Although behavioral evidence indicate that humans take counterfactual outcomes into account when making decisions, reinforcement learning models typically generate prediction errors only for the factual consequences. To investigate the process of counterfactual learning, we designed a strategic sequential investment task that overtly emphasized the counterfactual outcome. We extended the standard Q-learning model by incorporating both counterfactual gains and losses as potential learning signals. Inspired by the behavioral findings, we applied a wide range of modified versions of Q -learning models according to different hypotheses about how the counterfactual information may affect subsequent decisions.

[I] The experimental paradigm and data presented in this section was first published, in a different format, in: M. J. Tobia*, **R. Guo***, etc, “Neural systems for choice and valuation with counterfactual learning signals.,” *Neuroimage*, vol. 89, pp. 57–69, Apr. 2014. Copyright © 2013 Elsevier Inc. (*equally contributed)

4.3.2 Methods

4.3.2.1 Experimental paradigm

The strategic sequential investment task was designed to investigate the potential use of counterfactual consequences in terms of fictive prediction error (FPE) signals during reward-based learning. On each trial, participants decided how much money to invest in a financial market, and then learned about the factual and counterfactual outcomes in succession. The task design included a complex state-space as shown in Figure 4.12, comprised of four possible paths, i.e., a sequence of three states. Each path had a different chance of gaining or losing money in the long run. Every particular state was uniquely identifiable by a different neutral visual background. As illustrated in Figure 4.12, the paths leading to the state 4 and the state 6 were associated with long-term gains, with state 4 being the most lucrative.

Participants completed 80 rounds of the strategic sequential investment task where each round started at state 1, consisted of three decisions and ended in state 4, 5, 6, or 7 (Figure 4.12). On each trial, participants chose an amount of money to invest (0, 1, 2 or 3 Euros). Their experienced path through the virtual maze was determined by the magnitude of their investments rather than the outcome of the trial. Investing 0 or 1 Euro was defined as *risk-averse* investments, which led the subjects to odd-numbered states, i.e., state 3, 5 and 7. These states were non-lucrative, losing states. Investing 2 or 3 Euro was defined as *risk-seeking* investments, which led the subjects to even-numbered states, i.e., state 2, 4 and 6. States 4 and 6 are lucrative, winning states. In order to identify and follow the optimally lucrative path, participants needed to make strategic decisions by accepting interim losses at state 2 so as to gain access to the most lucrative state 4. As such, decisions based on the expected values needed to take into account anticipated future rewards, rather than only considering reward from the current state.

Thirty males aged 18-30 years (mean = 23.8, SD = 3.2) were recruited from the student population at University of Hamburg. All experimental protocols were approved by the ethics committee of the medical association of Hamburg and carried out in accord with the Declaration of Helsinki.

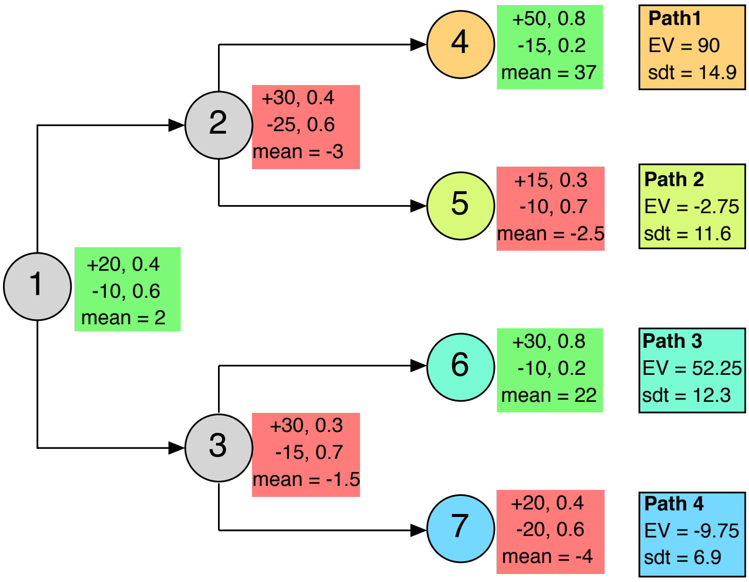

Figure 4.12 State space of the strategic sequential investment task. Numbered circles indicate the seven states and arrows denote the possible transitions. The seven states differ with respect to their winning and losing probabilities as well as the mean amount of monetary gains and losses. In each state, the underlying outcome is generated by a bi-Gaussian distribution with a standard deviation of 5. Panels next to each state provide information about the mean (top left) and probability (top right) of the win Gaussian along with the mean (middle left) and probability (middle right) of the loss Gaussian, while the mean outcome of the state is presented at the bottom of that panel. For example in state 1, mean=20*0.4 + (-10)*0.6 = 2. On a particular trial, the outcome equals to the price change multiplied by the amount of money invested by the subject. The right-most panels provide the total expected rewards (EV) and the standard deviations (std) for each possible state sequence (i.e., Path 1 to Path 4) under the assumption that every path is experienced equally. States where the state characteristics are indicated in green panels have positive mean outcomes, i.e. state 1, 4, and 6, whereas states with red panels have negative outcomes, i.e. 2, 3, 5, and 7. Each state is associated with a particular neutral background (see Figure 6.1 for an example).

4.3.2.2 Computational modeling

Q-learning is a model-free reinforcement-learning algorithm that learns the state-action values via the temporal difference (TD) between obtained and expected rewards. However, the TD prediction error only involves factual consequences about the selected action. In order to assess a counterfactual learning process, I modified the standard Q -learning model by incorporating counterfactual consequences into the valuation via a two-stage process.

The strategic sequential investment task consists of seven states each indicated by a unique background stimulus, where subjects choose among four actions of investing 0, 1, 2, or 3 Euro. We denote the action at trial t by $a_t \in \{0,1,2,3\}$ in the state $s_t \in \{1,2,\ldots,7\}$. For the Q-learning model, we initialize all the Q-values with 0. We denote the market change by o_t, which is drawn from the reward function of each state as described in the green or red panels in Figure 4.12. After having chosen an action a_t in the current state s_t, observed the successive state s_{t+1} and received reward $r_t = a_t \cdot o_t$, a standard Q-learning model updates the Q-value of the current state-action pair as following:

$$Q(s_t, a_t) \Leftarrow Q(s_t, a_t) + \alpha \underbrace{[\, r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(\, s_t, a_t)\,]}_{\text{TD}-\text{error}}, \tag{4.26}$$

where a' represents all the possible actions in the succeeding state. The learning rate α determines the speed of changes in behavior and the discount factor γ reflects the preference of short-term over long-term rewards. Notably, this Q-learning model ignores any counterfactual information provided during the task.

In the fictive phase, subjects observed the counterfactual outcome associated with a maximal bet of 3 Euro, as shown in Figure 6.1. We define the experienced fictive error differently for positive and negative market developments, because both behavioral and neuroimaging data suggest that their impact on decisions and neural activity might be different (Chandrasekhar et al., 2008; Fujiwara et al., 2009; Lohrenz et al., 2007; Pieters and Zeelenberg, 2007):

(1) When the market goes up and if less than 3 Euro is invested, the counterfactual loss is defined as the amount of money one could have won more, i.e., $f_+ = 3o_t - r_t$.

(2) When the market goes down and if more than 0 Euro is invested, the counterfactual gain is defined as the amount of money one would have lost more, i.e., $f_{-} = r_t - 3o_t$.

This was the information presented to the subjects at the fictive outcome phase as shown in Figure 6.1. For this reason, we integrate such counterfactual outcome into the Q-learning model as a second update:

$$Q(s_t, a_t) \Leftarrow Q(s_t, a_t) + \alpha_{FPE} \; [\underbrace{f_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(\, s_t, a_t)}_{\text{fictive TD-error}}], \quad (4.27)$$

where $\alpha_{FPE} \in \{\alpha_{+}, \alpha_{-}\}$ and $f_t \in \{f_{+}, f_{-}\}$. Respectively, α_{+} is the learning rate over counterfactual loss f_{+} when the market goes up, and α_{-} is the learning rate over counterfactual gain f_{-} when the market goes down. The introduction of these additional two parameters enables the model to update the expected values with counterfactual gains and losses differently.

After this two-stage update, actions are selected stochastically through a softmax function as following:

$$P(s_t, a_t) = \frac{\exp\,(\beta Q(s_t, a_t))}{\sum_{a=0}^{3} \exp\,(\beta Q(s_t, a))}. \quad (4.28)$$

In total, this model contains 5 free parameters: discount factor γ, standard Q-learning rate α, counterfactual loss learning rate α_{+}, counterfactual gain learning rate α_{-}, and inverse temperature parameter β. This model nests the standard Q-learning model, i.e., it is reduced to the standard Q-learning model when α_{+}= 0 and α_{-}= 0.

We denote this model, containing counterfactual outcome, as FPE-Q model. The two-stage update with Equation (4.26) and Equation (4.27) can be reduced into a single update of a risk-sensitive Q model where the reward is replaced by a utility function of the true reward:

$$Q(s_t, a_t) \Leftarrow Q(s_t, a_t) + \alpha[\, U(r_t) + \gamma \max_{a'} Q(s_{t+1}, a') - Q(\, s_t, a_t)\,]. \quad (4.29)$$

$U(r_t)$ is a subjective utility function as following:

$$U(r_t) = (1-p)r_t + pf_t, \tag{4.30}$$

and $0 \le p \le 1$. Details for deriving the Equation of (4.29) and (4.30) are shown in the supplementary material of this section.

4.3.3 Results

All the visits to state 4 by all participants during the final 10 rounds were characterized by a maximal investment (i.e., 3 Euro), suggesting that all the participants recognized the value of the state. The difference in their strategic decisions concerned their ability to take a sequence of actions that led to state 4. I categorized the subjects as risk-seeking and risk-averse in terms of an exploratory analysis. Risk-seeking subjects were defined as those who reached the most lucrative state (i.e., state 4) in at least 7 out of the last 10 rounds and invested 3 Euro there. 17 out of 30 subjects were risk-seeking and the rest of subjects were risk-averse. On average over 240 trials, risk-seeking subjects earned €18.91 ± €0.96 (mean ± SD), in comparison risk-averse subjects earned €3.25 ± €0.90 (mean ± SD). Risk-seeking subjects earned significantly more than risk-averse subjects (t=13.39, p=1.15e-12, two-sample t-test). An example from the risk-seeking subjects is shown in Figure 4.13 A and an example from the risk-averse subjects is shown in Figure 4.13 B. The probabilities of choosing the 4 decision paths were calculated in each bin of 10 rounds across time. The path leading to state 4 was the most lucrative followed by the path ending in state 6, whereas the other paths had negative expected values. The choice behavior demonstrated that individuals might have different regret sensitivities to counterfactual consequences. Crucially, subjects needed to risk money in states 2 and 3 in order to explore the whole state space and to exploit the lucrative paths 1 and 3.

The model's free parameters were individually fitted onto each subjects' choice behavior by maximum likelihood estimation. The FPE-Q model nests the standard Q model, so we can compare their goodness of fit with the likelihood ratio test. The FPE-Q model fited the behavioral data significantly better than the standard Q model (likelihood ratio test statistic and p value averaged across subjects: $\chi^2(2) = 49.87, p = 1.48e-11$). As shown in Table 4.4, the better model fit was also demonstrated by a smaller BIC score. Some subjects showed rapid switch towards a better action sequence even after two third of the experiment. This suggests that those subjects might have employed a valuation system that allowed a rapid switch in computational policy. For this reason, we also tested the model-based reinforcement-learning methods with a model-based Q model, referred to as the forward model in Table 4.4. This forward model directly learned the reward and transition functions of the task, but it did not yield better model fit in explaining the behavioral results.

Table 4.4 Model comparison and the best fitting parameters of each model. The best fitting model is indicated by *.

Model	BIC	Best fitting parameters (mean, SE)				
		α	β	γ	α_+	α_-
Q	527	0.25, 0.05	6.15, 0.88	0.80, 0.06	–	–
FPE-Q	487*	0.16, 0.01	6.78, 0.50	0.68, 0.05	0.03, 0.01	0.22, 0.04
Forward	514	0.21, 0.05	8.08, 2.28	0.32, 0.06	–	–

α : learning rate; β : exploration parameter; γ: discount factor; α_+: sensitivity to counterfactual loss; α_-: sensitivity to counterfactual gains.

The FPE-Q model explained the behavioral data better than standard Q models. The behavioral responsiveness to counterfactual gains was associated with optimal performance. On the one hand, risk-seeking subjects showed a significantly smaller learning rate (mean=0.11, SE=0.02) for counterfactual gains than risk-averse subjects (mean=0.36, SE0.06)(t=4.14, p=1.45e-04, two-sample t-test), as shown in Figure 4.14 A. On the other hand, risk-seeking subjects showed a significantly greater discount factor (mean=0.79, SE = 0.04) for the future rewards than risk-averse subjects (mean=0.55, SE=0.1)(t=2.4, p=0.01, two-sample t-test), as shown in Figure 4.14 B. These parameters indicate that risk-seeking subjects were more motivated to the long-term rewards. The transitions between states of the task were designed in a way that investing more while taking interim loss led to the most rewarding state. Thus, less sensitivity to counterfactual gains promoted optimal decision sequence of path 1. The results suggest that risk-seeking subjects did not regret that they risked more money in the current state and were more often transferred to the most lucrative path as a result.

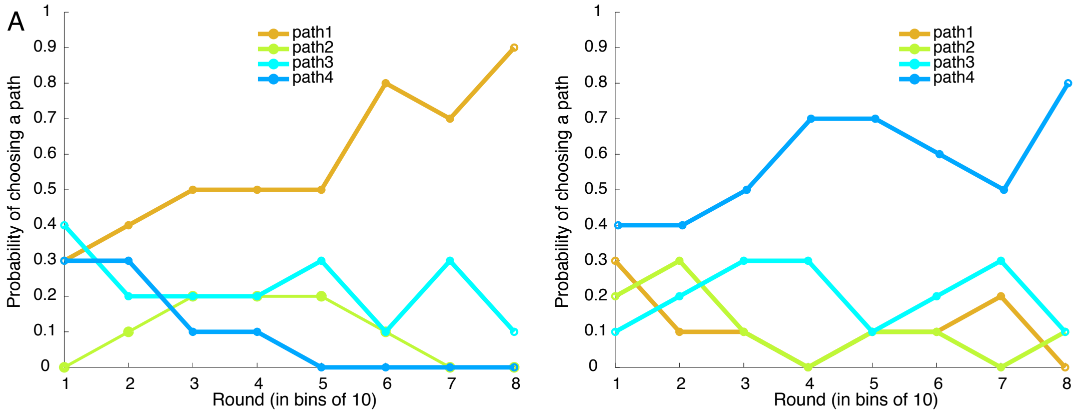

Figure 4.13 Behavioral results contrasting risk-seeking against risk-averse subjects. Choice data are binned into eight 10-round bins, where each round consists of three actions of investment and ends up with one path (see Figure 4.12 for details about each path of sequential decisions). The mean probability of subjects choosing each path is plotted across the time in four different colors. Path 1 is the most rewarding path followed by Path 3; Path 2 and Path 4 have negative payoffs. (A) One of the risk-seeking subjects who has been choosing the optimal policy most of time. (B) One of the risk-averse subjects who has been choosing path 4 most often, which is the worst policy during the task with respect to maximizing rewards.

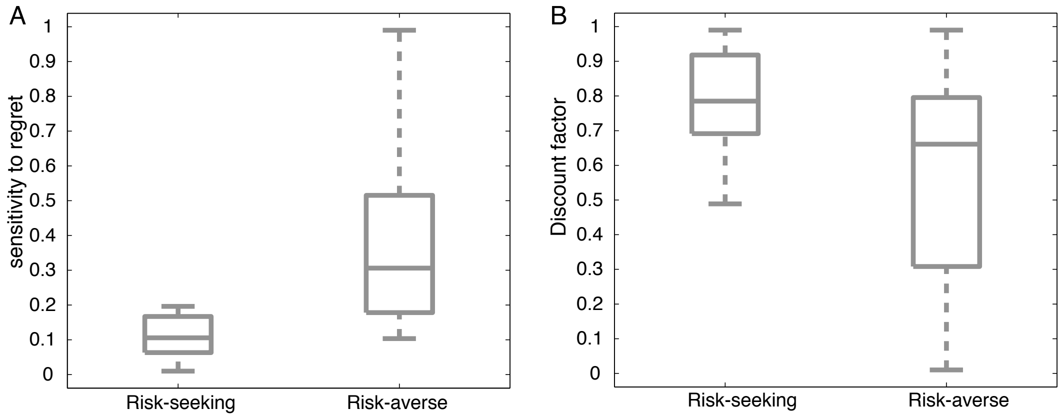

Figure 4.14 Best-fitting parameters from the FPE-Q model suggesting different decision strategies between the risk-seeking and risk-averse subjects. (A) The sensitivity to counterfactual gains (α_-). Risk-seeking subjects have significantly smaller α_-than risk-averse subjects. (B) The discount factor for the future reward (γ). Risk-seeking subjects have significantly greater γ than risk-averse subjects.

