A Bayesian Model of Cognitive Control

by

Jiefeng Jiang

Department of Psychology and Neuroscience
Duke University

Date:_____

Approved:

_____
Tobias Egner, Supervisor

_____
Katherine Heller

_____
Rick Hoyle

_____
Scott Huettel

_____
Michael Platt

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor
of Philosophy in the Department of
Psychology and Neuroscience in the Graduate School
of Duke University

2014

ABSTRACT

A Bayesian Model of Cognitive Control

by

Jiefeng Jiang

Department of Psychology and Neuroscience
Duke University

Date:_____

Approved:

_____
Tobias Egner, Supervisor

_____
Katherine Heller

_____
Rick Hoyle

_____
Scott Huettel

_____
Michael Platt

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Psychology and Neuroscience in the Graduate School of Duke University

2014

Abstract

"Cognitive control" describes endogenous guidance of behavior in situations where routine stimulus-response associations are suboptimal for achieving a desired goal. The computational and neural mechanisms underlying this capacity remain poorly understood. The present dissertation examines recent advances stemming from the application of a statistical, Bayesian learner perspective on control processes. An important limitation in current models consists of a lack of a plausible mechanism for the flexible adjustment of control over variable environments. I propose that flexible cognitive control can be achieved by a Bayesian model with a self-adapting, volatility-driven learning scheme, which modulates dynamically the relative dependence on recent (short-term) and remote (long-term) experiences in its prediction of future control demand. Using simulation data, human behavioral data and human brain imaging data, I demonstrate that this Bayesian model does not only account for several classic behavioral phenomena observed from the cognitive control literature, but also facilitates a principled, model-guided investigation of the neural substrates underlying the flexible adjustment of cognitive control. Based on the results, I conclude that the proposed Bayesian model provides a feasible solution for modeling the flexible adjustment of cognitive control.

Dedication

To *Xiaoqiao*.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to sincerely thank my colleagues, friends and family, without whom this dissertation, and my career in cognitive neuroscience would not have been possible. To Tobias Egner, for being the best advisor. To Nestor Schmajuk, for teaching me the basics of modeling. To Katherine Heller, for sharing her expertise in Bayesian methods. To Tim Behrens, for sharing his programs of the Bayesian model. To Rick Hoyle, Scott Huettel, Michael Platt and Chris Summerfield, for insightful comments and challenging questions that keep me thinking and substantially improve the quality of this dissertation. To Franziska Korb, Joseph King, Darinka Trubutschek, Anastasia Kiyonaga, Hanna Oh, Yu-Chin Chui and Amelia Abbott-Frey, for helpful discussions.

To Wen Jiang, Shenduo Li, Taoran Li, Jing Tao, Ziyu Zhang, Linghan Zhu, Tan Zi and many other friends, for the great times we spent together and laughters we share. To Xiaoqiao and my parents, for all of their love, support and inspiration. And finally to my grandma, may your memory last forever.

1. Cognitive Control as Statistical Inference

"Cognitive control" describes the ability to guide one's behavior and mental states in line with internal goals. A key characteristic of cognitive control is thought to be flexibility: control processes must be capable of dynamically adapting (both qualitatively and quantitatively) to ongoing changes in the environment. How this type of contextual regulation of control occurs (in the absence of an all-knowing homunculus) is a key question in current cognitive psychology and neuroscience research. This dissertation attempts to address this question using a Bayesian approach, which behaves similar to the well-known reinforcement learning algorithms with a flexible learning rate (see below). Behavioral and brain imaging studies reported in this dissertation suggest that this Bayesian approach is not only able to simultaneously account for some key behavioral phenomena in the field of cognitive control, but also capable of guiding research to discover neural substrates underlying the flexibility in cognitive control. The structure of this dissertation is organized as follows: In chapter 1, I first review recent efforts in modeling the control processes using both Bayesian and non-Bayesian approaches. Based on the reviews, I then propose that Bayesian methods form a potent solution to model flexible cognitive control. Chapter 2 starts with detailed description of a novel Bayesian model for simulating flexible cognitive control and an experimental design that promotes flexible adjustment of cognitive control. The description is then followed by some proof-of-principle validations of the model using computer-generated

data based on the experimental design. Chapter 3 documents 3 behavioral studies to further validate the Bayesian model's potential of accounting for several key behavioral phenomena. In chapter 4, I present a functional magnetic resonance imaging (fMRI) study that employs this Bayesian model to explore the neural basis underlying flexible cognitive control. In chapter 5, the key findings in this dissertation are summarized. Limitation and directions of future research are also discussed.

1.1 Cognitive control as 'guided' information processing

In interacting with our environment, we transform sensory input into internal representations and select cognitive or motor actions based on these representations and our current goals. Given the fact that there is an enormous amount of sensory information and many possible actions available in contrast to only a few desired responses, appropriate action selection is a difficult task. To simplify this task, stimuli and actions that are frequently paired become mnemonically associated (e.g., via Hebbian learning) into stimulus-response (S-R) ensembles (or pathways) or more complex and extended action schemas (Norman & Shallice, 1986) that facilitate prompt reaction. Because much sensory information is processed in different pathways in parallel but only few actions can (or should) be taken simultaneously, stimulus representations and S-R pathways are believed to compete for being selected to drive behavior (Desimone & Duncan, 1995; Miller & Cohen, 2001; Norman & Shallice, 1986). The results of this competition are largely driven by the strength of associative

pathways: stronger (i.e., more frequently activated) pathways are more likely to win the competition than weaker or novel ones. Once selected (and executed), the strength of a particular pathway may be reinforced or reduced depending on the assessment of how well the selected actions have fulfilled the organism's intended goals (Balleine & Dickinson, 1998).

This competition mechanism (or "contention scheduling", see Norman & Shallice, 1986) can generate appropriate behavior in many situations, but strong, stereotyped pathways can also result in suboptimal and even hazardous actions in some situations. For example, a US citizen's habitual driving on the right side of the road may have serious consequences when performed in the UK. In this case, a set of weaker or even novel associations (e.g., driving on the left side of the road) must be biased to win the competition in order to achieve the organism's goals. This "top-down" biasing of information processing to favor goal-directed stimuli and actions is the essence of cognitive control (e.g., Norman & Shallice, 1986; Botvinick et al., 2001; Miller & Cohen, 2001). In present-day neuroanatomical models, cognitive control is closely tied to the prefrontal cortex (PFC), which is proposed to harbor temporary representations of current goals, goal-relevant stimuli and strategies (Badre, 2008; M. M. Botvinick, Braver, Barch, Carter, & Cohen, 2001; Braver & Barch, 2002; Duncan, 2001; Fuster, 2008; Koechlin, Ody, & Kouneiher, 2003; Miller & Cohen, 2001; Norman & Shallice, 1986). To implement control, representations of goals, context and related methods (like rules) are

thought to be actively maintained in the PFC, which sends biasing signals to posterior brain regions to guide the information flowing through the desired pathways and reach the selection of appropriate actions (e.g., Miller & Cohen, 2001).