4.3.4 Discussion

In this section, I introduced a strategic sequential investment task, which includes state-space structure and state-transition rules that the subjects could learn. The task was designed to investigate the effects of both counterfactual losses and gains on valuation and choice behavior. The task overtly presented the counterfactual outcome on each trial. Critically, learning could be accelerated when a subject reacted to the fictive prediction error or experienced regret.

The computational model incorporated the counterfactual outcome by computing a fictive error signal that was subsequently used to update the state-action values. The FPE-Q model explained the behavioral data significantly better than the pervious Q models or a random agent. The modeling results suggest that fictive temporal-difference prediction errors might be utilized as a computational learning signal. In addition, the FPE-Q model showed that subjects had different strategies, i.e., risk-seeking and risk-averse. The task design therefore allowed the characterization of the fictive prediction error as a valid learning signal with respect to its influence on behavior and its potential neural mechanism that would not be revealed by the standard Q model.

We expected that subjects who successfully exploited the task to maximize their long-term gains would demonstrate a different pattern of neural correlates compared to those subjects who failed to exploit the task. The subjects were scanned with fMRI while they performed this task. We investigated the neural signals related to the expected values and fictive errors using model-based fMRI analysis. The results are discussed in Chapter 6.

From a theoretical point of view, the FPE-Q model works the same as a risk-sensitive reinforcement-learning model that converges to a subjective estimation of the true expected value. This interpretation is consistent with utility theory (Bernoulli, 1954), which measures the subjective preference for a given outcome. The central hypothesis of expected utility is that subjects choose the highest expected utility instead of the highest expected value. Although the FPE-Q model is a post hoc model that describes the decisions by means of counterfactual valuation, we can formulate the two-stage update into a single calculation in the form of a risk-sensitive Q-learning model (Shen et al., 2014). By doing this, we can examine the properties and convergence of this model in terms of MDP.

4.3.5 Supplementary material: examining the property and convergence of the FPE-Q model

Firstly, at trial t, we denote $Q_m \coloneqq \max_{a'} Q(s_{t+1}, a')$, the factual outcome $r{:} = a_t \cdot o_t$ and the reference maximum outcome $\hat{r}{:} = 3 \cdot o_t$. Accordingly, the counterfactual outcome is $f_t = |r - \hat{r}|$. Then, the first-stage update, i.e., Equation (4.26), can be written as:

$$Q^1 = (1-\alpha)Q + \alpha(r + \gamma Q_m). \tag{4.31}$$

For a positive market development ($o_t \geq 0$), the second-stage update, i.e., Equation (4.27) is therefore as following:

$$Q^2 = (1-\alpha_+)Q^1 + \alpha_+(\,\hat{r} - r + \gamma Q_m), \tag{4.32}$$

which equals to

$$\begin{aligned} Q^2 &= (1-\alpha_+)[(1-\alpha)Q + \alpha(r + \gamma Q_m)] + \alpha_+(\hat{r} - r + \gamma Q_m) \\ &= (1-\alpha'_+)Q + \alpha'_+\left(r + \frac{\alpha_+}{\alpha'_+}(\hat{r} - 2r) + \gamma Q_m\right) \\ &= (1-\alpha'_+)Q + \alpha'_+(r_+ + \gamma Q_m), \end{aligned} \tag{4.33}$$

where $\alpha'_+ = \alpha_+ + \alpha - \alpha\alpha_+ \in [0,1]$ and $r_+ = r + \frac{\alpha_+}{\alpha'_+}(\hat{r} - 2r)$. If we denote $p_+{:} = \frac{\alpha_+}{\alpha'_+} \in [0,1]$, we have:

$$r_+ = (1-p_+)r + p_+(\hat{r} - r). \tag{4.34}$$

Similarly, for a negative market development ($o_t < 0$), we have:

$$\begin{aligned} Q^2 &= (1-\alpha_-)Q^1 + \alpha_-(\,r - \hat{r} + \gamma Q_m) \\ &= (1-\alpha'_-)Q + \alpha'_-\left(r - \frac{\alpha_-}{\alpha'_-}\hat{r} + \gamma Q_m\right) \\ &= (1-\alpha'_-)Q + \alpha'_-(r_- + \gamma Q_m), \end{aligned} \tag{4.35}$$

where $\alpha'_- = \alpha_- + \alpha - \alpha\alpha_- \in [0,1]$ and$r_- = r - \frac{\alpha_-}{\alpha'_-}\hat{r}$. If we denote $p_-{:} = \frac{\alpha_-}{\alpha'_-} \in [0,1]$, we have:

$$r_{-} = (1 - p_{-})r + p_{-}(r - \hat{r}). \tag{4.36}$$

Importantly, the final form of Equation (4.33) and Equation (4.35) are the same as the form of Equation (4.31), which is a standard Q-learning model. Therefore, the two-stage update with Equation (4.26) and Equation (4.27) can be reduced into a single standard Q-learning model, but the reward is replaced by a utility function of the true reward:

$$Q(s_t, a_t) \Leftarrow Q(s_t, a_t) + \alpha[\, U(r_t) + \gamma \max_{a'} Q(s_{t+1}, a') - Q(\, s_t, a_t)\,]. \tag{4.37}$$

In Equation (4.37), $U(r_t)$ is a subjective utility function as following:

$$U(r_t) = (1 - p)r_t + pf_t, \tag{4.38}$$

with $p = p_+, f_t = f_+$ when $o_t \geq 0$ and $p = p_-, f_t = f_-$ when $o_t < 0$.

This utility function is a linear weighing between the factual outcome r_t and the fictive outcome f_t. Both the r_t and f_t are explicitly defined in the task and directly presented to the subjects during the experiment. This utility function provides additional power in explaining the choice behavior with respects to different strategies, such as risk-seeking or risk-averse. In contrast, the standard Q-learning model can only predict risk-neural choice behavior, which is nested by the current model (i.e., $p = 0$) as well. As shown in Figure 4.13, subjects clearly had different risk-sensitivities during the task. This explains why the FPE-Q model fits the choice data better than the previous Q models.

4.4 Summary

Using different experimental paradigms we studied the influence of higher-order reward expectations, volatility of returns, and fictive prediction errors on the decision-making processes. We constructed computational models based on Markov decision processes and reinforcement learning. These models were used to quantify the choice behavior and to understand how humans learn in these tasks, especially when the task requires integrating a variety of information. The behavioral paradigms presented in Section 4.2 and 4.3 were further combined with fMRI experiments in Chapter 5 and 6, which allowed us to relate model variables and parameters to the neural activity.

Chapter 5: CONTEXTUAL MODULATION OF PREDICTION ERROR REPRESENTATIONS

Although the co-existence of both reward and saliency related signals in human ventral striatum has been confirmed, the precise interaction between these two signals has not been fully resolved. In this chapter, I approach this question computationally using the behavioral task described in Section 4.2 and model-based fMRI analysis.

At the neural level I found a co-existence of stimulus and reward prediction errors in the ventral striatum, suggesting that this region responds to general surprising perceptual events as well as unexpected reward delivery or omission. Furthermore, the activation patterns of the stimulus and the reward prediction errors are different under the experimental context. The amygdala correlated with the dynamic interaction between the stimulus-response and the action-outcome learning, suggesting that it might be negotiating between an initial emphasis on choosing the salient stimulus and pure reward-based choices later. In summary, this study highlights the roles of key parts of the decision-making network in learning stimulus- and reward-based choices.

5.1 Introduction

Human decisions sometimes appear to diverge from an individual's explicit desires. Both behavioral and neural evidence has suggested that instrumental and goal-directed systems control behavior concurrently and converge in human striatum. (Balleine and O'Doherty, 2010; Dayan and Balleine, 2002; Dickinson and Balleine, 2002; Rangel et al., 2008). In this chapter, we probe the dynamic interaction between instrumental and goal-directed systems, while having stimulus-response and response-outcome associations being learned in parallel.

We used the behavioral task described in Section 4.2.2, in which subjects had to choose a location of left or right where a stimulus (lottery ticket) would appear with a specific probability that was unknown to the subjects. If the subject made the correct choice, then and only then they would receive a reward with another specific probability. Thus, this task involves two learning objectives: (1) to learn where the stimulus is most likely to appear, i.e., stimulus-response learning, and (2) to learn where the reward is most likely to appear, i.e., response-outcome learning. Critically, we designed two conditions in which the stimulus location was either *unbiased* for the reward location or *biased*, in which the location with the larger stimulus probability was not the location with the higher reward probability. This creates a conflicting situation, which permits us to disentangle the influence of the two learning systems when they both operate but contradict each other.

We expected that subjects' choices would be initially dominated by instrumental stimulus-response learning because the task instructions emphasized that a reward could only be obtained if the stimulus appeared at the chosen location. However, with experience and gradually more knowledge about the probabilistic structure of the task subjects would transit to response-outcome learning and choose the location with the higher reward probability to maximize their payoff, even though that meant choosing the location with the smaller stimulus likelihood in the biased condition. We also predicted that in the latter case this transition from stimulus-response to response-outcome learning would occur slower.

Our findings revealed that both stimulus-response and response-outcome associations were learned with two separate reinforcement-learning models combined by a non-linear weighting function. Transitions from instrumental to goal-directed learning were slower in the biased condition. Both prediction errors correlated with activity in the ventral striatum. Whereas stimulus prediction errors (SPEs) elicited stronger correlations during the biased condition, reward prediction errors (RPEs) elicited stronger correlation in the unbiased condition. Furthermore, the neural response in amygdala correlated with the non-linear weighting function that modeled the influence of each learning system on the expected value for each trial. The results indicated that instrumental and goal-directed systems flexibly interacted during learning to maximize returns.

5.2 Methods

5.2.1 Participants

31 participants with normal or corrected-to-normal vision were recruited from the student population at University of Hamburg. They were screened according to the health and safety requirements for undergoing MRI scanning. Each participant was paid a base rate of €10 for participating in the experiment and a bonus depending on the amount of money won at the end of the experiment (mean 8.9€ ± SD 2.8€). The final analysis included 27 subjects (mean age, 26 years; age range, 20-36 years; 14 male and 13 female). 4 subjects were excluded from the analysis, one due to excessive head motion and three due to failure to perform on more than half trials during task. This study was conducted in accordance with the principles of the Declaration of Helsinki for subject participation in scientific studies and was approved by the local ethics committee (PV3661).

5.2.2 Experimental design

We set up two experimental conditions: (a) Unbiased condition: the two locations had equal stimulus (i.e., presentation of lottery) probability of 0.5, and (b) Biased condition: one location was associated with a higher stimulus probability of 0.7 and the other location was associated with a lower stimulus probability of 0.3. Conditional reward occurred probabilistically with 0.2 or 0.8 for correct predictions on the either location. Critically, the higher reward probability was always assigned to the location where the stimulus probability was lower in the biased condition. The biased condition is the experimental condition 2 presented in Chapter 4.2.2.2 and the unbiased condition is a new control condition.

The rationale behind this design was to provide distinct experimental contexts for the stimulus induced stimulus-response learning and the reward induced action-outcome learning. Note that the biased condition induces a conflict between stimulus and reward learning, in which the salient visual stimulus is not predictive of the higher reward probabilities, whereas in the unbiased condition the stimulus is behaviorally irrelevant.

5.2.3 Experimental task and procedure

Participants were presented with a cover story describing the lottery prediction task diagramed in Figure 5.1 A and were informed that they would receive what they earn. On each trial, there were two lottery boxes on the left and the right of the screen. Subjects were instructed to predict the location of the lottery ticket by making a button press on the response trigger pad with the right index or middle finger. If the lottery ticket appeared in their chosen location, they had a chance to win 1 Euro. If the lottery ticket was not in their chosen location, they would not receive a monetary reward. They were further informed that the lottery ticket would occur on each side with a specific probability. Similarly, whether a reward would be delivered after the lottery ticket appeared in the chosen location was also determined by a specific probability. As a consequence, subjects may or may not get a reward even though the lottery location had been correctly predicted.

Each trial started with a 2s inter trial interval, during which time a fixation cross was presented at the center of the screen. The two lottery boxes were then displayed and the subject had to make a choice. If no choice was made within 2s, a message “Too slow!” was displayed for a time-out of 4s and that particular trial was abandoned. The chosen box was highlighted in yellow with a triangle at the bottom, after which the lottery ticket in the form of a fractal image was shown for 1.5s. After a jittered time interval that lasted between 2s to 4s (uniform distribution) the outcome was presented for 1.5s. This consisted of either a coin (indicating a reward of 1€) or a crying face (indicating no reward). The jitter is a critical component for reducing the correlation between the BOLD signals induced by the stimulus and the reward events.

The two experimental conditions were organized into 8 blocks, as listed in Table 5.1: first, the stimulus and reward contingencies were switched between the left and the right locations in different blocks; second, each probability setup was used in two blocks. Each block was indicated to the subject by a different fractal image. The assignment of fractal images to blocks and the ordering of the blocks were fully counterbalanced across subjects. Each participant completed 8 blocks of 40 trials each, with self-paced breaks in between blocks. The payment was calculated as the average of the reward they won in each block.

The relative outcome is a normalized product of the stimulus likelihood and the conditional reward, which is the expected value of each location. The contingencies in every row are run for two blocks of 40 trials each. Altogether, the experiment consisted of 8 blocks.

5.2.4 Imaging data

5.2.4.1 Acquisition

fMRI data collection was conducted on a Siemens Trio 3T scanner with a 32-channel head coil. Each brain volume consisted of 40 axial slices acquired in a descending manner, with the following T2*-weighted echo planar imaging (EPI) protocol: repetition time, 2260 ms; echo time, 26 ms; flip angle, 80°; field of view, 220 mm, slice thickness, 2 mm; inter-slice gap, 1 mm. Orientation of the horizontal section was tilted at 30° to the anterior commissure – posterior commissure axis in order to improve signals in the medial orbitofrontal cortex (Deichmann et al., 2003).

Data for each subject were collected in 8 runs, each with volumes ranging from 175 to 188, and the first 4 volumes were discarded to avoid T1 saturation effects, i.e., the experimental task started at volume 5. In between runs, subjects were encouraged to take a self-paced break. In addition, a gradient echo field map (short TE, 5 ms; long TE, 7.46 ms; number of echos, 48; echo spacing, 0.73) was acquired prior to EPI scanning to measure the magnetic field inhomogeneity, and a high-resolution (1 mm^3 voxels) T1-weighted structural image was acquired after the experiment with an MP-RAGE pulse sequence. The entire experiment lasted for about one hour, with about 6.7 minutes active scanning during each run.

5.2.4.2 Preprocessing

fMRI data analysis was carried out using SPM8 (Wellcome Department of Imaging, Neuroscience, Institute of Neurology, London, UK). All images were corrected for differences in slice acquisition with the middle slice of the volume as a reference. A voxel displacement map (VDM) was calculated from the field map to account for the spatial distortion resulting from the magnetic field inhomogeneity.

Incorporating this VDM, the EPI images were then corrected for motion and spatial distortions through realignment and unwarping.

Each subject's anatomical image was manually checked and reoriented by setting the origin to the anterior commissure. The EPI images were then coregistered to this origin-corrected anatomical image. The anatomical image was segmented into grey matter, white matter, and CSF using SPM8's New Segment tool. These grey and white matter images were used with SPM8's DARTEL toolbox to create individual flow fields as well as a group anatomical template. The EPI images were then normalized to the MNI space using the respective flow fields through DARTEL's normalization tool. Finally, a Gaussian kernel of 8 mm full-width at half-maximum was used to smooth the EPI images. Details about preprocessing are also described in Section 3.2.1.

5.2.4.3 Model-based fMRI analysis

We conducted model-based statistical analysis of fMRI data by estimating the time courses of stimulus prediction error (δ_{SPE}), reward prediction error (δ_{RPE}), and hybrid decay (η) signal of the hybrid model from each subject's sequence of choices, observed stimulus, and reward information. The first level analysis design matrix for each of the 8 sessions consisted of: (1) two stick-function regressors representing the onset timings of the stimulus and the reward respectively; (2) three parametric regressors calculated from Equation (4.17) (4.19) and (4.21) of the hybrid model, where the stimulus event was modulated by η and δ_{SPE}, and the outcome event was modulated by δ_{RPE}; (3) 6 scan-to-scan motion parameters produced during realignment and a constant term as nuisance regressors. All the regressors were convolved with the canonical hemodynamic response function and entered into a general linear model in SPM8. The parametric modulator η was orthogonalized with respect to δ_{SPE}. As a consequence, any shared variance between the two parametric regressors was assigned to the δ_{SPE} regressor. We specified the model in this order to give δ_{SPE} the maximal explanatory power. For completeness, we also tested an additional GLM specifying the regressor for η first and assigning all the shared variance to η. Both of these models yielded almost identical statistical parametric maps. Therefore, here we only report the results from the GLM specifying the regressor for δ_{SPE} first and the modulator modulator η was orthogonalized with respect to δ_{SPE}.