In the laboratory, cognitive control is traditionally tested in interference (or "conflict") tasks such as the Stroop task (MacLeod, 1991), which entail conditions that require subjects to overcome a stronger habitual response in favor of a weaker (but correct) response. Consider, for instance, a variant of the Stroop task employed in the empirical section of this dissertation (Figure 1). This task requires a subject to respond to the gender of a face image, while ignoring a word label (either "male" or "female") that is overlaid on the image and which can be either congruent (e.g., "male" overlaid on a male face) or incongruent (e.g., "female" overlaid on a male face) with the face image (Egner, Etkin, Gale, & Hirsch, 2008). In order to arrive at the correct response during an incongruent trial, the subject has to overcome the highly automatic processing of the word-meaning in favor of categorizing the face's gender. Correct response selection on incongruent trials therefore requires the application of cognitive control in the PFC, strengthening the information flowing through the task-relevant processing pathway to win out over the task-irrelevant (though more habitual) one (M. M. Botvinick et al., 2001; Braver & Barch, 2002; Cohen, Dunbar, & McClelland, 1990). Accordingly, many neuroimaging studies of these types of tasks have documented higher activation in the PFC associated with higher conflict and control levels (Barch et al., 2001; M. M.

Botvinick, Cohen, & Carter, 2004; Ridderinkhof, Ullsperger, Crone, & Nieuwenhuis, 2004), and modulated activity in brain regions related to processing task relevant- and irrelevant stimuli (Egner & Hirsch, 2005a; King, Korb, von Cramon, & Ullsperger, 2010; Liu, Banich, Jacobson, & Tanabe, 2004; Wittfoth, Buck, Fahle, & Herrmann, 2006).

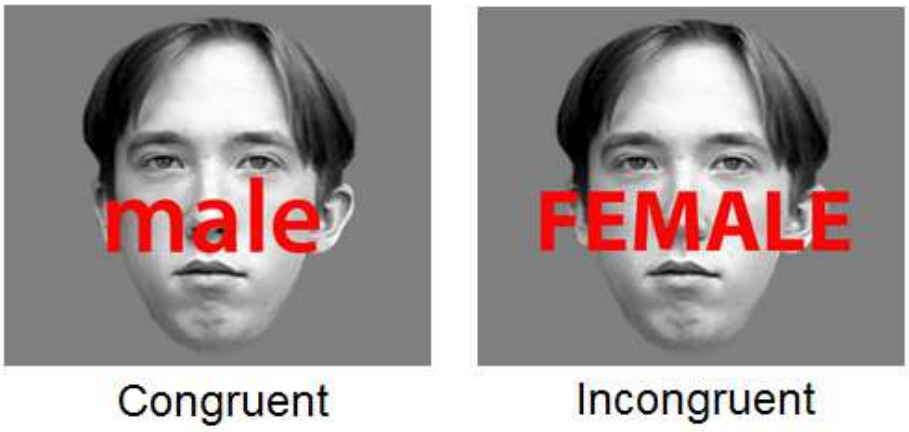


Figure 1: Example trials of a gender variant of the Stroop task.

One crucial question regarding this account, however, is how cognitive control itself is controlled. For example, when does cognitive control engage to bias competition of pathways? How does it change strength when more or less control is needed? And how is control withdrawn? In this dissertation, I argue (as have others before me, see Botvinick et al., 2001) that the regulation of cognitive control relies on the prediction of processing demands (e.g., anticipated conflict or congruency levels), which is derived from previous experience. In the following, two influential non-Bayesian models are reviewed: the conflict monitoring model (M. M. Botvinick et al., 2001), and the dual mechanisms of control model (Braver, 2012). Both models adjust the level of cognitive

control based on previous experience. Yet, as described below in detail, none of these models can explain how the brain flexibly incorporates and combines information across different time scales (short-term and long-term) to predict conflict. As a potential solution, this flexibility can be modeled using a Bayesian approach. Then I review basic concepts of Bayesian methods and several attempts to model various aspects of cognitive control using Bayesian models.

1.2 Classic Models for Cognitive Control

1.2.1 The Conflict Monitoring Model

The conflict monitoring model (M. M. Botvinick et al., 2001) treats the intervention of cognitive control as a reactive processing adjustment following the detection of conflict. This adjustment is achieved by incorporation of two systems: a conflict monitoring system that estimates the levels of conflict and sends signals to a control system, which in turn delivers biasing signals to information processing pathways. It is not entirely clear in the model whether control is originally recruited for dealing with conflict in the ongoing trial or for subsequent trials only (for discussion, see Egner, Ely, & Grinband, 2010), but the effects of conflict-driven control that are seen to support the model are typically measured by observing performance on the subsequent trial(s).

The specific mechanisms of the conflict monitoring system are made explicit in a neural network implementation (M. M. Botvinick et al., 2001), in which reaction time

(RT) was simulated as the time-point when the Hopfield energy (Hopfield, 1982) of one output node (out of two or more) reached a pre-defined threshold. This neural network implementation successfully simulated various landmark behavioral effects found in interference tasks. For example, the proportion incongruency effect (see Figure 2a), which describes the pattern that the larger the proportion of congruent trials is in a block, the higher the average interference effect is in that block (Logan & Zbrodoff, 1979; Tzelgov, Henik, & Berger, 1992), and the congruency sequence (or conflict adaptation, see Figure 2b) effect - a smaller interference effect (measured by subtracting mean RT of congruent trials from mean RT of incongruent or neutral trials) following an incongruent trial than after a congruent trial (Gratton, Coles, & Donchin, 1992), have both been simulated successfully by the conflict-monitoring model using a reinforcement learning algorithm that updates the prediction of congruency by incorporating (in)congruency at the current trial via a fixed learning rate α. Specifically, the prediction for the forthcoming trial is a linear combination of the (in)congruency at the current trial and the prediction concerning the current trial, with the rates of α and $(1 - \alpha)$, respectively. The model further proposes that the conflict monitoring system is housed in the anterior cingulate cortex (ACC) and the control system in the lateral PFC. These propositions have been supported by neuroimaging findings showing elevated activation in the ACC under conditions where conflict is high and control is assumed to be low (Barch et al., 2001; M. Botvinick, Nystrom, Fissell, Carter, & Cohen, 1999; Carter

et al., 1998; Kerns et al., 2004; MacDonald, Cohen, Stenger, & Carter, 2000; MacLeod & MacDonald, 2000) and enhanced activation in lateral PFC under conditions where conflict is low and control is assumed to be high (Egner & Hirsch, 2005a; Kerns et al., 2004; MacDonald et al., 2000), as well as increased functional connectivity between the lateral PFC and regions supporting task-relevant stimulus information in the posterior brain (Egner & Hirsch, 2005a).

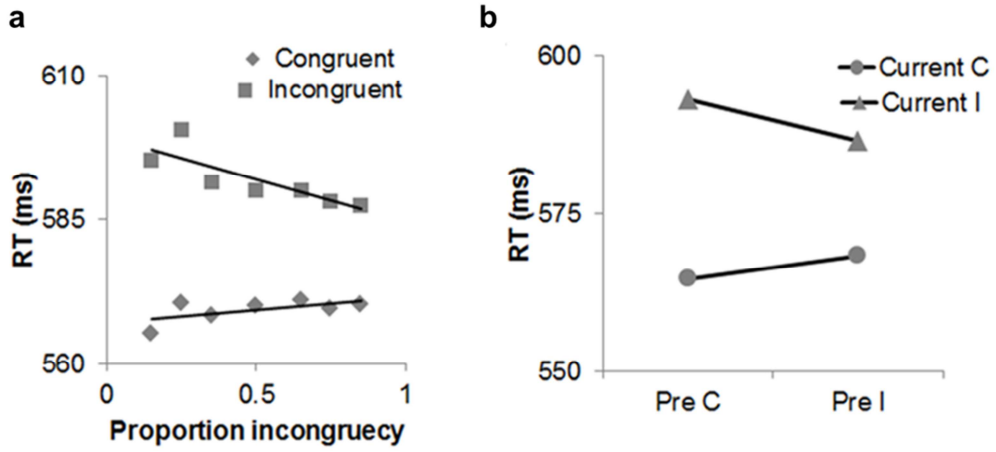


Figure 2: Data showing proportion incongruency (a) and conflict adaptation (b) effects. Pre C/Pre I = Preceded by a congruent/incongruent trial; Current C/Current I = current trial is congruent/incongruent.

Although the conflict monitoring model is able to simulate the phenomena of conflict adaptation and proportion incongruency effects (Botvinick et al (2001), simulation 2A and 2B) separately, a closer look at the simulation results suggests the model is not able to replicate these two effects using the same set of parameters. Specifically, in the simulation of conflict adaptation (simulation 2A), the best model has a learning rate of 0.5; while the learning rate is dramatically reduced to 0.05 when

simulating the decreasing interference effect as the proportion of incongruent trials increases in the simulation of proportion incongruency effects (simulation 2B). The 0.5 learning rate in simulating conflict adaptation effects essentially represents a phasic or transient mechanism relying more on recent experience, as the last trial weighs as much as all previous trials combined; while the 0.05 learning rate reflects a more tonic or sustained mechanism incorporating temporally more remote or extended information that allows for the proportion of incongruent trials to be learnt. The fact that the conflict-monitoring model cannot simulate both of these effects simultaneously is problematic, given that they are supposed to reflect the same basic phenomenon (conflict-driven control) and that conflict adaptation and proportion incongruency effects do in fact co-occur in a single task-setting (e.g., Torres-Quesada et al., 2013; also see section 3.1), a finding which the conflict-monitoring model is clearly unable to capture.

1.2.2 The Dual Mechanisms of Control Model

The more recent dual-mechanisms of control model may have the potential to overcome this problem, as it specifically accommodates control effects that operate over different times scales, by incorporating both a "reactive" and a "proactive" control mechanism (Braver, 2012; Braver, Gray, & Burgess, 2007; De Pisapia & Braver, 2006). The key difference between these two mechanisms lies in their time scales and their relation to stimulus onsets. Specifically, the reactive mechanism accounts for transient changes of cognitive control after a stimulus has been encountered (e.g., following conflict),

whereas the proactive mechanism monitors long-term changes of conflict density and applies changes to cognitive control before the onsets of incoming stimuli. Although operating on different time scales, these two mechanisms cooperate to modulate cognitive control. To test the feasibility of this model, De Pisapia & Braver (2006) conducted a color-naming Stroop fMRI study, which included three types of blocks with varying proportions of incongruent trials. The authors found that in ACC and left dlPFC the conflict-related (i.e. incongruent – congruent, at trial level) activity was highest when most trials were congruent (and proactive control presumably low), suggesting a reactive, short-term/phasic type of control being applied; whereas in the right dlPFC, the sustained, block-wise activation was the highest when most trials were incongruent, suggesting the wielding of a proactive, long-term/tonic type of control. The authors furthermore found that a model in which both ACC and the dlPFC units had a reactive and a proactive component could simulate both the phasic and tonic activation patterns found in the fMRI data. This dual mechanisms of control model represents a novel approach to understanding cognitive control, but there is presently little empirical evidence to support the idea of two conflict monitoring units working on different time scales in the ACC. It is also unclear whether this model can simulate both long-term (e.g. proportion incongruency) and short-term (e.g. conflict adaptation) regulation of control simultaneously, and it would be more parsimonious if both types of control were

integrated into a single mechanism. In the following section, I aim to sketch out how such integration can be achieved.