We calculated first-level single subject contrasts for each regressor of interest. The resulting contrast images were raised to the second-level group analysis as random effects, where one-sample t-tests were conducted for significant effects across the subjects. Activations were tested with a whole-brain corrected threshold of $p < 0.05$ (Forman et al., 1995) using the 3dClustSim program in AFNI (Cox, 1996) corrected with the following parameters: voxelwise p value 0.001, cluster threshold 0.05, 10000 simulations, 146519 voxels (91x109x61 3D grid, 2x2x3 mm^3) in a mask on the whole brain. Based on the results of the Monte Carlo simulation, a minimum cluster size of 30 voxels was the threshold for significance. For displaying purposes we chose a significance threshold of $p < 0.001$ uncorrected.

To further show how well the parametric modulators fit the data, we plotted percent signal change (PSC) using the rfxplot toolbox (Gläscher, 2009). For each subject, average PSCs were extracted from an 8 mm sphere centered on the peak voxel of region of interest identified by the second-level group analysis. For the η parametric modulator, trials were split into 4 bins according the quartile values of η. The events in each bin were modeled as an onset regressor and the parameters (PSC) for these newly created regressors were then estimated. The PSC of these regressors from each bin were used as a measure indicating the average magnitude of the BOLD response.

5.3 Results

5.3.1 Behavioral analysis

On average over 40 trials, subjects earned €6.9 ± €1.1€ (mean ± SD) in the biased condition, which was significantly less than what would have been expected under chance performance (€7.6) across subjects ($t_{(26)}$=3.4, p=0.001, one-tailed t-test). This suggests a rather strong influence of the misleading stimulus likelihood, which results in subjects' ignorance of the profitable choices. By comparison, subjects earned €10.9 ± €2.4 (mean ± SD) in the unbiased condition, which significantly exceeded chance performance by €0.9 on average ($t_{(26)}$=2, p=0.03, one-sample t-test) suggesting that when the stimulus likelihood was uninformative and reward probabilities were learned directly, they easily dominated a randomly behaving agent.

We visualized data from the experiment by plotting the probability of choosing the left location as a function of experimental conditions. Assuming a probability-matching response, the ideal reward prediction model thus predicts that the left choice proportion will match the relative reward probability on the left, resulting in the green dot shown in Figure 5.1 B, whereas a stimulus learner model will predict the choice according to the stimulus probability as shown in the blue dot in Figure 5.1B. These data suggest that subjects showed sensitivity to both sources of information. Subjects preferred the higher stimulus probability but lower reward side in the biased condition of block 1 and block 2. In contrast, they preferred the higher reward location in the unbiased condition of block 3 and 4. Furthermore, choice behavior was consistently symmetric across location-counterbalanced blocks, e.g. subjects showed same proportion of right choices in block 1 as that of left choices in block 2. These results suggest that choice decisions were modulated by both stimulus likelihood and relative outcome. Subjects tended to trade off reward for the stimulus predictability.

To further explore subjects' learning process, we collapsed the data across left and right locations from respective counterbalanced blocks of the same experimental condition and examined the choice data across trials. As shown in Figure 5.2 A, the misleading side with higher stimulus probability in the biased condition was chosen more frequently than in the unbiased condition. The mean percentage of suboptimal choice (e.g., choosing left in block 1) clearly decreased linearly as a function of trial in the unbiased condition, 46% in the first trial quarter vs. 41% in the final quarter ($t_{(107)}$=2, p=0.02, paired t-test), whereas in the biased condition no such decrease was observed (2 (experimental conditions) by 4 (time bins) repeated measures ANOVA test showed that there was a significant main effect of experimental conditions F(1,107)= 36.259, p=2.47e-08 and a significant interaction effect F(3,321)= 3.299, p=0.022. The main effect of time was not significant F(3,321)=0.497, p=0.661). This showed that in the unbiased condition the percent of non-optimal choices decreased, whereas in the biased condition it did not.

Comparable results were observed with respect to response time as shown in Figure 5.2 B. Response times were significantly longer for the unbiased condition (mean ± SE = 594 ± 7.6ms) than for the biased condition (mean ± SE = 570 ± 7.3ms) (2 (experimental conditions) by 4 (time bins) repeated measures ANOVA test showed that there is a significant main effect of experimental conditions, F(1,107)=7.442, p=0.007). This slower response in unbiased condition indicates that a greater mental effort is involved in overcoming the stimulus bias and making decisions for the optimal payoff. This interpretation is supported by the fact that optimal choices in the unbiased condition exhibited longer response time than non-optimal (stimulus-oriented) choices ($t_{(107)}$ = 4.48, p = 9.56e-06).

5.3.2 Computational model-based analysis

We fitted each of our 4 computational models discussed in Section 4.2.2.3 to subjects' trial-by-trial choice using Hierarchical Bayesian Analysis and evaluated relative goodness of fit with DIC measures. Parameter estimates and DIC measures for all the models are summarized in Table 5.2. The DIC scores showed that the behavioral data strongly favored the hybrid model. Furthermore, the ability of each model to account for the pattern of subjects' choice behavior is displayed in Figure 5.3, demonstrating that the hybrid model was performing best in predicting subjects' choices. The model-predicted probability of choosing left had been split into five equal-sized bins. The proportion of subjects' left choices increased linearly with the hybrid model's action probability while the other models failed to capture the behavioral variations.

Furthermore, the best-fitting hybrid trade-off parameter η decayed more quickly in the unbiased condition, suggesting a faster transition to purely reward-based choices than in the biased condition (Figure 5.4 A). The slope of η for the unbiased condition was significantly larger than the slope of η for the biased condition ($t_{(26)}$=8.4, p=3.4e-09, one-tailed paired t-test), whereas the offsets of η for both conditions were not significantly different (p>0.06, paired t-test). As can be seen in Figure 5.4 B, performance of the subjects in the form of the mean total reward accrued was strongly positively correlated with their individual best-fitting slope of η (r=0.68,p=1.2e-08).

Together, these results suggest that subjects simultaneously learned the stimulus and the reward contingencies based on separate prediction errors and dynamically adjusted their decision strategy towards the reward-seeking choice. Thus, we used the hybrid model to inform all our further fMRI analysis. For each subject, we generated regressors for the stimulus prediction error, reward error and decaying trade-off parameter using the Maximum a posterior of the parameters' group posterior distribution as listed in Table 5.2. The group parameters were used to generate trial-by-trial time series, because unregularized parameter estimates from individuals tend to be too noisy to obtain reliable neural results (Daw, 2011).

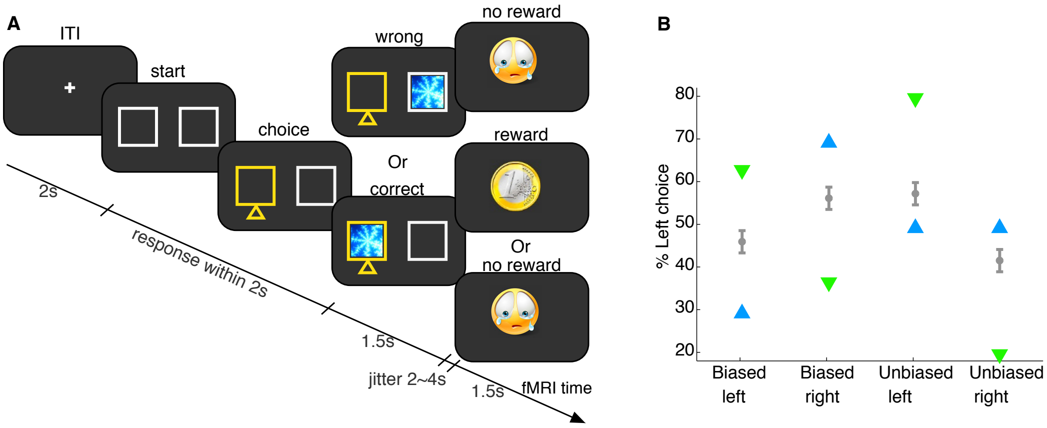

Figure 5.1 Experimental design and choice behavior. (A) Experimental task illustrated in the fMRI timing: subject had to make a choice when the two white boxes were displayed. Here the left was chosen and highlighted in yellow, after which the lottery ticket (fractal image) appeared either on the left or on the right. In this case, if the lottery was on the left, subjects would either win a coin (indicating a reward of 1 Euro) or receive a crying face (indicating a no reward), but if the lottery was on the right, subjects would get a crying face for sure. (B) Proportion of subjects choosing left, mean indicated in gray with error bars (SEM) averaged over all subjects and all trials. Triangles are theoretical predictions as listed in Table 5.1. The ideal prediction from a stimulus learner is shown in blue and follows the probabilities listed as 'Stimulus likelihood' in Table 5.1. The ideal prediction from a reward learner is shown in green and follows the probabilities listed as 'Relative outcome' in Table 5.1. Experimental choices lie in between of the pure stimulus and reward learners, suggesting a dual contribution of both influences. Abscissa labels indicate the experimental condition and the higher rewarding location.

Table 5.1. Stimulus likelihood and the conditional reward for each of the two experimental conditions, counterbalanced between the left and right locations. Relative outcome are the product of stimulus likelihood and conditional reward

Experimental conditions	Block number	Stimulus likelihood L, R	Conditional reward L, R	Relative outcome L, R
Biased	1	0.3, 0.7	0.8, 0.2	0.63, 0.37
	2	0.7, 0.3	0.2, 0.8	0.37, 0.63
Unbiased	3	0.5, 0.5	0.8, 0.2	0.8, 0.2
	4	0.5, 0.5	0.2, 0.8	0.2, 0.8

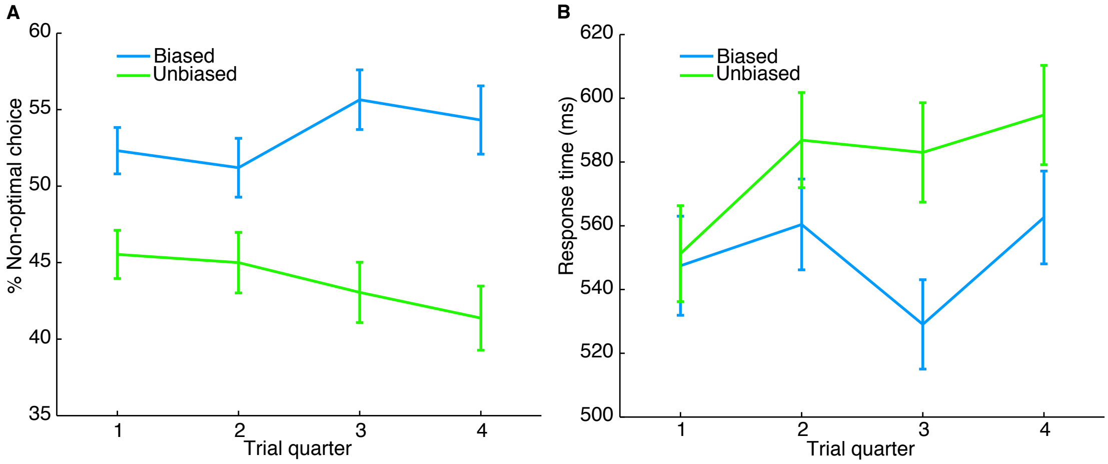

Figure 5.2 Behavioral results contrasting learning under biased and unbiased conditions. (A) Choice data binned into four 10-trial bins. The misleading side with higher stimulus probability in the biased condition shown in blue was chosen more frequently than in the equal condition shown in green. The mean percentage of non-optimal choice decreased linearly as a function of trial in the unbiased condition. (B) Response time binned into four 10-trial bins. The response time from the second to fourth trial quarters was significantly longer for the unbiased condition (green) than for the biased condition (blue). Error bars indicate SEM.

Table 5.2 Model comparison and the best fitting parameters of each model. The best fitting model is indicated by *.

Model	Comparison DIC	Best fitting parameters (Biased, Unbiased)			
		Learning rate	Temperature	Offset of η	Slope of η
Stimulus	11424	0.16, 0.11	1.13, 1.76	–	–
Reward	11439	0.56, 0.47	0.38, 0.70	–	–
Forward	11307	0.04, 0.02	6.68, 13.35	–	–
Hybrid	11164*	0.32, 0.50	2.07, 0.82	0.97, 0.92	0.09, 0.30

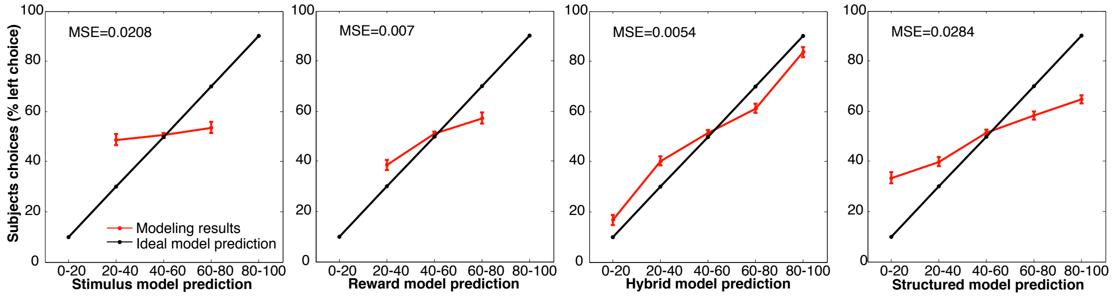

Figure 5.3 Performance of each model in capturing the variations in subjects' choice behavior. Actual choice probability plotted against fitted model choice probability (binned 20% wide), averaged across subjects (error bars represent SEM). The hybrid model is performing best in predicting subjects' choices. The proportion of subjects' left choices fit with the hybrid model's predicted action probabilities, but not with the other three models.

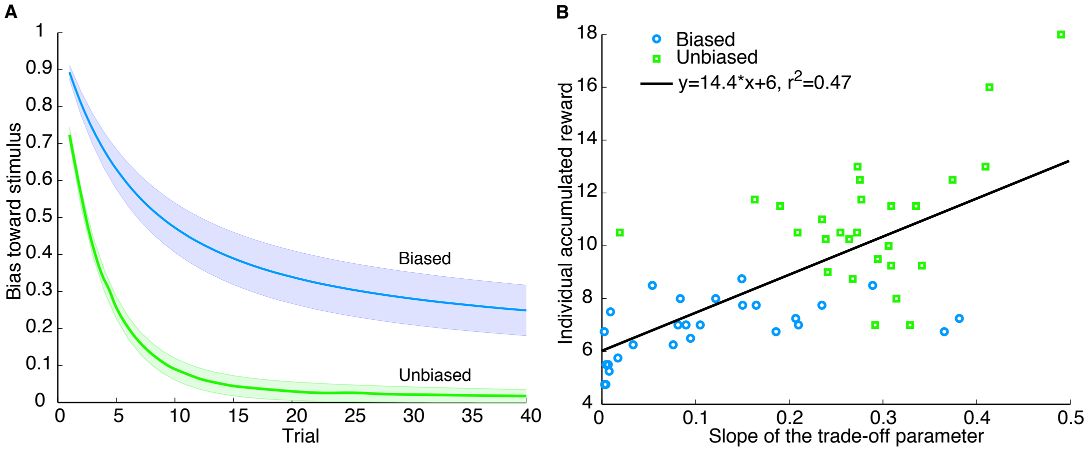

Figure 5.4 (A) Hybrid exponential decay parameter (η) in two experimental conditions. η for the unbiased condition (green) decayed significantly faster than η for the biased condition shown (blue). Shading corresponds to the SEM. (B) Scatter plot of individual accumulated reward in either experimental condition against the individual best fitting slope of η. Each data point represents one subject, blue circle denotes the biased condition and green square denotes unbiased condition. Task performance increases with the slope of η parameter. Participants earned significantly more for unbiased condition than for the biased condition.