1.2.3 Cognitive Control as Statistical Inference

In the computer simulations of both conflict-monitoring and dual-mechanism models, short-term information (e.g. congruency at the current trial) and long-term information (e.g. congruency at earlier trials) were integrated using a fixed weight. Other computational cognitive control models using reinforcement learning (Blais, Robidoux, Risko, & Besner, 2007) and Hebbian learning (Verguts & Notebaert, 2008, 2009) have also used fixed parameters in their simulations of various behavioral phenomena of conflict-control. Although these simulations matched empirical data well, the use of a fixed weight for information integration elicits two important, yet unanswered questions: (1) how is the weight determined? And (2) (How) does the weight change when the (experimental) environment changes? To answer these questions, I argue that the weight should be self-adapting based on how reliable the short-term and long-term information is. The idea of a self-adapting learning rate is not a new concept: in classical conditioning, there have been models that use the novelty of stimuli to affect learning rate (Jiang, Schmajuk, & Egner, 2012; Pearce & Hall, 1980; Rescorla & Wagner, 1972; Schmajuk, Lam, & Gray, 1996). In these models, novelty guides changes in learning rate, which in turn updates the association between conditioned and unconditioned stimuli. Since there are no conditioned and

unconditioned stimuli in typical interference tasks, these models cannot be directly applied to simulating cognitive control processes. However, Bayesian models provide a natural solution of dynamically updating predictions based on integrating prior, temporally remote (long-term information) with recent observations (short-term information). Accordingly, several recent studies have employed Bayesian methods to model aspects of cognitive control, which are reviewed in the next section, preceded by an overview of Bayesian methods.

1.3 Overview of Bayesian Methods

Bayes' theorem can be written as follows:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad \text{(Eq 1)}$$

Where X, Y are random variables (e.g. sensory input, internal states, motor output, etc). Unlike conventional variables, the value of a random variable can vary due to randomness. Thus, a random variable is often represented in the probabilistic distribution of its possible values. This equation means that the conditional probability of Y given X could be calculated using the probabilities of X, Y, and the conditional probability of X given Y. This equation is especially useful when $P(Y|X)$ (posterior probability) is difficult to estimate but $P(X|Y)$ is relatively easy to obtain. For example, when X is an observation and Y is an internal state which cannot be observed directly, one can infer the state of Y based on X and $P(X|Y)$ using the Bayes' theorem. Thus, Bayes' theorem can be used to infer, for instance, the distributions of conflict-control, based on

the congruency observed. The estimated internal states can then be used to predict congruency in forthcoming trials. Bayesian methods have been widely applied in cognitive neuroscience studies (e.g., (Bach & Dolan, 2012; Vilares & Kording, 2011)), and a comprehensive review of studies using Bayesian methods is beyond the scope of this dissertation. Instead, I focus on Bayesian models that employed a graphical representation, because it provides a natural representation of dependence on previous information.

A graphical representation of a Bayesian model (Pearl, 1988) consists of a set of nodes and a set of edges connecting pairs of nodes. A node represents a variable in a Bayesian model, such as conflict, or observed congruency. In addition, a node is associated with the probability distributions of the variable represented. An edge represents a relation (reflecting conditional independencies) between two nodes and can be either directed or undirected. For example, a directed edge from node A to node B means that the value of variable B depends on variable A, but not vice versa. The edge is also associated with a distribution on which the estimation of parameters and Bayesian inference is based. This distribution encodes interactions between the two variables connected. Specifically, a directed edge from node A to node B is associated with a conditional probability distribution $p(B|A)$, which encodes how variable A influences variable B.

For example, the temporal dependency of conflict between trials can be formally represented using a Bayesian model (Figure 3). In this model f_i, and o_i denote predicted conflict level and observed congruency at trial i, respectively. f_i is quantified as the probability that the forthcoming trial is incongruent, ranging from 0 to 1. o_i is a binary variable in which 0 and 1 encode a congruent and an incongruent trial, respectively. The temporal dependency is represented by edges from the states at the current trial to the states at the next trial. An edge from predicted conflict level to observation is added to estimate f_i using Bayesian inference.

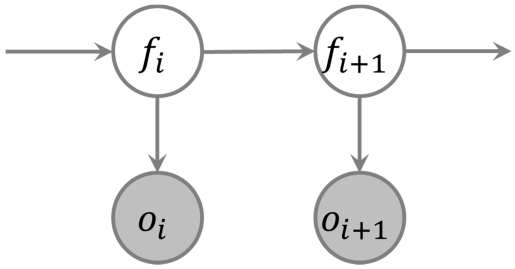


Figure 3: A basic Bayesian model of conflict-control. The model entails 2 variables, conflict (f) and observation (o, shown in grey indicating this variable is observable) for each trial. The directed edges indicate the information flow.

This Bayesian model does not only allow dependency between variables to be incorporated, but also significantly reduces the amount of computation needed to infer the states of these variables. Based on the structure of the graphical representation and the Markov property which states that each variable's future value is conditionally independent of the past, given its present value, the joint distribution in this model of cognitive control can be factorized as:

$$p(f_{i+1}, o_{i+1}|o_1 \ \ldots, o_i) = \int p(f_i)p(f_{i+1}|f_i)p(o_{i+1}|f_{i+1})\, df_i \qquad \text{(Eq 2)}$$

The conditional probability distribution on the left side contains many variables. Without any prior knowledge of the model structure, it is intractable to calculate, or even store this distribution. However, the Markov condition decomposes this distribution into a product of three much simpler distributions, each of which is easy to store and compute. Specifically, the Markov condition shows that the prediction (states at trial $i + 1$) based on all previous information is equivalent to the prediction based only on the previous trial. In other words, relevant historical information is integrated into the states of the most recent trial. Thus, storage and computation involving older trials are not necessary.

In sum, the graphical representation incorporates dependency between model variables; and its structure greatly reduces the computational and storage burden of estimating posterior probabilities.

1.4 Bayesian Models for Cognitive Control

Recently, Bayesian models with graphical representation have demonstrated great potential in modeling cognitive functions using behavioral data (Mozer, Colagrosso, & Huber, 2002; Reynolds & Mozer, 2009; Shenoy, Rao, & Yu, 2010; Shenoy & Yu, 2011; Tenenbaum, Griffiths, & Kemp, 2006; Tenenbaum & Xu, 2000; Vossel et al., 2013; Yu, Dayan, & Cohen, 2009) and brain imaging data (Behrens, Woolrich, Walton, & Rushworth, 2007; den Ouden, Daunizeau, Roiser, Friston, & Stephan, 2010; Ide, Shenoy,

Yu, & Li, 2013). In the following, I review recent studies using Bayesian models with graphical representation to model various aspects of cognitive control, including speed-accuracy trade-off (section 1.4.1), conflict effects in the Eriksen flanker task (section 1.4.2), and response inhibition (section 1.4.3). The Bayesian models reviewed below all attempted to account for decisions/behavior using within-trial and/or inter-trial simulations. In the within-trial simulations, those models accumulated evidence from independent sources via Bayesian integration. The decision/behavior best supported by the evidence was then selected by the models. In the between-trial simulations, predictions of stimuli were based on trial-history information integrated via Bayes' rule. The predictions were then used as the initial evidence for within-trial simulations.

1.4.1 Bayesian Modeling of Speed-accuracy Trade-off

Some recent studies demonstrate the feasibility of using generative Bayesian models to simulate both within-trial dynamics and across-trial sequential effects in cognitive control. One such study applied a Bayesian model to explaining the dynamics of the speed-accuracy tradeoff and its dependency on trial-history from two speeded discrimination tasks (Mozer et al., 2002). Subjects responded to a target letter by pressing one button and responded to other letters by either pressing a second button (discrimination task) or doing nothing (go/no-go task). The Bayesian model used in this study operates at both within- and across-trial levels. At the within-trial level, the posterior distribution at a given time point encoded the probability of making one

response vs. the other. This posterior distribution of responses depended on the prior distribution of responses and the sensory input. At the beginning of each trial, the posterior distribution is the same as the prior distribution. This posterior distribution of responses is then subject to (Bayesian) updating after each time point to favor the response suggested by visual input. Because this updating was performed at every time point, the influence of visual input accumulated with time, guiding the posterior distribution of response to gradually shift from the prior distribution to a distribution which is biased toward the correct response to the visual stimulus. Therefore, the effect of cognitive control is reflected by the change of the posterior distribution based on accumulation of sensory input. The simulated RT was the time that optimized the cost of sensory input accumulation against the probability of an incorrect response. The simulated accuracy was estimated from the posterior distribution at the simulated RT. For both the discrimination task and the go/no-go task, within-trial simulation successfully replicated accuracy and RT patterns from empirical data under different target to non-target ratios. Based on these simulations, the authors argued that speed-accuracy trade-off is optimal in these tasks in that it minimizes a cost that combines time pressure and the certainty of perception. At the across-trial level, the initial prior was also updated after each trial using a similar rule as used in within-trial simulation to account for sequential effects of response priming. The across-trial simulation successfully captured complex RT and accuracy patterns when trials were grouped

based on the trial history, up to 4 trials preceding the current trial. In this study, then, a Bayesian model was used to simulate perceptual decision-making as an integration of prior information and visual input, which could naturally be modeled using the prior distribution and the likelihood distribution, respectively.

1.4.2 Bayesian Modeling of the Eriksen Flanker Task

To investigate different mechanisms that may account for generating conflict in the Eriksen flanker task (Eriksen & Eriksen, 1974, also see Figure 4), a study by Yu and colleagues (Yu et al., 2009) applied two rival Bayesian models to behavioral data from a "deadline version" of the flanker task (Gratton, Coles, Sirevaag, Eriksen, & Donchin, 1988; Servan-Schreiber, Bruno, Carter, & Cohen, 1998). Here, subjects were pressured to make fast responses (in order to beat an experimenter-imposed deadline) to a target (center) letter (either "S" or "H") that is flanked by distractors (either "S" or "H"). Thus a trial could be either congruent (e.g. "SSSSS" or "HHHHH") or incongruent (e.g. "HHSHH" or "SSHSS", see). Bayesian models were used to simulate two different potential sources of conflict, namely "compatibility bias" (a prior assuming that more than half of the trials were congruent) and "spatial uncertainty" (where perception of one letter was interfered with by nearby letters). Both models adopted the same hierarchical design with 3 levels: The highest level encoded the (in)congruency of a trial; below the congruency level was the stimulus level, in which there were 3 nodes, each representing a letter; at the bottom was a level of 3 nodes, each encoding activity of a

group of neurons whose receptive fields were centered on a particular letter. The difference between the two models lay in how prior knowledge was applied: in the compatibility bias model, the prior assumed there were more congruent trials than incongruent trials, and each node at the level of neuronal activity was only influenced by the letter it represented and random noise. This model simulated a situation in which the conflict monitoring process was biased to an expectation of low conflict and receptive fields of neuron groups were narrow. By contrast, the prior of the spatial uncertainty model assumed a 50/50 distribution of congruent and incongruent trials, but here a neuron's activity was influenced by not only the letter its receptive field centered on, but also the neighboring letter(s). This model simulated a situation in which neurons have large receptive fields and the interference is caused by the ambiguous neuronal signals containing information of different letters.

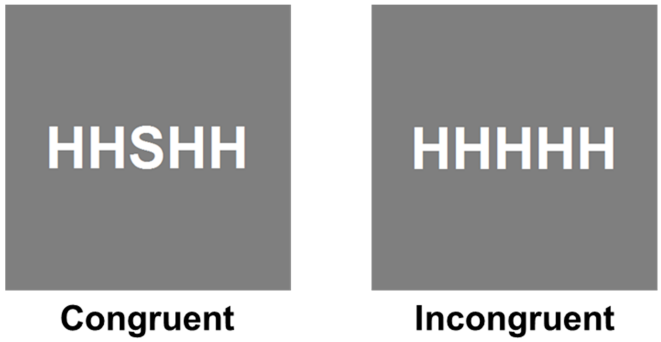


Figure 4: Example stimuli of an Eriksen fanker task.

Within-trial and between-trial simulations were then conducted using both models. To simulate within-trial dynamics, the simulation was divided into multiple

time steps. The two models operated as neural decoders: they estimated visual input and congruency based on simulated neuronal activity. Thus, the key part of the simulation is the joint posterior distribution of letter and congruency conditioned on neuronal activity, which started with the prior distribution of congruency and visual input, and was then updated at every time step based on Bayes' rule. A simulated response was made when the marginal posterior distribution of one letter exceeded a pre-defined threshold. Both models successfully replicated RT distributions acquired from empirical studies using deadline Eriksen flanker tasks. The two models were also extended to allow for across-trial updates of the prior distribution of congruency. The extended models were able to simulate conflict adaptation and proportion incongruency effects in Eriksen flanker tasks (Yu et al., 2009).