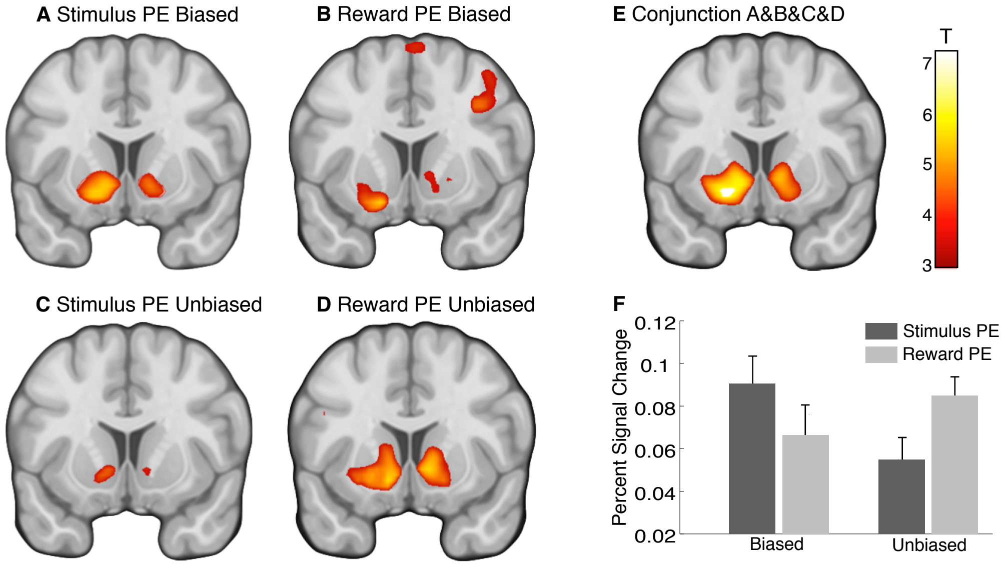

Figure 5.5 (A-D) Coronal view of the map of the t statistics for tests of neural modulation by the δ_{RPE} and δ_{SPE} of the chosen action, showing different effects in the ventral striatum under respective experimental manipulations. (E) Map of t statistics for the neural modulation by a conjunction of the δ_{RPE} and δ_{SPE} showing the overlapping voxels from each prediction error of the entire experiment. (F) The mean percent signal change for the parametric modulators encoding δ_{RPE} and δ_{SPE} in their overlapping voxels showing a significant interaction effect (F(1,27)=6.650, p=0.013). Error bars indicate SEM across subjects. Results are shown at the peak of the conjunction image (-12, 12, -6), p<0.001 uncorrected.

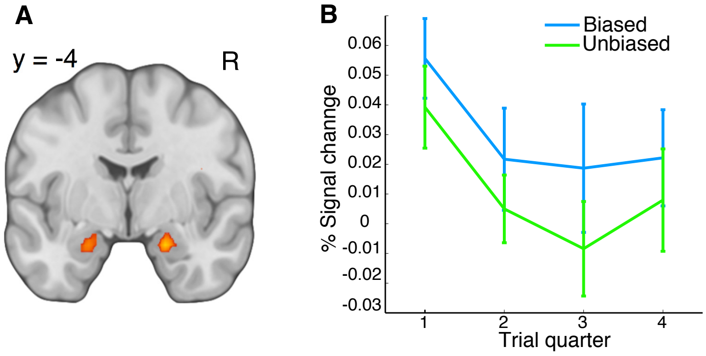

Figure 5.6 (A) Map of the t statistics for tests of neural modulation by the hybrid trade-off parameter η. Results are shown at (-20,-4-18), p<0.001 uncorrected. (B) The mean percent signal change for the parametric modulator encoding η. Variable values are binned into the 25th, 50th, 75th, and 100th percentiles of the parametric modulator. Error bars indicate SEM across subjects.

Table 5.3 Statistical results for the contrast of the parametric regressors

Contrast	Region	Hemi	x	y	z	Peak T
Stimulus prediction error	Putamen	L	-16	8	-9	6.05
	Caudate	R	12	10	-6	4.87
	Inferior occipital gyrus	L	-22	-96	-6	10.27
		R	24	-92	-6	10.45
Reward prediction error	Putamen	L	-22	14	-9	7.14
		R	22	16	-9	6.54
	Insula	L	-30	22	-6	7.45
		R	38	22	-3	8.24
	Middle frontal gyrus	R	40	14	39	5.5
	Superior frontal gyrus	R	8	16	60	4.41
Conjunction between both error signals	Nucleus accumbens	L	-12	12	-6	6.86
		R	8	12	-3	6.01
Decaying trade-off parameter	Amygdala	L	-20	-4	18	4.65
		R	24	0	-21	6.24
	Fusiform gyrus	L	-34	-48	-15	6.84
		R	34	-36	-21	7.77
	IPS/Angular gyrus	R	34	-56	48	6.29
	IPS/Superior parietal lobe	L	-30	-62	45	5.42

All peaks are corrected for whole-brain comparison threshold of $p < 0.05$ (voxelwise multiple comparisons corrected with 3dClustSim) Abbreviations: Hemi = Hemisphere; L = Left; R = right; (x, y, z), MNI coordinates; IPS, intraparietal sulcus.

5.3.3 fMRI results

Our experimental design allowed us to separately assess the neural correlates of stimulus-response association and action-outcome learning. Critically, the biased condition of the experiment provided a conflict where the biased stimulus interfered with the learning of reward association. Correlation of the two prediction error regressors was low (mean correlation coefficient = 0.1, SE=0.0047, p=1.2e-11), so we were confident to identify dissociable neural correlates for each regressor, if they existed.

Coordinates and significance levels for all contrasts assessing parametric modulation are shown in Table 5.3. We first tested for areas showing changes in activity related to the stimulus prediction error and the reward prediction error. We found a co-existence of both the stimulus and the reward prediction errors in the ventral striatum, suggesting that this region responds to surprising perceptual events as well as unexpected reward delivery or omission. Interestingly, the activation patterns of the stimulus and the reward prediction errors were different for the experimental manipulations: The stimulus prediction error was stronger in the biased condition whereas the reward prediction error was stronger in the unbiased condition (Figure 5.5 A-D). This result presumably reflects the fact that subjects went for the stimulus in the biased condition, but tended to optimize performance in the unbiased condition. A conjunction analysis (Figure 5.5 E) confirmed that these two effects occurred in overlapping voxels. We performed additional *post hoc* analyses to look more closely at the overlapping voxels from each prediction error. In particular, we extracted PSC (using the rfxplot toolbox) from a spherical region of interest (radius: 8 mm) centered on the voxel identified in the conjunction analysis (Figure 5.5 E) and conducted a 2 (experimental conditions) by 2 (prediction errors) repeated-measure ANOVA. As shown in Figure 5.5 F, there was a significant interaction (F(1,27)=6.650, p=0.013) confirming that activation was indeed stronger for δ_{SPE} in the biased condition, whereas δ_{RPE} elicited stronger activation in the unbiased condition. These results indicate that shifting the focus from reward to stimulus learning modulates the prediction error representations in ventral striatum.

We next tested for areas showing changes in activity related to the parametric modulation of the decaying trade-off parameter and found significant correlates in the amygdala. Figure 5.6 A shows this activation profile, with Figure 5.6 B showing the PSC for the correlates from each experimental condition. The findings suggest that the amygdala is initially activated during stimulus learning but its activation quickly fades away as the choice behavior transitions to pure reward-based choices. Importantly, the slope of the decay of amygdala activation (Figure 5.6 B) resembles the difference between experimental conditions during the model-based analysis of choice behavior (Figure 5.4 A), namely a faster decay of the trade-off parameter η in the unbiased condition.

To rule out a prominent transitional function from other regions (i.e., IPS, occipital visual area and anterior visual area) that showed correlates with the decaying trade-off parameter, we also conducted *post hoc* analysis on each identified region in the brain and extracted their time course of the PSC. After fitting an exponential function to the time course of the PSC from each region, only the slopes from the amygdala showed significant difference between the two experimental conditions. The slope of the PSC (mean = 0.18, SD = 0.06) from the unbiased condition in amygdala is significantly bigger than the slope (mean = 0.08, SD = 0.04) from the biased condition ($t_{(26)}$=2.8, p=0.0048, paired-ttest). Thus, while other regions, including the fusiform gyrus and the IPS, also exhibited a decay of activation that correlated with the trade-off parameter η, it is only in the amygdala that we observed a difference in decay slopes resembling the observed behavioral dynamics.

5.4 Discussion

We used a probabilistic learning and decision-making task to investigate the neural signatures of stimulus prediction error and reward prediction error associated with the respective stimulus-response and action-outcome learning. Our behavioral analysis demonstrated that participants successfully acquired knowledge about the reward probability in the unbiased condition, in which the stimulus exerted no bias to either choice action. However, in the biased condition, participants failed to overcome the stimulus-reward conflict and consistently preferred the non-optimal choice action. Choice behavior was most consistent with the predictions for the hybrid model, combing stimulus-based and reward-based influences. The hybrid model explained choice behavior significantly better than either the stimulus model or the reward model alone. This indicates that participants implicitly computed both stimulus and reward expectations and relied on these estimates to make decision. We also found that the stimulus-induced choice bias declined over the course of continuous learning in the unbiased condition, along with more cognitive effort indicated by a slower response time. This dynamic transition from initial stimulus bias to reward-seeking preferences was captured by a parameter eta decaying across time. In the imaging data, we found trial-by-trial correlations of the stimulus and reward prediction errors coexisting in the ventral striatum along with amygdala activity mediating the dynamic learning. The fMRI data, together with the computational modeling, therefore allowed us to assess a trial-by-trial parametric signal of latent expectation estimates during learning.

Our finding of reward prediction errors in ventral striatum is consistent with many previous accounts about reward-based learning, where BOLD response in ventral striatum correlated with a reward prediction error in a variety of Pavlovian and instrumental conditioning task (Delgado et al., 2005; Li and Daw, 2011; Montague et al., 2004; Niv and Schoenbaum, 2008; O'Doherty et al., 2003, 2004). More recently, a growing body of evidence indicates the existence of more than one prediction error signal in the brain (Daw et al., 2011; Diuk et al., 2013; Gläscher et al., 2010). Our results extend beyond these previous studies by showing that the same neural populations that encode reward prediction errors might be recruited for encoding value-nonspecific stimulus learning signals. The different neural activation patterns of two prediction error signals under experimental manipulation further raise the possibility that reward context might be extended to the unrewarded stimuli and therefore induce biased behavioral reactions.

Indeed, recent physiological findings from experiments in primates have suggested that dopamine neurons respond to unrewarded physically salient stimuli in highly rewarded contexts (Kobayashi and Schultz, 2014). Our results extend these findings to humans and provide evidence for contextual modulation of striatal BOLD response, which reflects dopaminergic release (D'Ardenne et al., 2008; Knutson and Gibbs, 2007; Schott et al., 2008). Physiological studies have also shown that dopamine neurons can process both primitive and cognitive rewards, thus providing a common instructive signal for both reward-seeking and information-seeking behavior (Bromberg-martin and Hikosaka, 2009). In this context, the neural correlates of stimulus prediction errors might suggest that the ventral striatum encodes a general instructive signal to get the motor system ready for any potential reward. Furthermore, these correlates also complement the previous studies seeking to dissociate value and saliency signals in the human brain (Litt et al., 2011; Mcclure et al., 2003; Zink et al., 2003). These studies have presented evidence of the ventral striatal activations that correlated with saliency computation at the time of decision-making. Salience was defined as any unexpected stimuli or intrinsic motivation related to attention and arousal. Some of the findings are analogous to our results, which suggest a general role of ventral striatum in responding to unexpected perceptual events as well as to reward.

Furthermore, many early studies have identified a role for the amygdala in stimulus-reward learning (Baxter and Murray, 2002; Roesch et al., 2010; Seymour and Dolan, 2008; Whalen and Phelps, 2009). Our findings, BOLD activity in the amygdala correlating with an exponential decay function, may be interpreted in the context of previous human Pavlovian-instrumental transfer (PIT) studies (Prévost et al., 2012; Talmi et al., 2008). These studies demonstrated that instrumental actions are subject to motivational influences from incidental Pavlovian conditioned stimuli, and the amygdala activation is associated with this influence. In those studies, subjects were first trained with separate Pavlovian and instrumental trials and then took a PIT test when both the conditioned stimulus and instrumental options are presented in parallel. However, in our experimental design, we directly expose subjects to the stimulus and reward without any prior knowledge. It is conceivable that they start by considering only the stimulus as potentially rewarding and take actions to obtain the stimulus correctly. In this way, the stimulus gains a Pavlovian conditioning influence on response. This incentive influence gradually vanishes in the unbiased condition when subjects become more aware of the goal-directed reward. In contrast, the stimulus conditioning continues to strongly influence decisions in the biased condition. This different degree of influence is mapped to amygdala activity and suggests a computational role for amygdala in reporting the motivational influence of stimulus on a trial-by-trial basis. The computational role of the decaying function in combing the two prediction errors in the ventral striatum further suggests a functional link between amygdala and ventral striatum in a dynamic learning process. A recent study has computationally characterized learning signals in the amygdala by using a reversal learning task(Li et al., 2011). It suggested that human amygdala codes for cue-reinforcer associations, which is complementary to the striatum's coding of prediction error during associative learning. Our study complements this finding and suggests that amygdala associability coding is not specific to aversive tasks.

A comparison of our results and the literature on cognitive control also provides some insights on the role of the neural networks of attention, particularly cingulate cortex, parietal sulcus and insula, in valuation and decision-making (Bugg and Crump, 2012; Petersen and Posner, 2012). Our task shares common features with paradigms that engage cognitive control (i.e., tasks that require subjects to select relevant information despite their tendency to select goal-irrelevant information). Studies using the flanker interference task, where stimulus location cues the likelihood of incongruent trials, have shown that location-based conflict contexts can implicitly prime retrieval and implementation of top-down attentional control (King et al., 2012). Our data extends these findings to learning processes and provides evidence that enhanced attentional resources may be gradually allocated away from goal-irrelevant stimulus during learning. The reduced conflict in the unbiased condition in comparison to the biased condition suppresses the impact of distracting information on learning.

It is worth noting that the current study does not allow us to answer whether the stimulus triggers an intrinsic rewarding effect or the stimulus interrupts reward learning as a perceptual effect. We did not control subjects' eye fixation in the task and the fractal images stimuli were rich enough to induce strong neural activity in visual cortex. We cannot explore the critical role of visual attention in the computation of value signals in our current task design. There have been studies showing that primary visual cortex encodes expectations about stimulus location (Alink et al., 2010; Sharma et al., 2003). We cannot rule out the possibility that the observed stimulus-driven decision is due to visual attention guided reward coding. Future research might replace the simple lottery prediction task in current study with a psychophysical discrimination task, which can directly manipulate the attentional cognitive effort. This would be a possible approach to further dissociate perceptual effects from the reward-based learning process.

In conclusion, our results highlighted the roles of ventral striatum and amygdala for the decision-making in learning stimulus-response and action-outcome associations and trading off both associations against each other.

Chapter 6: NEURAL SYSTEM FOR VALUATION WITH FICTIVE PREDICTION ERROR[II]

Counterfactual learning refers to the consideration of events that did not occur in comparison to those actually happened in order to determine optimal actions. It can be formulated as computational learning signals, which are referred to as fictive prediction errors. In this chapter, I investigate the functional neural systems involved in counterfactual learning using the behavioral task described in Section 4.3. The purpose of this fMRI study is to determine how the fictive prediction errors contribute to the neural representations of state-action values.

The model-based fMRI analysis suggests that the expected value computed from the fictive prediction error model robustly modulated BOLD signal in the ventral medial prefrontal cortex and orbital frontal cortex. Overlapping neural substrates in the ventral striatum processes both factual and fictive prediction errors. These findings demonstrate that fictive prediction error signals can be an important component of valuation for decision-making. The brain system involved in learning from reward predictions also supports the learning of counterfactual outcomes via fictive prediction error.

[II] Adapted with permission from M. J. Tobia*, **R. Guo***, etc, "Neural systems for choice and valuation with counterfactual learning signals," *Neuroimage*, vol. 89, pp. 57–69, Apr. 2014. Copyright © 2013 Elsevier Inc. (*equally contributed)

6.1 Introduction

Numerous studies have examined the effects of fictive error signals on subsequent choices, as well as the shared neural substrates for processing reward prediction error and fictive error signals (Sommer et al., 2009). But less attention has been given to elucidating how fictive prediction error signals shape the expected values that mediate choice during reinforcement learning, especially during strategic sequential choices for which optimal performance requires accepting interim losses in order to maximize long-term gains. For these reasons, the goal of this experiment was to investigate the functional-neuroanatomical systems involved in valuation that incorporates a fictive prediction error (FPE) signal. In particular, this experiment was designed to address two issues: (1) whether or not FPE signals improve computations of expected value during reinforcement learning, and (2) whether or not the FPE signals correlate with neural activity in the prefrontal and subcortical regions in the brain.

A counterfactual loss, i.e., an amount of reward that was not acquired, occurs on winning trials as a missed opportunity for which an alternative action would have returned a greater reward, and is associated with subjectively experienced regret (Camille et al., 2004; Coricelli et al., 2005). Counterfactual loss promotes choice repetition (Boorman et al., 2013; Nicolle et al., 2011) as well as choices that spontaneously deviate from an established preference (Boorman et al., 2009). On the one hand, it can be used to optimize choice strategy (Li and Daw, 2011; Lohrenz et al., 2007). On the other hand, it may lead to increased subsequent risk taking (Brassen et al., 2012; Büchel et al., 2011).

A counterfactual gain, i.e., an amount of punishment that was not suffered, occurs on losing trials as a reduced cost for which an alternative action would have cost more, and is associated with subjectively experienced relief (Nicolle et al., 2011). Counterfactual gains may lead to differential changes in cognitive performance such as accelerated response times during decision-making (Fujiwara et al., 2009), although the precise nature of these effects is not well elucidated in the literature. For example, Lohrenz and colleagues (Lohrenz et al., 2007) included the fictive error stemming from the counterfactual gain in their analysis of fictive learning signals, but found that it did not significantly predict subsequent choice behavior. This suggests that fictive error signals from counterfactual gains and losses may have dissociable effects on learning and choice behavior.