In another study, a Bayesian spotlight diffusion model was proposed to account for various aspects of the Eriksen flanker task (White, Brown, & Ratcliff, 2012). Specifically, a spotlight diffusion model (White, Ratcliff, & Starns, 2011) was used to simulate attentional mechanisms, and Bayesian belief-updating was employed to account for the information processing mechanisms involved in task performance (i.e., how evidence for response selection was accumulated within a trial). A spotlight was used to simulate the locus of attention, within which all information was selected as evidence for the decision-making process. At the beginning of a trial, the spotlight covered both the center target and the flankers, such that the overall evidence was

driven predominantly by the flanker stimuli. The spotlight then gradually narrowed to only cover the central letter, resulting in the evidence being biased toward the target information. At each time point, beliefs were updated to incorporate new evidence about what the correct response was, using Bayesian evidence integration. Accordingly, the model predicts that responses are biased to the flankers at the beginning of a trial and then gradually shift to be dominated by target information. By fitting the model to empirical data, it was shown that this Bayesian spotlight diffusion approach could successfully account for the relationship between RT and accuracy in the Eriksen flanker task (White et al, 2012).

1.4.3 Bayesian Modeling of Response Inhibition

Bayesian models have also been employed in investigating inhibitory control (Ide et al., 2013; Shenoy et al., 2010; Shenoy & Yu, 2011). In these studies, subjects performed a stop-signal task, in which a habitual response based on a (frequent) go signal needed to be suppressed when a (rare) stop signal was presented at varying intervals after the go signal. Bayesian models were used to provide a rational account (i.e. behavior guided by optimizing a cost function) for various behavioral patterns observed in the stop-signal task. Here, one Bayesian model was used to simulate beliefs about the appropriate action to take, and a second Bayesian model was used to simulate beliefs about perceiving a stop signal. Within each trial, both beliefs started with the (true) prior probability of go/stop trials in the task, and were then updated based on

visual input using Bayes' rule at every time step. For each time step, a cost function was calculated based on the beliefs and possible actions available. An action (i.e., button-press vs. withholding response) was selected by minimizing the cost function. This within-trial Bayesian model successfully simulated the pattern of increased error rates as the interval between the onset of the go signal and the onset of the stop-signal increased. It could also simulate the commonly observed faster responses in error stop trials compared to successful go trials (Shenoy et al., 2010; Shenoy & Yu, 2011). A third Bayesian model was employed for simulating between-trial effects, predicting the likelihood of encountering a stop-signal in the forthcoming trial. This prediction was a linear combination of the prior probability of encountering a stop-trial and the posterior probability of encountering a stop-trial based on trial history. These two probabilities were integrated via fixed weights. The prediction significantly correlated with RTs, where high probability of encountering a stop signal predicts slower responses, suggesting more inhibitory control being exerted (Ide et al., 2013). Additionally, this model successfully simulated post-stop-trial response slowing, and increased RTs and error rates when the proportion of stop-trials increased (Shenoy et al., 2010; Shenoy & Yu, 2011). Based on these Bayesian predictions, fMRI data recorded during the stop-signal task (Ide et al., 2013) further revealed that the dorsal ACC encodes prediction error (presence/absence of stop signal - prediction).

1.5 Towards Modeling the Flexibility of Cognitive Control

Although the models reviewed above were successful in simulating various effects of cognitive control, the fixed parameters used in these simulations raise several concerns. First, it is unclear whether the way in which these parameters were determined in the models reflects the mechanisms of parameter-selection in the brain. This is especially unlikely (or impossible) in cases where optimal parameters were fit from data: here, a model determines the parameters after acquiring all data, in contrast to the brain having to determine the parameters on the fly. Second, even with a model that could potentially provide on-the-fly simulation (e.g. the model in Figure 3), the lack of a mechanism for online adjustment of cognitive control would result in sub-optimal performance in a non-stationary environment, because without such a mechanism, the parameters that determine the level of cognitive control (e.g. the learning rate) have to stay constant throughout the experiment. Given the high flexibility required of cognitive control, it is unclear that those fixed parameters are globally optimal across various experimental configurations. In fact, it has been shown that fixed learning rates are suboptimal in non-stationary environments (Behrens et al., 2007; den Ouden et al., 2010; Vessel et al., 2013). In the next chapter, I therefore propose a Bayesian model that resolves these two concerns. This model simulates the on-the-fly selection of parameters in the brain by making estimates only based on previous trial information. It also models the flexibility of cognitive control by incorporating a component that accounts for

changes in the experimental environment. This Bayesian model has the potential of tackling the research question of how a parsimonious, unified mechanism can flexibly adjust cognitive control to adapt to various environments. By applying this model to empirical data, I demonstrate that it can simulate various key phenomena of cognitive control in a Stroop task in Chapter 3. Facilitated by this Bayesian model, I show that neural substrates supporting this flexible mechanism and their connections can be explored using brain imaging techniques (Chapter 4).

2. A Bayesian Model of Cognitive Control

In this chapter, I propose a Bayesian model that can account for the flexibility of cognitive control over conflict in a non-stationary environment. The modeling done relies on the ability to perform statistical inference, taking the perspective that the regulation of cognitive control should be considered as a process of predicting the optimal amount of cognitive control required in a given context. To achieve this contextual flexibility, the model estimates future conflict from previous experience and, importantly, it does so via a weighed integration of longer-term and short-term estimations of conflict distributions, with the integration weights being adjusted on the basis of the (belief about) volatility of the environment. For instance, in a stable environment (e.g. when most trials are congruent / incongruent), the weights are biased to historically remote/long-term information because an occasional oddball trial (e.g., an incongruent stimulus in a largely congruent trial history context) is unlikely to reflect a true change in the environmental statistics. When the environment is fast-changing (e.g. when the proportion of congruent trials varies frequently over time), however, the weights are biased to more recent information. This is because older information is likely to be outdated, and an unexpected trial type may indicate a true change of conflict likelihood in the environment. In order to assess the stability of the environment, I extend the model in Figure 3 by adding a volatility variable (denoted by v), the belief of which in turn determines the weights of integration (Figure 5). The structure of this

model is identical to the model of Behrens et al., (2007). The mathematical formulation of the model is described in details in section 2.1. This model is also an example of a hierarchical model, in which the information flows in one direction, that is, there are no reciprocal edges from nodes at lower levels to nodes at higher levels. Hierarchical Bayesian models have been widely used in modeling cognitive functions such as categorization (Tenenbaum & Xu, 2000; Xu & Tenenbaum, 2007) and visual cognition (Lee & Mumford, 2003; Summerfield, Behrens, & Koechlin, 2011). This model yields a probability distribution over the predicted conflict level variable. In the implementation of this model, predicted conflict level is approximated using the probability of encountering an incongruent trial. In other words, the predicted conflict level is higher if the next trial is deemed more likely to be incongruent. Both variables are then used to determine the amount of control needed, which is reflected in sequential effects such as conflict adaptation and longer-term effects like proportion incongruency. Another way to understand the relation between volatility and predicted conflict level can be drawn from Yu and Dayan (2005): predicted conflict level encodes the probability (distribution) that the forthcoming trial is incongruent. The variance of this distribution (i.e. change of probability) is determined by the volatility. This chapter is organized as follows: section 2.1 describes this model and its implementation in details. Sectoion 2.2 presents proof-of-principle validations of the proposed model.