To investigate the effects of both counterfactual losses and gains on valuation and choice behavior, we designed a strategic sequential investment task that overtly presented the counterfactual outcome on each trial, including the action-contingent state transition rules. The experimental paradigm is inspired by the task used by Lorenz and colleague (Chiu et al., 2008; Lohrenz et al., 2007). However, the task is substantially modified to include a temporal structure and to promote counterfactual learning. I developed a computational model that incorporates the counterfactual outcome by computing a FPE signal in the framework of Q-learning. This FPE signal differs from the fictive error signal studied by Lohrenz and colleagues (Lohrenz et al., 2007) in that it is computed using the temporal difference between expected values and counterfactual consequences, rather than simply the difference between an obtained and unobtained outcome. The FPE signal is subsequently used to update the state-action values. We expected that incorporating counterfactual learning signals into Q-learning would facilitate model performance, and that the expected values would modulate the BOLD signal in ventral medial prefrontal cortex and orbital frontal cortex (vmPFC/OFC) during decision-making, while both factual and fictive prediction errors are processed by overlapping neural substrates in the ventral striatum.

6.2 Methods

6.2.1 Experimental paradigm

The strategic sequential investment task included a complex state-space as shown in Figure 4.12 and described in detail in Section 4.3.2.1. The task was presented to participants in the scanner as an event-related design, shown in Figure 6.1, with 5 stimulus events during each trial. Each trial started with the presentation of a state indicated by a unique visual background cue and a randomly initialized response meter indicating the amount of money to invest on that particular trial (i.e., choice phase). Participants could move the indicator bar on the response meter using an MR-compatible mouse according to the value of their desired investment (0-3 Euro). This stimulus remained onscreen for 3000 ms as the choice phase. This was followed by a brief 500 ms anticipation phase, and then factual and counterfactual outcomes were presented in succession.

The outcome presentation (3000-5000 ms, jittered) informed participants of the amount of money that had been gained or lost on that trial, indicated by a stack of coins. The counterfactual outcome presentation (3000-5000 ms, jittered) informed participants of an additional amount of money that could have been won or lost if the maximum investment, i.e. 3 Euros, was selected. This was symbolized by a second stack of coins that highlighted the difference between factual and counterfactual outcomes. Previously, Lohrenz and colleagues (Lohrenz et al., 2007) employed a similar task design but did not explicitly present the counterfactual outcome on each trial. Instead, they computed a fictive error signal implicitly, based on the difference between factual obtained reward and what would have been obtained if a maximal investment had been selected on that trial.

Each trial concluded with the presentation of a state transition stimulus for 2700 ms, which highlighted whichever of the two possible subsequent states had been selected based on the magnitude of subject's investment. The two possible state transitions were shown simultaneously at the lower and upper portion of the display in a random fashion. This transition event was substituted by an additional feedback stimulus after the third trial of each round (the task always returned to state 1 as the first trial of each round), which indicated the total amount of money gained or lost over the previous three decisions. This multi-trial feedback phase is not presented in Figure 6.1.

Two short rounds of practice trials familiarized participants with the task stimuli and the mouse controls for indicating their choice, and for making ratings of win expectancies, but did not reveal information about the actual win probabilities or expected values that defined each state and path from the task. The first round of practice trials was self-paced. The second round of practice trials was presented at the same speed as the task was going to be presented during fMRI scanning. In the test session in the scanner, the task was presented to participants in 30 rounds during each of 8 scanning runs and ended up with 240 trials in total.

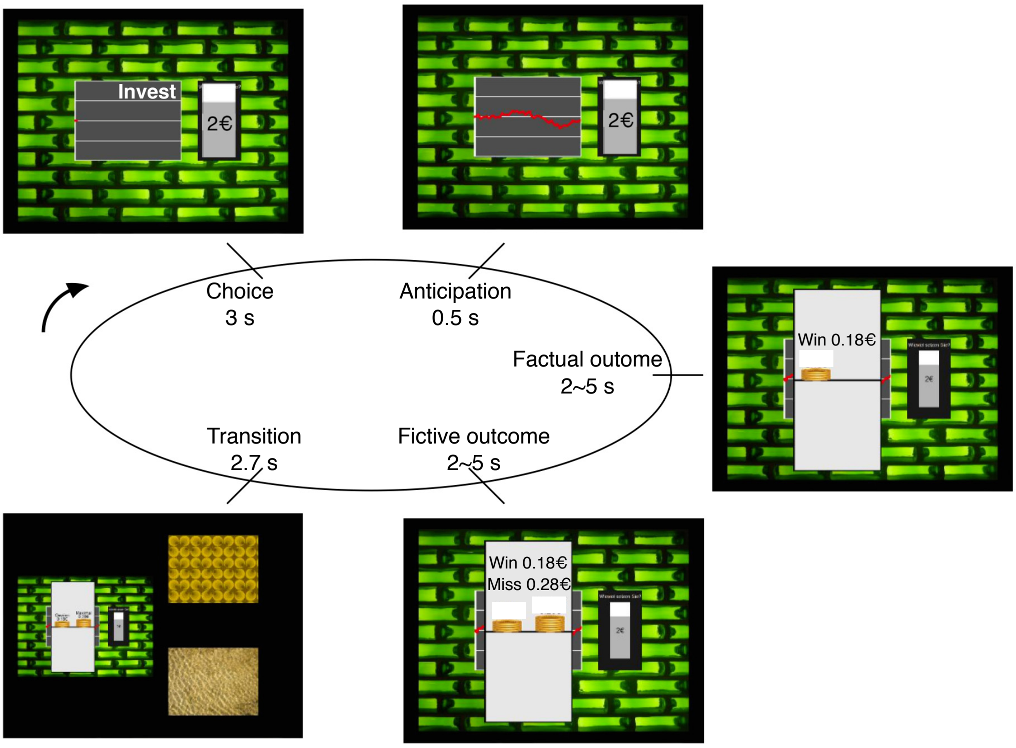

Figure 6.1 Illustration of the task presentation and fMRI timing within a trial, i.e., within a state. Each state was associated with a particular neutral background. By this background picture, participants could learn during the 240 trials of the experiment to associate each state with an expected reward value. In the beginning of each trial, participants decided how much to invest, i.e. 0, 1, 2, or 3 Euros in the stock market of the current state, shown as 'choice' phase on top left next to the arrow. The amount of the investment was indicated in the bar right to the market chart. During a brief 'anticipation' phase, participants observed how the market had developed. Then, they learned in the 'factual outcome' phase how much they won or lost, which was the product of their investment and the market change. The outcome was presented in numbers but also visualized by a positive or negative stack of coins. In the following 'fictive outcome' phase, subjects learned how much they would have won or lost if they would have invested the maximum of 3 Euros. This phase was included to foster counterfactual comparisons, which resulted in a fictive prediction error, i.e. the difference between the factual and the counterfactual outcome. Participants started each round in state 1 and were then transferred through the state space following a transition rule that was unknown to them. In particular, Investments of 0 or 1 Euro led to an odd-numbered state, whereas investments of 2 or 3 Euro led to an even-numbered state in the state structure of Figure 4.12. At the end of each trial, the two possible next states were shown to the participant in the 'transition' phase in random vertical order. Then subjects were transferred to the state according to their decision. After 3 trials, a round ended and subjects were informed about the total win or loss of this round (i.e., the total amount won or lost after experiencing a path of three states), and were transferred back to state 1.

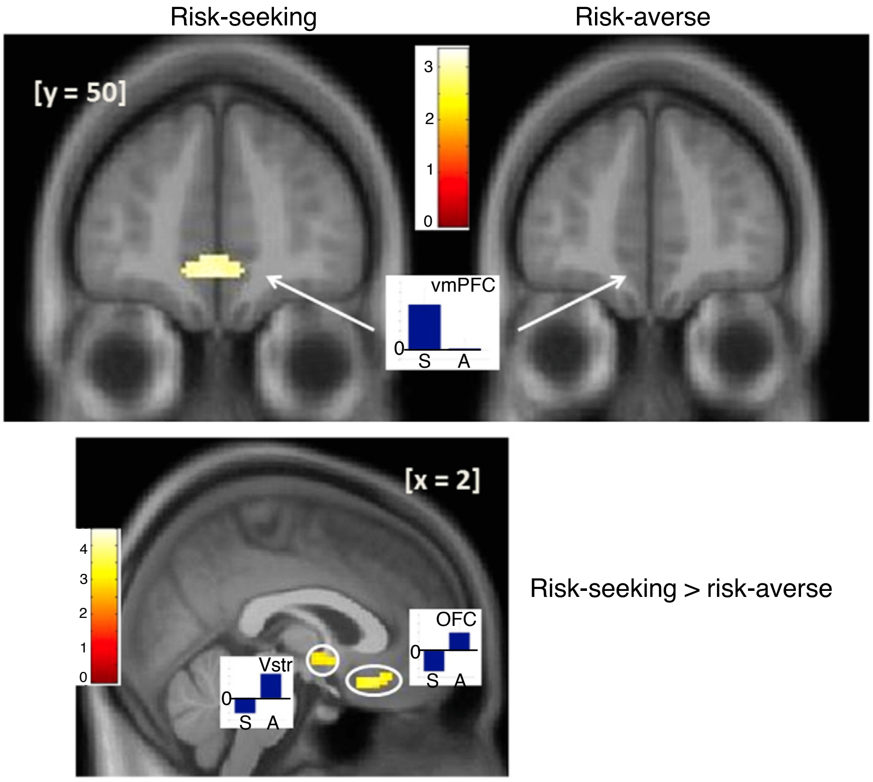

Figure 6.2 Expected value and counterfactual losses in successful learning. The bar plots in blue show the group averaged percent signal change taken from the peak voxel of each cluster. Top: risk-seeking subjects (S) showed a significant modulation in the vmPFC (left) whereas the risk-averse subjects (A) did not. Bottom: risk-seeking and risk-averse subjects demonstrated differential modulation by the counterfactual losses in the ventral striatum (Vstr) and posterior OFC. The risk-seeking subjects demonstrated a negative modulation in both regions, and the risk-averse subjects showed a positive modulation. Note that these results were achieved with regressors derived from the FPE-chosen-and-better-*Q* model.

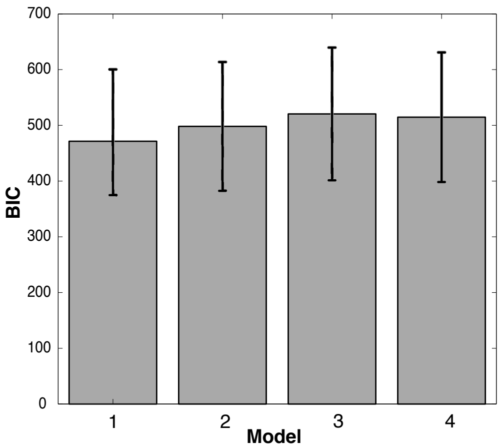

Figure 6.3 Models compared with BIC scores. A smaller BIC indicates a better fit. The models compared are: 1. FPE-*Q* model, 2. FPE-chosen-and-better- *Q* model, 3. FPE-better- *Q* model, 4. FPE-max- *Q* model. Bars indicate mean BIC scores averaged across 30 subjects. Error bars indicate SEM.

6.2.2 Computational modeling

We tested the FPE- Q model as described in the Section 4.3.2.2 and three variations of FPE-Q models according to different hypotheses about how counterfactual outcomes might influence reward-based learning. The FPE- Q model updates the expected value of the chosen state-action $Q(s_t, a_t)$ twice with both equation (4.26) and (4.27). In the three variant models, different subset of possible actions in the visited state is updated at the second stage. If we rewrite Equation (4.27) as:

$$Q(s_t, a^*) \Leftarrow Q(s_t, a^*) + \alpha_{FPE} \, [\underbrace{f_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a^*)}_{\text{fictive TD-error}}], \quad (6.1)$$

a^* denotes different subset of the possible actions. Again, as defined in Section 4.3.2.2 all the possible actions are $a \in \{0,1,2,3\}$, so we have $a^* \subseteq a$. Each model can be described according to a^*:

1. FPE-Q model: the expected value of the chosen action in the current state is updated, that is $a^* = a_t$.

2. FPE-chosen-and-better-Q model: the expected values of the chosen action and the actions that would have yielded better reward are both updated, that is, $a^* = \{a \geq a_t\}$, if $\mathrm{o_t} \geq 0$ and $a^* = \{a \leq a_t\}$, if $o_t < 0$.

3. FPE-better-Q model: the expected values of the actions that would have yielded better reward than the chosen action are updated, that is, $a^* = \{a > a_t\}$, if $o_t \geq 0$ and $a^* = \{a < a_t\}$, if $o_t < 0$.

4. FPE-max-Q model: the expected values of the actions that would have yielded the best reward is updated, that is, $a^* = 3$, if $o_t \geq 0$ and $a^* = 0$ if $o_t < 0$.

We fitted each of the candidate models into the behavioral data using maximum likelihood estimation and compared the goodness of fit with BIC measures. As shown in Figure 6.3, the mean BIC scores averaged across all the 30 subjects suggest that each model explained the behavioral data equally well. There was no significant difference among the BIC scores of each model.

6.3 Results

6.3.1 fMRI data protocol and processing

All MR images were acquired with a Siemens Trio 3T scanner with a 32-channel head coil. Structural MRI were recorded from each participant using a T1 weighted magnetization-prepared rapid gradient-echo sequence with a voxel resolution of 1x1x1 mm^3, coronal orientation, phase-encoding in left-right direction, FoV = 192x256 mm, 240 slices, 1100 ms inversion time, TR = 2300 ms, TE = 2.98 ms, and 90 flip angle. Functional MR time series were recorded using a T2* weighted EPI sequence with TR = 2380 ms, TE = 25 ms, voxel size = 2x2x2 mm^3, FoV = 204x204 mm, skip factor = 0.5, anterior-posterior phase encode, 40 slices acquired in descending order. The acquisition time was approximately 8 minutes per scanning run. The preprocessing was done with the procedure described in Section 3.2.1 and Figure 3.2. In addition, the motion correction is done with the *Spike analyzer.*

We used the regressors derived from the FPE-chosen-and-better-Q model in all our fMRI analysis, because the regressors derived from this model yielded the most robust statistical results. First level analyses included onset regressors for each stimulus event and a set of model-derived parametric modulators generated using each subject's best-fitting parameter (see Figure 6.1 for event phase): the time series of Q-values for the selected action and the choice value of investment (0-3 Euro) at choice phase, factual TD at factual outcome phase, and fictive error signals f_t at counterfactual outcome phase. All regressors were convolved with a canonical hemodynamic response function. Coincident parametric modulators at the same event onset were serially orthogonalized as implemented by default in SPM. For instance, the Q-value regressor was orthogonalized with respect to the choice value regressor. This was done to prevent the first level GLM from allowing variance that was common for both regressors become undetected. In addition, a set of regressors was included for each participant to censor EPI images with large, head movement related spikes in the global mean.

Second level analyses consisted of a one-way analysis of variance (ANOVA). To control for false positives at the group level, 3dClustSim in AFNI (Cox, 1996) was used to determine two different thresholds to apply to cortical and subcortical clusters. The simulation for cortical clusters included all brain voxels (whole-brain correction). The simulation for subcortical clusters (subcortical volume correction) was performed inside a mask (2870 voxels) of the caudate (head, body, tail), nucleus accumbens, and putamen. Both simulations used a single-voxel threshold of $p < 0.005$ and a smoothness of 8 mm^3. Results of the simulation showed that a minimum cluster size of 156 and 32 contiguous voxels yielded a corrected $p < 0.05$ for cortical and subcortical clusters, respectively. These empirically derived thresholds are more conservative with respect to false positive results compared to those recommended by Lieberman and Cunningham (2009), which were chosen to provide an appropriate balance between false alarm and missing true effect for whole-brain corrections. Standardized MNI coordinates are reported with the z-scored peak voxel value (z) and cluster sizes (n).

6.3.2 fMRI results

Firstly, Q-values derived from the FPE-chosen-and-better-Q model modulated the BOLD signal changes in vmPFC ([6, 52, -10], z=3.3, n=960), while the Q-values derived from the standard Q model failed to predict significant BOLD signal changes throughout the entire brain. These results indicate that the counterfactual learning signals were incorporated into the representation of expected value. Secondly, the factual TD error robustly modulated activity of the ventral striatum ([-12, 2, -12], z = 11.9, n=669). Finally, The exploratory subdivision of the subjects into risk-seeking and risk-averse (see behavioral results discussed in Section 4.3.3) yielded interesting effects in the fMRI data that lend themselves to further interpretation of the neural mechanism of valuation processing with counterfactual learning signals. Whereas risk-seeking subjects showed a statistically significant correlation with Q-values in the vmPFC ([0, 50, − 8], z = 2.85, n = 210) at the time of choice, risk-averse subjects showed a very weak representation of expected values indicated by a non-significant correlation with Q-values. In addition, the neural response to the counterfactual losses (i.e., f_+) was significantly different between the risk-seeking and risk-averse subjects. A region of the right ventral striatum ([8, − 4, − 6], z = 3.8, n = 171) and medial OFC ([2, 26, − 16], z = 3.3, n = 171) was negatively modulated by f_+ for counterfactual losses in the risk-seeking subjects, but positively modulated in the risk-averse subjects, shown in Figure 6.2.

6.4 Discussion

The results of this experiment demonstrated that counterfactual learning signals improved Q-learning model fit, and this improved model predicted BOLD signal changes correlated with expected value and reward prediction. Expected value computed from the FPE-chosen-and-better-Q model robustly modulated activity in the vmPFC and OFC. In addition, the FPE-chosen-and-better-Q model showed that risk-seeking and risk-averse subjects differentially utilized counterfactual outcome, and it produced differential correlations with expected value and fictive error signal in the vmPFC/OFC and ventral striatum, respectively.

There is accumulating evidence suggesting that humans indeed incorporate counterfactual consequences into subsequent decisions and that counterfactual consequences modulate neural activity (Boorman et al., 2009; Brassen et al., 2012; Büchel et al., 2011; Coricelli et al., 2005; Li and Daw, 2011; Loomes and Sugden, 1982b; Nicolle et al., 2011). However, none of these studies have incorporated fictive prediction error signals into valuation for strategic decisions that maximize long-term gains despite of interim losses. For example, Li and Daw employed counterfactual outcomes in their study of value-based choices (Li and Daw, 2011). They used a Rescorla-Wagner learning model to estimate expected value, which by definition does not take into account the temporal structure of future anticipated rewards. In addition, neither Lohrenz and colleagues (Lohrenz et al., 2007) nor Chiu and colleagues (Chiu et al., 2008) included the fictive prediction errors from counterfactual gains or losses into their Q-learning model. Instead, these two studies only used a separate regression analysis to determine whether a fictive error signal from counterfactual losses predict changes in subsequent decisions.

Lohrenz and colleagues (Lohrenz et al., 2007) used a temporal-difference prediction error for the factual reward only (as noted in the method section of their manuscript) and is therefore the same as the standard Q-learning model used for comparison in our experiment. In our computational models, we have taken the fictive error signals (computed as the difference between the obtained and unobtained outcomes) as a counterfactual outcome and further computed a fictive temporal-difference error within a two-stage Q learning framework. This additional update with counterfactual outcome helped our models to identify whether fictive error signals contribute to long-term valuation. In this study, the FPE models nested the standard Q-learning model. If participants had not incorporated counterfactual information into their valuation processes, the learning rates for both FPE related parameters, i.e., $\alpha_{FPE} = \{\alpha_{+}, \alpha_{-}\}$, would have been zero. The FPE models are then reduced to a standard Q-learning model and the expected values should not differ among the models. To the contrary, learning rates for both FPE parameters were significantly greater than zero, expected values from FPE models were significantly different from those of the standard Q-learning model, and the FPE models explained choice behavior significantly better, suggesting that participants utilized counterfactual information for valuation and decision-making.

Furthermore, risk-seeking and risk-averse subjects processed the FPE signal differently. Participants who successfully exploited the task to maximize long-term gains demonstrated a different pattern of brain activity compared to the participants who failed to discover or exploit the task. According to Q-learning, participants that were able to exploit the task and select the optimal path (i.e., risk-seeking subjects) did so by maximizing the long-term expected value of their actions. Their representation of expected value was more strongly influenced by counterfactual losses than that of the group of risk-averse subjects. Previously, counterfactual losses have been associated with increased risk taking (Brassen et al., 2012; Büchel et al., 2011), possibly due to the averseness of subjectively experiencing regret at the missed opportunity. In the task design of current study, increased risk taking would lead to the optimal path and hence greater sensitivity to counterfactual losses is indeed advantageous. Previous literature concerning the effects of counterfactual consequences on choice behavior has focused on the interaction of cognitive and emotional effects of counterfactual losses (Sommer et al., 2009). For example, counterfactual losses lead to increased risk taking, and are strongly associated with subjectively experienced regret. Experiencing regret in the face of a missed opportunity is dependent on the structural and functional integrity of the vmPFC (Camille et al., 2004), and adjusting behavior in order to strategically reduce anticipated regret involves activation of the posterior OFC (Coricelli et al., 2005). These differences in subjectively experienced emotions and behavioral biases, suggest that counterfactual gains and losses may contribute differently to valuation.

The differential effects of counterfactual gains and losses may be related to the volatility and risk inherent to the environment. Counterfactual losses may lead to increased riskiness when volatility is low, but may not exert an influence on choice when volatility or risk is high and ambiguous. Counterfactual gains may lead to more conservative choices when volatility and risk are high or unknown, with relatively small effects when volatility and risk are low and unambiguous (Fujiwara et al., 2009; Henderson and Norris, 2013). Risk and volatility were each ambiguous in the current experimental task, and the nature of the environment involved frequent losses. Importantly however, the learning rates associated with the counterfactual learning signals dissociated risk-seeking subjects from risk-averse subjects, with risk-seeking subjects utilizing counterfactual losses more so than risk-averse subjects, who used counterfactual gains more so than risk-seeking subjects. This is consistent with previously reported effects of missed opportunities and regret related choices on subsequent decisions where more optimal decision-making was associated with responsiveness to counterfactual losses specifically (Lohrenz et al., 2007).

The factual TD error term from the standard Q model accounted for BOLD signal changes in the ventral striatum and appeared nearly identical to the modulation effect of fictive prediction error from the FPE-chosen-and-better-Q model. However, risk-seeking subjects demonstrated a significant stronger modulation by fictive prediction error in the ventral striatum. This suggests that the mismatch between responses to factual and counterfactual consequences in an overlapping region of the ventral striatum may be a potential neural mechanism for computing and incorporating counterfactual learning signals during valuation.

The neural representation of expected value during choice behavior is strongly associated with vmPFC activation in humans. Previously, Gläscher and colleagues examined the neural representation of expected values during both action-outcome and stimulus-response learning using a Q-learning model (Gläscher et al., 2009). They found that expected value for both types of choices was significantly correlated with BOLD signal changes in the vmPFC. Consistent with their findings, our study identified a distributed neural system involved in the representation of expected value that was anchored in the vmPFC. The BOLD response from OFC and vmPFC, modulated by Q-values in this study, are often cited as part of a valuation system, however, they are each recently acknowledged as important nodes in a long-term memory system for associative information (Euston et al., 2012; Rushworth et al., 2011). It may be that valuation, decision-making, and episodic memory systems interact or share functional anatomy, which is consistent with the type of processing necessary for learning associations among context, action, events, and consequences. As such, seemingly incompatible models of memory and decision-making may be mutually informative in the development of neurobiological plausible models of large-scale neurocognitive brain function.

In summary, model comparison demonstrated that counterfactual processing occurs during reward-based learning when such information is available. These findings showed that the effects of counterfactual consequences on decision-making can be mediated by a direct effect on action-state values. The results demonstrated that counterfactual learning is an important component of valuation and reward-based learning.

Chapter 7: CONCLUSIONS AND GENERAL DISCUSSION

This thesis has used Markov decision process and reinforcement-learning models to investigate the neural correlates of multiple valuation systems in human decision-making. A key insight is that multiple valuations share a common computational mechanism of reward prediction error.

Chapter 4 provides three new experimental paradigms designed for testing hypotheses formulated by computational models discussed in Chapter 2. The experimental results broadly support the idea that error-correction via reinforcement can be adapted to learn different aspects of reward-related information. Combined with model-based fMRI analysis discussed in Chapter 3, I have explored the brain systems involved in learning, prediction, and decision-making. Chapter 5 and 6 present fMRI experimental results suggesting that the brain regions such as the ventral striatum might encode different types of prediction error signals according to the specific context of learning.

In this final chapter, I briefly discuss some key principles from these different streams of research and interpret them in a general context of perceptual and economic decision-making. The work presented in this thesis also provides a basis for future theoretical and experimental investigations of a number of issues, some of which I now review.

7.1 Multiple prediction errors in human basal ganglia

The main focus of this thesis revolves around the hypothesis that different valuation systems share a common neurobiological mechanism of error-correction via reinforcement. To this end, I have conducted three behavioral and two fMRI studies. Altogether, these results provide evidence showing that the BOLD signals in the ventral striatum correlate with a variety of prediction errors. Here I briefly summarize the main findings of these studies and highlight the advancement that they contributed to our understanding of computations in the human brain.

7.1.1 Summary of behavioral results

I have presented the results from three new reward-based learning paradigms appropriate for investigating factors other than the mean expected reward that may affect decisions. Each paradigm is designed and analyzed on the basis of reinforcement learning.

1. Chapter 4.1 shows that in a two-armed bandit task where a visual stimulus alternates between the left and right sides of the screen according to specific Markov chains, the subjects have to somehow deal with the information provided by the higher-order conditional probabilities to optimize their predictions.

2. For a hierarchical two-armed bandit task where subjects have to learn the reward distributions conditioned on the stimulus likelihood, Chapter 4.2 presents data from 5 experimental conditions, each of which has the stimulus and the reward potentially driving choices in the opposite directions to various degrees. The results demonstrate that both stimulus likelihood and reward probabilities influence decisions. The influence of the stimulus leads the choices toward the option of lower expected values. Chapter 5 replicates these behavioral findings with shorter experimental block of trials and interprets the data as supporting the hypothesis that decisions are dynamically shifted from mainly stimulus-driven to more reward-oriented.

3. Chapter 4.3 presents a strategic sequential investment task, which includes both state-space structure and state-transition rules that the subjects can learn. The task is designed to examine counterfactual learning and risk

sensitivity. The behavioral results suggest that subjects may integrate counterfactual information into the valuation of subsequent choices. Individual choice strategies can be interpreted as different sensitivities to the risk inherent to the task.

7.1.2 Summary of modeling results

This thesis contains two variations on the Rescolar-Wagner model and one variation on the Q model. These new models are developed to make more realistic predictions by addressing perceptual or emotional complexities ignored by standard models. Models are compared according to their ability in explaining the choice behavior.

1. Higher-order prediction model, which adapted Rescolar-Wagner learning to infer either first-order (Equation (4.10)) or second-order (Equation(4.13)) interdependencies in the temporal structure of stimuli. This model quantitatively accounts for higher-order sequential learning, especially when choice behavior are apparently inconsistent with predictions from the existing zero-order model that only relate decisions to expected values.

2. Hybrid model (Equation (4.20)), which combined two Rescolar-Wagner learning models with an exponential decay function. This model embodies a novel hypothesis about the dynamic transition from stimulus-response learning to action-outcome learning. This hypothesis is tested with the experiment designed to dissociate the influences of stimulus and outcome during learning.

3. FPE- Q model (Equation (4.27) and (4.37)), which incorporated an empirically defined fictive error into the Q-learning model. This model incorporates empirical assumptions that counterfactual consequences influence valuation. It establishes parallels between the theories of expected utility and counterfactual learning. Therefore, the model explains risk-sensitive choice behavior, which the previous risk-neural Q model fails.

7.1.3 Summary of fMRI results

1. The neural correlates of reward prediction error in the ventral striatum during both studies presented in Chapter 5 and 6 are consistent with established evidence on the reward-related dopamine system of human basal ganglia.

2. In parallel to the reward prediction error, respective neural correlates of a stimulus prediction error and a fictive prediction error are demonstrated in the ventral striatum. The major conclusion from these two studies is that the ventral striatum correlates with different prediction-error-like signals in either a perceptual or an emotional context. This finding leads to the understanding that the human basal ganglia may use a common prediction-error mechanism to estimate expectations of a wide variety of information.

3. The weighting function derived from the hybrid model correlates with the BOLD signals in the amygdala, suggesting that the amygdala may flexibly weight the interaction between instrumental and goal-directed learning.

4. The expected values derived from the FPE-Q model correlate with BOLD signals in the vmPFC/OFC, suggesting that these regions are involved in processing counterfactual valuation.

In the end, we briefly consider what can be described in terms of multiple prediction errors. The prediction error is a measure of how surprising the observed outcome deviates from the expectation. Such surprising event can essentially drive learning to diminish the prediction error. The *stimulus prediction error* and the *fictive prediction error* can both be interpreted as an *information prediction error*. Seeking information on the environmental stimulus probabilities or the counterfactual consequences can prepare the learning agent for anticipating the correct reward or punishment. The idea that the basal ganglia encode information prediction error leads to other interesting questions: How is the different information distinguished? Do the same neural populations in the basal ganglia compute and distinguish different information prediction error? Or is the information processed into a common signal in other cortical regions with the neural populations encoding prediction error only performing general-purpose computation? Another issue is timing. If a task showing hierarchical structure requires multiple prediction errors that may coincide in time, how are these prediction errors encoded? Are the subgroups of the same neuron population simultaneously assigned to encode one of the prediction errors? Or does the whole neuron population encode all the prediction errors in serial at a finer time scale? Future work may expand on these ideas and re-examine some of the data presented in this thesis.

7.2 Future directions: expectation, value, and attention

Traditional reinforcement learning studies mainly use straightforward two-armed bandit tasks, which only require estimation of the immediate reward. However, temporal extent of the outcome may also affect decision. Consider the task discussed in Chapter 6, subjects have to learn to take a small amount of loss so as to reach a larger amount of gain in the future. Future experiments can be designed to examine the distinction between short-term and long-term predictions.

Rational decision-makers should always choose the option with the higher expected reward. However, under some circumstances humans seem to violate this rational optimization strategy. Consider the task discussed in Chapter 5, the expected reward for different options is determined through an interaction of stimulus probability and reward estimation. Future work may extend on the idea of stimulus-reward interaction and starts to examine whether there is a link between this stimulus expectation and the visual attention. We can replace the stimulus presentation with a perceptual discrimination task. Such an experimental design will require both *perceptual* and *economic* decision-making during the task.

Lastly, a computational framework combining both perceptual uncertainty and reward ultility for characterizing human risk-sensitive choice behavior is still absent. It remains unclear whether the perceptual and economic estimations are converted into a common encoding signal that is used as decision variables. A further direction of research is to integrate the perceptual and economic aspects of decision-making via a POMDP framework and investigate their cooperative or competitive neural mechanisms during learning using the model-based fMRI analysis.

Bibliography

Abler, B., Walter, H., Erk, S., Kammerer, H., and Spitzer, M. (2006). Prediction error as a linear function of reward probability is coded in human nucleus accumbens. Neuroimage *31*, 790–795.

Ahn, W.-Y., Krawitz, A., Kim, W., Busmeyer, J.R., and Brown, J.W. (2011). A Model-Based fMRI Analysis with Hierarchical Bayesian Parameter Estimation. J. Neurosci. Psychol. Econ. *4*, 95–110.

Alink, A., Schwiedrzik, C.M., Kohler, A., Singer, W., and Muckli, L. (2010). Stimulus Predictability Reduces Responses in Primary Visual Cortex. *30*, 2960–2966.

Andersson, J.L., Hutton, C., Ashburner, J., Turner, R., and Friston, K. (2001). Modeling geometric deformations in EPI time series. Neuroimage *13*, 903–919.

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. Neuroimage *38*, 95–113.

Baldi, P., and Itti, L. (2010). Of bits and wows: A Bayesian theory of surprise with applications to attention. Neural Networks *23*, 649–666.

Balleine, B.W. (1992). Instrumental performance following a shift in primary motivation depends on incentive learning. J. Exp. Psychol. Anim. Behav. … *18*, 236–250.

Balleine, B.W., and O'Doherty, J.P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. Neuropsychopharmacology *35*, 48–69.

Balleine, B.W., Daw, N.D., and O'Doherty, J.P. (2009). Neuroeconomics: Multiple forms of value learning and the function of dopamine. In Neuroeconomics, pp. 367–387.

Barto, A.G., Sutton, R.S., and Anderson, C.W. (1983). Neuronlike elements that can solve difficult learning control problems. IEEE Trans. Syst. Man. Cybern. *13*, 835–846.

Baxter, M.G., and Murray, E.A. (2002). The amygdala and reward. Nat. Rev. Neurosci. *3*, 563–573.

Bell, D.E. (1981). Regret in decision making under uncertainty. Oper. Res. *30*, 961–981.

Bellman, R.E. (1957). Dynamic Programming (Princeton Unversity Press).

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. Econometrica *22*, 23–36.

Berridge, K.C. (2012). From prediction error to incentive salience : mesolimbic computation of reward motivation. *35*, 1124–1143.

Bishop, C. (2006). Pattern recognition and machine learning (Springer-Verlag New York).

Boorman, E.D., Behrens, T.E.J., Woolrich, M.W., and Rushworth, M.F.S. (2009). How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action. Neuron *62*, 733–743.