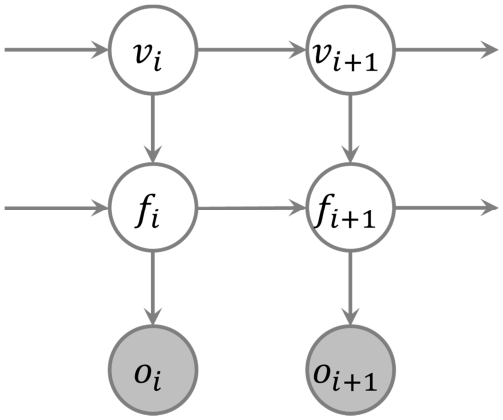


Figure 5: The graphical representation of the Bayesian model of flexible conflict-control. The model uses 3 variables, volatility (v), conflict (f), and observation (o, shown in grey indicating this variable is observable) for each trial. The directed edges indicate the information flow.

2.1The structure of the Bayesian Model

The graphical representation of the proposed Bayesian model is shown in Figure 5. Each row represents a variable, namely volatility (v), conflict (f), and observation (o). Each column represents the state of the variables in a trial. The generative model (how the distribution of each variable is determined by other variable(s) and / or parameter(s)) is shown as follows (see below for definition of distribution parameters):

$$\begin{aligned} &\mathrm{v_{i+1}} \sim \mathrm{N}(\mathrm{v_i}, \sigma_\mathrm{v}) \\ &\mathrm{f_{i+1}} \sim \mathrm{Beta}(\mathrm{f}_i 2^{v_{i+1}} - 2\mathrm{f}_i + 1, -\mathrm{f}_i 2^{v_{i+1}} - 2\mathrm{f}_i + 2^{v_{i+1}} - 1) \\ &\mathrm{o_{i+1}} \sim \mathrm{Bernoulli}(\mathrm{E}(\mathrm{f_{i+1}})) \end{aligned} \qquad \text{(Eq 3)}$$

This Bayesian model simulates a trial sequence of an experiment on a trial-by-trial basis. Within each trial, the simulation takes the form of a predict-update algorithm that is used in Kalman filters (Johan Masreliez & Martin, 1977). In other words, the simulation of each trial $i + 1$ contains two steps. The first step makes prediction of the

states of v_{i+1} and f_{i+1} before the stimulus is presented. These predictions are used to account for the behavioral patterns observed in the empirical studies reported in chapter 3 (see below) and brain activity presented in chapter 4 (see below). The second step updates / filters the belief of the states of v_{i+1} and f_{i+1} given the observed congruency o_{i+1}. These two steps are repeated for each trial to generate trial-by-trial estimates.

In the first step, the model initially predicts a joint distribution based on the model's previous states and 2 transition distributions:

$$p(\sigma_v, v_{i+1}, f_{i+1}|o_1, \dots, o_i)$$
$$= \iint p(v_{i+1}|v_i, \sigma_v)\, p(f_{i+1}|f_i, v_{i+1}) p(\sigma_v, v_i, f_i|o_1, \dots, o_i) dv_i df_i \quad \text{(Eq 4)}$$

Where $\sigma_v$constrains how fast v can change over time. Specifically, $p(v_{i+1}|v_i, \sigma_v) \sim N(v_i, \sigma_v)$. In other words, the transition distribution $p(v_{i+1}|v_i, \sigma_v)$ is Gaussian distribution with the mean of v_i and the standard deviation (SD) of σ_v. This transition distribution assumes that v is most likely to remain in its previous states, although it can also possibly drift to another state. σ_v determines how likely it is for v_i to shift to a new state. Because in the experiments below the volatility altered between conditions and trials, σ_v is set as a variable and used the Bayesian model to infer its state. Different from other variables, σ_v is considered as unchanged throughout an experiment. Thus, as can be seen below, it does not have a subscript indicating its temporal state or a transition distribution that governs its temporal shift. The update of

its estimate at each trial reflects the model 's change of its belief of σ_v, rather than the change of the true underlying σ_v per se. The transition distribution $p(v_{i+1}|v_i, \sigma_v)$ determines how the estimate of v_i is used to predict its future state, which is in turn employed to compute the predicted conflict level, as described below.

$p(f_{i+1}|f_i, v_{i+1})$ describes how f can change over time.

$p(f_{i+1}|f_i, v_{i+1})$~Beta(α, β). This distribution is a beta distribution, with its parameters α and β in the following form:

$$\begin{cases} f_i = \dfrac{\alpha - 1}{\alpha + \beta - 2} \\ v_{i+1} = \log_2(\alpha + \beta) \end{cases} \quad \text{(Eq 5)}$$

There are two main reasons for using a beta distribution: first, the predicted conflict level was defined as the probability of encountering an incongruent stimulus in the upcoming trial, so the possible value of f_{i+1} should be limited to a range of 0 to 1, which is also the range of values that a beta distribution is defined on. Second and more important, with the current set-up, the probability density function of $p(f_{i+1}|f_i, v_{i+1})$ takes the following form:

$$p(f_{i+1}|f_i, v_{i+1}) = \text{constant} \cdot f_{i+1}^{\alpha-1} \cdot (1 - f_{i+1})^{\beta-1} \quad \text{(Eq 6)}$$