Boorman, E.D., Rushworth, M.F., and Behrens, T.E. (2013). Ventromedial prefrontal and anterior cingulate cortex adopt choice and default reference frames during sequential multi-alternative choice. J. Neurosci. *33*, 2242–2253.

Boorman, L., Kennerley, A.J., Johnston, D., Jones, M., Zheng, Y., Redgrave, P., and Berwick, J. (2010). Negative blood oxygen level dependence in the rat: a model for investigating the role of suppression in neurovascular coupling. J. Neurosci. *30*, 4285–4294.

Bornstein, A.M., and Daw, N.D. (2012). Dissociating hippocampal and striatal contributions to sequential prediction learning. Eur. J. Neurosci. *35*, 1011–1023.

Botvinick, M., Nystrom, L.E., Fissell, K., Carter, C.S., and Cohen, J.D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. Nature *402*, 179–181.

Botvinick, M.M., Cohen, J.D., and Carter, C.S. (2004). Conflict monitoring and anterior cingulate cortex: An update. Trends Cogn. Sci. *8*, 539–546.

Brassen, S., Gamer, M., Peters, J., Gluth, S., and Büchel, C. (2012). Don't Look Back in Anger! Responsiveness to Missed Chances in Successful and Nonsuccessful Aging. Science (80-.). *336*, 612–614.

Bray, S., Rangel, A., Shimojo, S., Balleine, B., Doherty, J.P.O., and O'Doherty, J.P. (2008). The neural mechanisms underlying the influence of pavlovian cues on human decision making. J. Neurosci. *28*, 5861–5866.

Bromberg-martin, E.S., and Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. Neuron *63*, 119–126.

Brooks, S.P., and Gelman, A. (1997). General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Stat. *7*, 434–455.

Büchel, C., Brassen, S., Yacubian, J., Kalisch, R., and Sommer, T. (2011). Ventral striatal signal changes represent missed opportunities and predict future choice. Neuroimage *57*, 1124–1130.

Bugg, J.M., and Crump, M.J.C. (2012). In support of a distinction between voluntary and stimulus-driven control : a review of the literature on proportion congruent effects. *3*, 1–16.

Camille, N., Coricelli, G., Sallet, J., Pradat-Diehl, P., Duhamel, J.-R., and Sirigu, A. (2004). The involvement of the orbitofrontal cortex in the experience of regret. Science (80-.). *304*, 1167–1170.

Chandrasekhar, P.V.S., Capra, C.M., Moore, S., Noussair, C., and Berns, G.S. (2008). Neurobiological regret and rejoice functions for aversive outcomes. Neuroimage *39*, 1472–1484.

Chiu, P.H., Lohrenz, T.M., and Montague, P.R. (2008). Smokers ' brains compute , but ignore , a fictive error signal in a sequential investment task. *11*, 514–520.

Clark, J.J., Hollon, N.G., and Phillips, P.E.M. (2012). Pavlovian valuation systems in learning and decision making. Curr. Opin. Neurobiol. *22*, 1054–1061.

Cleeremans, A., and Dienes, Z. (2008). Computational models of implicit learning. In Cambridge Handbook of Computational Psychology, (Cambridge, UK: Cambridge University Press), pp. 396–421.

Cooper, J.C., Dunne, S., Furey, T., and O'Doherty, J.P. (2012). Human Dorsal Striatum Encodes Prediction Errors during Observational Learning of Instrumental Actions. J. Cogn. Neurosci. *24*, 106–118.

Coricelli, G., Critchley, H.D., Joffily, M., O'Doherty, J.P., Sirigu, A., and Dolan, R.J. (2005). Regret and its avoidance: a neuroimaging study of choice behavior. Nat. Neurosci. *8*, 1255–1262.

Cowles, M.K., and Carlin, B.P.C. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. J. Am. Stat. Assoc. *91*, 883–904.

Cox, R.W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. *29*, 162–173.

Crockett, M.J., Clark, L., Apergis-Schoute, A.M., Morein-Zamir, S., and Robbins, T.W. (2012). Serotonin Modulates the Effects of Pavlovian Aversive Predictions on Response Vigor. Neuropsychopharmacology *37*, 2244–2252.

D'Ardenne, K., McClure, S.M., Nystrom, L.E., and Cohen, J.D. (2008). BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. Science *319*, 1264–1267.

Daltrozzo, J., and Conway, C.M. (2014). Neurocognitive mechanisms of statistical-sequential learning: what do event-related potentials tell us? Front. Hum. Neurosci. *8*, 437.

Daw, N.D. (2011). Trial-by-trial data analysis using computational models. In Decision Making, Affect, and Learning: Attention and Performance XXIII, pp. 1–26.

Daw, N.D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *8*, 1704–1711.

Daw, N.D., Doherty, J.P.O., Dayan, P., Seymour, B., and Dolan, R.J. (2006). Cortical substrates for exploratory decisions in humans. Nature *441*, 876–879.

Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., and Dolan, R.J. (2011). Model-based influences on humans ' choices and striatal prediction errors. Neuron *69*, 1204–1215.

Dayan, P. (2008). The role of value systems in decision making. In Decision Making, the Human Mind, and Implications for Institutions, C. Engel, and W. Singer, eds.

Dayan, P., and Balleine, B.W. (2002). Reward, Motivation,and Reinforcement Learning. Neuron *36*, 285–298.

Dayan, P., and Berridge, K.C. (2014). Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. Cogn. Affect. Behav. Neurosci.

Dayan, P., and Daw, N.D. (2008). Decision theory, reinforcement learning, and the brain. Cogn. Affect. Behav. Neurosci. *8*, 429–453.

Dayan, P., and Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. Curr. Opin. Neurobiol. *18*, 185–196.

Deichmann, R., Gottfried, J.., Hutton, C., and Turner, R. (2003). Optimized EPI for fMRI studies of the orbitofrontal cortex. Neuroimage *19*, 430–441.

Delgado, M.R., Miller, M.M., Inati, S., and Phelps, E.A. (2005). An fMRI study of reward-related probability learning. Neuroimage *24*, 862–873.

Dickinson, A., and Balleine, B. (2002). The Role of Learning in the Operation of Motivational Systems. In Stevens' Handbook of Experimental …, pp. 497–534.

Ding, L., and Gold, J.I. (2013). The basal ganglia ' s contributions to perceptual decision making. Neuron *79*, 640–649.

Diuk, C., Tsai, K., Wallis, J., Botvinick, M., and Niv, Y. (2013). Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia. J. Neurosci. *33*, 5797–5805.

Dolan, R., and Dayan, P. (2013). Goals and habits in the brain. Neuron *80*, 312–325.

Doll, B.B., Simon, D. a, and Daw, N.D. (2012). The ubiquity of model-based reinforcement learning. Curr. Opin. Neurobiol. *22*, 1075–1081.

Doya, K., Ito, M., and Samejima, K. (2011). Model-based analysis of decision variables. In Decision Making, Affect, and Learning: Attention and Performance XXIII, (Oxford University Press), p. 190.

Epstude, K., and Roese, N.J. (2008). The functional theory of counterfactual thinking. Pers. Soc. Psychol. Rev. *12*, 168–192.

Estes, W.K. (1948). Discriminative Conditioning. II. Effects of a Pavlovian Conditioned Stimulus upon a Subsequently Established Operant Response. J. Exp. Psychol. *38*, 173.

Euston, D.R., Gruber, A.J., and McNaughton, B.L. (2012). The Role of Medial Prefrontal Cortex in Memory and Decision Making. Neuron *76*, 1057–1070.

Fearnley, J.M., and Lees, A.J. (1991). Ageing and Parkinson's disease: substantia nigra regional selectivity. Brain *114*, 2283–2301.

Fiorillo, C.D., Tobler, P.N., and Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. Science (80-.). *299*, 1898–1902.

Forman, S.D., Cohen, J.D., Fitzgerald, M., Eddy, W.F., Mintun, M.A., and Noll, D.C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. Magn. Reson. Med. *33*, 636–647.

Fujiwara, J., Tobler, P.N., Taira, M., Iijima, T., and Tsutsui, K.I. (2009). A parametric relief signal in human ventrolateral prefrontal cortex. Neuroimage *44*, 1163–1170.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2003). Bayesian Data Analysis, 2nd ed. (CRC Press).

Gershman, S.J., Pesaran, B., and Daw, N.D. (2009). Human Reinforcement Learning Subdivides Structured Action Spaces by Learning Effector-Specific Values. *29*, 13524–13531.

Geurts, D.E.M., Huys, Q.J.M., den Ouden, H.E.M., and Cools, R. (2013). Serotonin and aversive Pavlovian control of instrumental behavior in humans. J. Neurosci. *33*, 18932–18939.

Gläscher, J. (2009). Visualization of group inference data in functional neuroimaging. Neuroinformatics *7*, 73–82.

Gläscher, J., and O’Doherty, J.P. (2010). Model-based approaches to neuroimaging: combining reinforcement learning theory with fMRI data. Wiley Interdiscip. Rev. Cogn. Sci. *1*, 501–510.

Gläscher, J., Hampton, A.N., and O’Doherty, J.P. (2009). Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. Cereb. Cortex *19*, 483–495.

Gläscher, J., Daw, N., Dayan, P., and O’Doherty, J.P.O. (2010). States versus rewards : dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron *66*, 585–595.

Glimcher, P. (2010). Foundations of neuroeconomic analysis (Oxford University Press, London).

Glimcher, P.W., and Fehr, E. (2014). Neuroeconomics: decision making and the brain, 2nd edition (Academic Press).

Gold, J.I., and Shadlen, M.N. (2002). Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. Neuron *36*, 299–308.

Gold, J.I., and Shadlen, M.N. (2007). The neural basis of decision making. Annu. Rev. Neurosci. 535–574.

Gonzalez-Castillo, J., Saad, Z.S., Handwerker, D. a, Inati, S.J., Brenowitz, N., and Bandettini, P. a (2012). Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis. Proc. Natl. Acad. Sci. U. S. A. *109*, 5487–5492.

Graybiel, A.M., Aosaki, T., Flaherty, A.W., and Kimura, M. (1994). The basal ganglia and adaptive motor control. Science (80-.). *265*, 1826–1831.

Haber, S.N., and Knutson, B. (2010). The reward circuit: linking primate anatomy and human imaging. Neuropsychopharmacology *35*, 4–26.

Hampton, A.N., Bossaerts, P., and O’Doherty, J.P. (2006). The Role of the Ventromedial Prefrontal Cortex in Abstract State-Based Inference during Decision Making in Humans. J. Neurosci. *26*, 8360–8367.

Hare, T.A., Doherty, J.O., Camerer, C.F., Schultz, W., and Rangel, A. (2008). Dissociating the Role of the Orbitofrontal Cortex and the Striatum in the Computation of Goal Values and Prediction Errors. *28*, 5623–5630.

Hart, S., and Mas-Colell, A. (2003). Regret-based continuous-time dynamics. Games Econ. Behav. *45*, 375–394.

Hebart, M.N., and Gläscher, J. (2014). Serotonin and dopamine differentially affect appetitive and aversive general Pavlovian-to-instrumental transfer. Psychopharmacology (Berl).

Hebb, D. (1949). The organization of behavior: a neuropsychologcial theory (New York Wiley-Interscience).

Heeger, D.J., Huk, A.C., Geisler, W.S., and Albrecht, D.G. (2000). Spikes versus BOLD: what does neuroimaging tell us about neuronal activity? Nat. Neurosci. *3*, 631–633.

Henderson, S.E., and Norris, C.J. (2013). Counterfactual thinking and reward processing: An fMRI study of responses to gamble outcomes. Neuroimage *64*, 582–589.

Holland, P.C., and Gallagher, M. (2004). Amygdala – frontal interactions and reward expectancy. Curr. Opin. Neurobiol.

Howard, R.A. (1960). Dynamic Programming and Markov Processes (The MIT Press).

Huelsenbeck, J.P., and Crandall, K. a (1997). Phylogeny estimation and hypothesis testing using maximum likelihood. Annu. Rev. Ecol. Syst. *28*, 437–466.

Huettel, S. a., and McCarthy, G. (2004). What is odd in the oddball task? Prefrontal cortex is activated by dynamic changes in response strategy. Neuropsychologia *42*, 379–386.

Huettel, S.A., Mack, P.B., and Mccarthy, G. (2002). Perceiving patterns in random series : dynamic processing of sequence in prefrontal cortex. Nat. Neurosci. *5*, 485–490.

Huettel, S.A., Song, A.W., and McCarthy, G. (2005). Decisions under uncertainty: probabilistic context influences activation of prefrontal and parietal cortices. J. Neurosci. *25*, 3304–3311.

Hutton, C., Bork, A., Josephs, O., Deichmann, R., Ashburner, J., and Turner, R. (2002). Image distortion correction in fMRI: A quantitative evaluation. Neuroimage *16*, 217–240.

Huys, Q.J.M., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R.J., and Dayan, P. (2011). Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. PLoS Comput. Biol. *7*.

Ito, M., and Doya, K. (2011). Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. Curr. Opin. Neurobiol. *21*, 368–373.

Ivry, R., and Knight, R.T. (2002). Making order from chaos: the misguided frontal lobe. Nat. Neurosci. *5*, 394–396.

Jezzard, P., and Balaban, R.S. (1995). Correction for geometric distortion in echo planar images from B0 field variations. Magn. Reson. Med. *34*, 65–73.

Joel, D., Niv, Y., and Ruppin, E. (2002). Actor-critic models of the basal ganglia: New anatommical and computational perspective. Neural Networks *15*, 535–547.

Jongsma, M.L. a, Eichele, T., Van Rijn, C.M., Coenen, A.M.L., Hugdahl, K., Nordby, H., and Quiroga, R.Q. (2006). Tracking pattern learning with single-trial event-related potentials. Clin. Neurophysiol. *117*, 1957–1973.

Kaelbling, L., Littman, M., and Cassandra, A. (1998). Planning and acting in partially observable stochastic domains. Artif. Intell. *101*, 99–134.

Kim, H., Shimojo, S., and O'Doherty, J.P. (2006). Is avoiding an aversive outcome rewarding? Neural substrates of avoidance learning in the human brain. PLoS Biol. *4*, 1453–1461.

King, J.A., Korb, F.M., and Egner, T. (2012). Priming of Control: Implicit Contextual Cuing of Top-down Attentional Set. J. Neurosci. *32*, 8192–8200.

Kirino, E., Belger, a, Goldman-Rakic, P., and McCarthy, G. (2000). Prefrontal activation evoked by infrequent target and novel stimuli in a visual target detection task: an event-related functional magnetic resonance imaging study. J. Neurosci. *20*, 6612–6618.

Klein, S.A. (2001). Measuring , estimating , and understanding the psychometric function : A commentary. *63*, 1421–1455.

Knutson, B., and Gibbs, S.E. (2007). Linking nucleus accumbens dopamine and blood oxygenation. Psychopharmacology (Berl). *191*, 813–822.

Kobayashi, S., and Schultz, W. (2014). Reward contexts extend dopamine signals to unrewarded stimuli. Curr. Biol. *24*, 56–62.

Konda, V., and Tsitsiklis, J. (2003). Actor-critic algorithms. SIAM J. Control Optim. *42*, 1143–1166.

Kruschke, J.K. (2010). Doing Bayesian data analysis: A tutorial with R and BUGS (Academic Press).

Leanne, T., Hogarth, L., and Theodora, D. (2011). Prediction and uncertainty in human Pavlovian to instrumental transfer. J. Exp. Psychol. Learn. Mem. Cogn. *37*, 757–765.

Lee, M.D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. J. Math. Psychol. *55*, 1–7.

Lee, S.W., Shimojo, S., and O'Doherty, J.P. (2014). Neural computations underlying arbitration between model-based and model-free learning. Neuron *81*, 687–699.

Levy, D.J., and Glimcher, P.W. (2012). The root of all value: a neural common currency for choice. Curr. Opin. Neurobiol. *22*, 1027–1038.

Li, J., and Daw, N.D. (2011). Signals in human striatum are appropriate for policy update rather than value prediction. J. Neurosci. *31*, 5504–5511.

Li, J., Schiller, D., Schoenbaum, G., Phelps, E.A., and Daw, N.D. (2011). Differential roles of human striatum and amygdala in associative learning. Nat. Neurosci. *14*, 1250–1252.

Lieberman, M.D., and Cunningham, W. a. (2009). Type I and Type II error concerns in fMRI research: Re-balancing the scale. Soc. Cogn. Affect. Neurosci. *4*, 423–428.

Litt, A., Plassmann, H., Shiv, B., and Rangel, A. (2011). Dissociating valuation and saliency signals during decision-making. Cereb. Cortex 95–102.

Liu, X., Powell, D.K., Wang, H., Gold, B.T., Corbly, C.R., and Joseph, J.E. (2007). Functional dissociation in frontal and striatal areas for processing of positive and negative reward information. J. Neurosci. *27*, 4587–4597.

Logothetis, N.K. (2008). What we can do and what we cannot do with fMRI. Nature *453*, 869–878.

Logothetis, N.K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. Nature *412*, 150–157.

Lohrenz, T., McCabe, K., Camerer, C.F., and Montague, P.R. (2007). Neural signature of fictive learning signals in a sequential investment task. Proc. Natl. Acad. Sci. U. S. A. *104*, 9493–9498.

Loomes, G., and Sugden, R. (1982a). Regret theory: an alternative of rational choice under uncertainty. Econ. J. *92*, 805–824.

Loomes, G., and Sugden, R. (1982b). Regret Theory: An Alternative of Rational Choice Under Uncertainty. Econ. J. *92*, 805–824.

Lungu, O. V., Wächter, T., Liu, T., Willingham, D.T., and Ashe, J. (2004). Probability detection mechanisms and motor learning. Exp. Brain Res. *159*, 135–150.

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2012). The BUGS Book: A practical introduction to Bayesian analysis.

Madden, D.J., Whiting, W.L., Provenzale, J.M., and Huettel, S. a. (2004). Age-related Changes in Neural Activity during Visual Target Detection Measured by fMRI. Cereb. Cortex *14*, 143–155.

Magri, C., Schridde, U., Murayama, Y., Panzeri, S., and Logothetis, N.K. (2012). The amplitude and timing of the BOLD signal reflects the relationship between local field potential power at different frequencies. J. Neurosci. *32*, 1395–1407.

Maier, A., Wilke, M., Aura, C., Zhu, C., Ye, F.Q., and Leopold, D. a (2008). Divergence of fMRI and neural signals in V1 during perceptual suppression in the awake monkey. Nat. Neurosci. *11*, 1193–1200.

Marchiori, D., and Warglien, M. (2008). Predicting human interactive learning by regret-driven neural networks. Science *319*, 1111–1113.

McClure, S.M., Berns, G.S., and Montague, P.R. (2003). Temporal Prediction Errors in a Passive Learning Task Activate Human Striatum. Neuron *38*, 339–346.

Mcclure, S.M., Daw, N.D., and Montague, P.R. (2003). A computational substrate for incentive salience. Trends Neurosci. *26*, 423–428.

Miller, M.B., Valsangkar-Smyth, M., Newman, S., Dumont, H., and Wolford, G. (2005). Brain activations associated with probability matching. Neuropsychologia *43*, 1598–1608.

Montague, P.R., Dayan, P., and Sejnowsk, T.J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. J. Neurosci. *16*, 1936–1947.

Montague, P.R., Hyman, S.E., and Cohen, J.D. (2004). Computational roles for dopamine in behavioural control. Nature *431*, 760–767.

Montague, P.R., Dolan, R.J., Friston, K.J., and Dayan, P. (2012). Computational psychiatry. Trends Cogn. Sci. *16*, 72–80.

Mukamel, R., Gelbard, H., Arieli, A., Hasson, U., Fried, I., and Malach, R. (2005). Coupling between neuronal firing, field potentials, and FMRI in human auditory cortex. Science *309*, 951–954.

Neyman, J., and Pearson, E.S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. *231*, 289–337.

Nicolle, A., Fleming, S.M., Bach, D.R., Driver, J., and Dolan, R.J. (2011). A regret-induced status quo bias. J. Neurosci. *31*, 3320–3327.

Nilsson, H., Rieskamp, J., and Wagenmakers, E.-J. (2011). Hierarchical Bayesian parameter estimation for cumulative prospect theory. J. Math. Psychol. 55, 84–93.

Niv, Y., and Schoenbaum, G. (2008). Dialogues on prediction errors. Trends Cogn. Sci.

Niv, Y., Edlund, J. a, Dayan, P., and O'Doherty, J.P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. J. Neurosci. *32*, 551–562.

Nomoto, K., Schultz, W., Watanabe, T., and Sakagami, M. (2010). Temporally extended dopamine responses to perceptually demanding reward-predictive stimuli. J. Neurosci. *30*, 10692–10702.

O'Doherty, J.P. (2004). Reward representations and reward-related learning in the human brain: insights from neuroimaging. Curr. Opin. Neurobiol. *14*, 769–776.

O’Doherty, J.P., Dayan, P., Friston, K., Critchley, H., and Dolan, R.J. (2003). Temporal difference models and reward-related learning in the human brain. Neuron *28*, 329–337.

O’Doherty, J.P., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R.J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science (80-.). *304*, 452–454.

Olds, J. (1958). Self-stimulation of the brain. Science (80-.). *127*.

Ostwald, D., Spitzer, B., Guggenmos, M., Schmidt, T.T., Kiebel, S.J., and Blankenburg, F. (2012). Evidence for neural encoding of Bayesian surprise in human somatosensation. Neuroimage *62*, 177–188.

Paulus, M.P., Feinstein, J.S., Tapert, S.F., and Liu, T.T. (2004). Trend detection via temporal difference model predicts inferior prefrontal cortex activation during acquisition of advantageous action selection. Neuroimage *21*, 733–743.

Pavlov, I.P. (1927). Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex. Oxford Univ. Press.

Perfors, A., Tenenbaum, J.B., Griffiths, T.L., and Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. Cognition *120*, 302–321.

Pessiglione, M., Petrovic, P., Daunizeau, J., Palminteri, S., Dolan, R.J., and Frith, C.D. (2008). Subliminal Instrumental Conditioning Demonstrated in the Human Brain. Neuron *59*, 561–567.

Petersen, S.E., and Posner, M.I. (2012). The Attention System of the Human Brain: 20 Years After. Annu. Rev. Neurosci. *35*, 73–89.

Pieters, R., and Zeelenberg, M. (2007). A Theory of Regret Regulation 1.0. J. Consum. Psychol. *17*, 29–35.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Proc. 3rd Int. Work. Distrib. Stat. Comput.

Poldrack, R.A., Mumford, J.A., and Nichols, T.E. (2011). Handbook of functional MRI data analysis (Cambridge university press).

Prévost, C., Mccabe, J. a., Jessup, R.K., Bossaerts, P., and O'Doherty, J.P. (2011). Differentiable contributions of human amygdalar subregions in the computations underlying reward and avoidance learning. Eur. J. Neurosci. *34*, 134–145.

Prévost, C., Liljeholm, M., Tyszka, J.M., and O'Doherty, J.P. (2012). Neural correlates of specific and general Pavlovian-to-Instrumental Transfer within human amygdalar subregions: a high-resolution fMRI study. J. Neurosci. *32*, 8383–8390.

Prévost, C., Mcnamee, D., Jessup, R.K., Bossaerts, P., and O'Doherty, J.P. (2013). Evidence for model-based computations in the human amygdala during Pavlovian conditioning. Plos Comput. Biol. *9*.

Puterman, M.L. (1994). Markov Decision Processes: Discrete Stochastic Dynamic Programming (New York: John Wiley & Sons, Inc.).

Ranganath, C., and Rainer, G. (2003). Neural mechanisms for detecting and remembering novel events. Nat. Rev. Neurosci. *4*, 193–202.

Rangel, A., Camerer, C., and Montague, P.R. (2008). A framework for studying the neurobiology of value-based decision making. Nat. Rev. Neurosci. *9*, 545–556.

Rao, R.P.N. (2010). Decision making under uncertainty: a neural model based on partially observable markov decision processes. Front. Comput. Neurosci. *4*, 146.

Rescorla, R.A. (1987). Pavlovian Conditioning. Am. Psychol. *43*, 151–160.

Rescorla, R., and Wagner, A. (1972). A theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. In Classical Conditioning II: Current Research and Theory, (New York), pp. 64–99.

Roesch, M.R., Calu, D.J., Esber, G.R., and Schoenbaum, G. (2010). Neural correlates of variations in event processing during learning in basolateral amygdala. J. Neurosci. *30*, 2464–2471.

Roese, N.J., Park, S., Smallman, R., and Gibson, C. (2008). Schizophrenia involves impairment in the activation of intentions by counterfactual thinking. Schizophr. Res. *103*, 343–344.

Rushworth, M.F.S., Noonan, M.P., Boorman, E.D., Walton, M.E., and Behrens, T.E. (2011). Review Frontal Cortex and Reward-Guided Learning and Decision-Making. Neuron *70*, 1054–1069.

Schönberg, T., Daw, N.D., Joel, D., and Doherty, J.P.O. (2007). Reinforcement Learning Signals in the Human Striatum Distinguish Learners from Nonlearners during Reward-. *27*, 12860–12867.

Schott, B.H., Minuzzi, L., Krebs, R.M., Elmenhorst, D., Lang, M., Winz, O.H., Seidenbecher, C.I., Coenen, H.H., Heinze, H., Zilles, K., et al. (2008). Mesolimbic functional magnetic resonance imaging activations during reward anticipation correlate with reward-related ventral striatal dopamine release. J. Neurosci. *28*, 14311–14319.

Schultz, W. (1998). Predictive Reward Signal of Dopamine Neurons. J. Neurophysiol. *80*, 1–27.

Schultz, W. (2013). Updating dopamine reward signals. Curr. Opin. Neurobiol. *23*, 229–238.

Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. Science (80-.). *275*, 1593–1599.

Schwarz, G. (1978). Estimating the dimension of a model. Ann. Stat. *6*, 461–464.

Seymour, B., and Dolan, R. (2008). Emotion, decision making, and the amygdala. Neuron *58*, 662–671.

Seymour, B., Doherty, J.P.O., Dayan, P., Koltzenburg, M., Jones, A.K., Dolan, R.J., Friston, K.J., and Frackowiak, R.S. (2004). Temporal difference models describe higher-order learning in humans. *429*, 664–667.

Sharma, J., Dragoi, V., Tenenbaum, J.B., Miller, E.K., and Sur, M. (2003). V1 neurons signal acquisition of an internal representation of stimulus location. Science *300*, 1758–1763.

Shen, Y., Tobia, M.J., Sommer, T., and Obermayer, K. (2014). Risk sensitive reinforcement learning. Neural Comput. *26*, 7.

Shiffrin, R.M., Lee, M.D., Kim, W., and Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical bayesian methods. Cogn. Sci. *32*, 1248–1284.

Simon, D.A., and Daw, N.D. (2011). Neural correlates of forward planning in a spatial decision task in humans. J. Neurosci. *31*, 5526–5539.

Singh, K.D. (2012). Which “neural activity” do you mean? fMRI, MEG, oscillations and neurotransmitters. Neuroimage *62*, 1121–1130.

Sirotin, Y.B., and Das, A. (2009). Anticipatory haemodynamic signals in sensory cortex not predicted by local neuronal activity. Nature *457*, 475–479.

Smith, K.S., Berridge, K.C., and Aldridge, J.W. (2011). Disentangling pleasure from incentive salience and learning signals in brain reward circuitry. *108*.

Sommer, T., Peters, J., Gläscher, J., and Büchel, C. (2009). Structure–function relationships in the processing of regret in the orbitofrontal cortex. Brain Struct. Funct. *213*, 535–551.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. J. R. Stat. Soc. Ser. B (Statistical Methodol. *64*, 583–639.

Squires, K., Wickens, C., Squires, N., and Donchin, E. (1976). The effect of stimulus sequence on the waveform of the cortical event-related potential. Science (80-.). *193*, 1142–1146.

Stadler, W., Klimesch, W., Pouthas, V., and Ragot, R. (2006). Differential effects of the stimulus sequence on CNV and P300. Brain Res. *1123*, 157–167.

Sutton, R.S., and Barto, A.G. (1998). Reinforcement Learning: An Introduction (Cambridge, MA: MIT Press).

Talmi, D., Seymour, B., Dayan, P., and Dolan, R.J. (2008). Human pavlovian-instrumental transfer. J. Neurosci. *28*, 360–368.

Team, S.D. (2014). Stan modeling language users guide and reference manual, Version 2.5.0.

Thorndike, E.L. (1933). A proof of the Law of Effect. Science (80-.). *77*, 173–175.

Tindell, A.J., Smith, K.S., Berridge, K.C., and Aldridge, J.W. (2009). Dynamic Computation of Incentive Salience : “ Wanting ” What Was Never “ Liked .” *29*, 12220–12228.

Tobia, M.J., Guo, R., Schwarze, U., Boehmer, W., Gläscher, J., Finckh, B., Marschner, A., Büchel, C., Obermayer, K., and Sommer, T. (2014). Neural systems for choice and valuation with counterfactual learning signals. Neuroimage *89*, 57–69.

Tolman, E.C. (1948). Cognitive maps in rats and men. Psychol. Rev. *55*, 189–208.

Turk-browne, N.B., Scholl, B.J., Chun, M.M., and Johnson, M.K. (2009). Neural evidence of statisticallLearning: Efficient detection of visual regularities without awareness. J. Cogn. Neurosci. *21*, 1934–1945.

Tversky, A., and Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. Science (80-.).

Valentin, V. V, Dickinson, A., and O'Doherty, J.P. (2007). Determining the neural substrates of goal-directed learning in the human brain. J. Neurosci. *27*, 4019–4026.

Wang, X.-J., and Krystal, J.H. (2014). Computational Psychiatry. Neuron *84*, 638–654.

Watkins, C.J.C.H., and Dayan, P. (1992). Q-learning. Mach. Learn. Spec. Issue Reinf. Learn. *8*.

Whalen, P.J., and Phelps, E.A. (2009). The Human Amygdala (New York: Guilford Press).

Wilks, S.S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Stat. *9*, 60–62.

Wise, R.A., and Rompre, P.-P. (1989). Brain dopamine and reward. Annu. Rev. Psychol. *40*, 191–225.

Wit, S. de, Watson, P., Harsay, H. a., Cohen, M.X., van de Vijver, I., and Ridderinkhof, K.R. (2012). Corticostriatal Connectivity Underlies Individual Differences in the Balance between Habitual and Goal-Directed Action Control. J. Neurosci. *32*, 12066–12075.

Wunderlich, K., Smittenaar, P., and Dolan, R.J. (2012). Dopamine enhances model-based over model-free choice behavior. Neuron *75*, 418–424.

Zeelenberg, M., Beattie, J., Pligt, J. Van Der, and Vries, N.K. de (1996). Consequences of regret aversion: effects of expected feedback on risky decision making. Organ. Behav. Hum. Decis. Process. *65*, 148–158.

Zhang, J., and Rowe, J.B. (2015). The neural signature of information regularity in temporally extended event sequences. Neuroimage *107*, 266–276.

Zink, C.F., Pagnoni, G., Martin, M.E., Dhamala, M., and Berns, G.S. (2003). Human striatal response to salient nonrewarding stimuli. J. Neurosci. *23*, 8092–8097.

APPENDICES

APPENDIX 1 PREPUBLICATION

Chapter 6 (page 107-118) of the dissertation were included in the publication:

M. J. Tobia*, R. Guo*, U. Schwarze, W. Boehmer, J. Gläscher, B. Finckh, A. Marschner, C. Büchel, K. Obermayer, and T. Sommer, "Neural systems for choice and valuation with counterfactual learning signals.," Neuroimage, vol. 89, pp. 57–69, Apr. 2014. (*equally contributed) Copyright © 2013 Elsevier Inc.

The work of this publication was part of the interdisciplinary consortium "Bernstein Fokus: Neuronale Grundlagen des Lernens", which centered on theoretical and experimental collaborations. I was funded by this project as Wissenschaftliche Mitarbeiterin during my PhD at the group of Prof. Klaus Obermayer.

In this paper, I hypothesized that the fictive prediction error can be incorporated into reinforcement-learning models and contribute to the computation of expected values. To test this hypothesis, I constructed computational models and analyzed the behavioral data based on the models. More comprehensive modeling and behavioral analysis were presented in this thesis (e.g., Page 79-85, 113), which *were not* included in the publication. M.J. Tobia analyzed the fMRI data using my modeling results in terms of model-based fMRI analysis (Figure 6.2 on Page 112).

Details of each author's contribution to the publication:

M.J. Tobia and R. Guo performed the research, analyzed the data and wrote the paper.

U. Schwarze, B. Finckh, A. Marschner contributed in fMRI data collection and the dietary depletion of either tryptophan or tyrosine/phenylalanine to manipulate serotonin (5HT) and dopamine (DA), respectively. The results related to depletion manipulation in the publication *were not* included in the dissertation.

W. Boehmer, J. Gläscher, C. Büchel, K. Obermayer, T. Sommer designed the research and contributed reagents and analytic tools.

APPENDIX 2 EIDESSTATTLICHE ERKLÄRUNG

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle Ausführungen, die anderen veröffentlichten oder nicht veröffentlichten Schriften wörtlich oder sinngemäss entnommen wurden, habe ich kenntlich gemacht.

Die Arbeit hat in gleicher oder ähnlicher Fassung noch keiner anderen Prüfungsbehörde vorgelegen.

STATEMENT OF AUTHORSHIP

This dissertation is the result of my own work and includes nothing, which is the outcome of work done in collaboration except where specifically indicated in the text. It has not been previously submitted, in part or whole, to any university of institution for any degree, diploma, or other qualification.

Rong Guo

May 2015, Berlin