Which can be interpreted as the likelihood function of observing $\alpha - 1$ incongruent trials and $\beta - 1$ congruent trials with the underlying proportion incongruency f_i. Thus, a larger $p(f_{i+1}|f_i, v_{i+1})$ suggests the prediction of conflict is better supported by temporally more-extended trial-history information. Following this

interpretation, according to the equations 5 and 6, v_{i+1} controls the length of the trial sequences of this likelihood function. A larger v_{i+1} suggests a longer trial sequence to take into account, which in turn indicates more dependence on long-term information, or a more stable condition. In other words, a larger v_{i+1} leads to a narrower spread of $p(f_{i+1}|f_i, v_{i+1})$, which constrains f_{i+1} from drifting far from its previous state. Another interpretation of $p(f_{i+1}|f_i, v_{i+1})$ can be linked to the learning rate used in many reinforcement learning models: a narrower spread of $p(f_{i+1}|f_i, v_{i+1})$ results in smaller difference between f_i and f_{i-1} that resembles the effect of a smaller learning rate, compared to a wider spread of $p(f_{i+1}|f_i, v_{i+1})$ determined by a smaller v_{i+1}. However, it is counterintuitive to have a larger volatility value when the environment is more stable. Thus when reporting volatility, a linear transform was used to make more volatile task settings correspond to larger volatility estimates while preserving the quantitative patterns of the results (see below). Furthermore, f_i is the mode of $p(f_{i+1}|f_i, v_{i+1})$, indicating that the predicted conflict level is most likely to reflect its previous state. Thus, $\iint p(v_{i+1}|v_i, \sigma_v)\, p(f_{i+1}|f_i, v_{i+1}) dv_i df_i$ can be viewed as how the prediction is made by incorporating various learning rates. The third term in the integral, $p(\sigma_v, v_i, f_i|o_1, \ldots, o_{i-1})$ is the belief in the previous trial and also represents the weights $p(\sigma_v, v_{i+1}, f_{i+1}|o_1, \ldots, o_i)$ in making the prediction (Eq. 3). After the joint distribution $p(\sigma_v, v_{i+1}, f_{i+1}|o_1, \ldots, o_i)$ is calculated, the estimates of volatility and conflict are computed as the mean of their corresponding marginalized distributions. The observed

congruency o_{i+1} is then predicted to have a Bernoulli distribution with a probability of $E(f_{i+1})$ of incongruency, where $E(f_{i+1})$ denotes the mathematical expectation of f_{i+1}.

In the first step, I adopted a numeric implementation for this Bayesian model to avoid the complexity of developing an analytical implementation: the range of each of the variables of σ_v, v and f was divided into multiple segments with equal length. For example, f_i was represented using an array ranging from 0 to 1 and with a step size of 0.02 (that is, 51 cells). The value of each cell represented the probabilistic density at that point. Similarly, the joint probabilistic distribution was represented by a 3D array with 3 dimensions of σ_v, v_{i+1} and f_{i+1}. $p(f_{i+1}|f_i, v_{i+1})$ was represented using a 3D array with 3 dimensions of f_{i+1}, f_i and v_{i+1}. And $p(v_{i+1}|v_i, \sigma_v)$ was represented using a 3D array with 3 dimensions of v_{i+1}, v_i and σ_v. All the aforementioned calculations were performed on these arrays. Specifically, step 1 took the form of:

$$p(\sigma_v, v_{i+1}, f_{i+1}|o_1, \dots, o_i) = \sum_{v_i} \sum_{f_i} p(v_{i+1}|v_i, \sigma_v)\, p(f_{i+1}|f_i, v_{i+1}) p(\sigma_v, v_i, f_i|o_1, \dots, o_i) \quad \text{(Eq 7)}$$

After the first step, $p(\sigma_v, v_{i+1}, f_{i+1}|o_1, \dots, o_i)$ was divided by its sum across σ_v, v_{i+1} and f_{i+1}so that the cells in $p(\sigma_v, v_{i+1}, f_{i+1}|o_1, \dots, o_i)$ summed to 1. The marginalization was done by collapsing the other dimensions. The mean was approximated using a weighed sum, $\sum xp(x)$.

In the second step (i.e. after the congruency of trial i + 1 is observed), the belief of variables is updated using the observed congruency in the following manner:

$$p(\sigma_v, v_{i+1}, f_{i+1}|o_1, \dots, o_{i+1}) = p(\sigma_v, v_{i+1}, f_{i+1}|o_1, \dots, o_i)p(f_{i+1}|o_{i+1}) \quad \text{(Eq 8)}$$

Where

$$p(f_{i+1}|o_{i+1}) = \begin{cases} 1 - f_{i+1}, \text{if } o_{i+1} \text{ is incongruent} \\ f_{i+1}, \text{if } o_{i+1} \text{ is congruent} \end{cases} \quad \text{(Eq 9)}$$

is the prediction error between the predicted conflict level and the true congruency.

In theory, the initial $p(\sigma_v, v_{i+1}, f_{i+1})$ (the dependence is removed to reflect that no knowledge from the trial sequence has been applied yet) can take any form to reflect the prior knowledge of the trial sequence. In the implementation of this model, the initial $p(\sigma_v, v_{i+1}, f_{i+1})$ was set to a uniform distribution, indicating unbiased belief of the statistical characteristics of the task. As the experiment proceeded (i.e., more trials were processed), this initial distribution was updated to a distribution that with more information of the trial sequence. Note that similar to many models, this initial distribution requires sampling from the trial sequence for proper update. Thus a "burn-in" block is usually added to a trial sequence for this Bayesian model to update its belief to adapt to the trial sequence. And the trials in this burn-in period are sometimes excluded from analysis. This is because that there is no guarantee that the subjects and the Bayesian model share the same prior knowledge. However, after the burn-in period,

it is safe to assume that both the subjects and the model have refreshed their belief to better approximate the task. Hence the model can then be used to account for behaviors and brain activity.

As mentioned before, the definition of v would result in a counterintuitive representation of having a lower v in a more volatile setting. Thus, to avoid confusion, I transformed the condition-specific volatility estimates before presenting them:

$$\tilde{v}_j = \max(v) + \min(v) - v_j \quad \text{(Eq 10)}$$

Where max(v) and min(v) are the maximum and minimum group-level mean of volatility estimates across all conditions, respectively. After this transform, more volatile conditions have higher $\tilde{v}_j$s. Note that because all v_js were transformed using the same constants, this transform had no bias on the results of analyses reported below.

2.2 Proof-of-principle Validations

To test whether this Bayesian model is able to learn from a trial sequence and, more importantly, whether it is able to detect changes of volatility in an experimental manipulation, I conducted 3 validation studies. In section 2.2.1, I qualitatively describe the behavior of this model in learning a single trial sequence. In section 2.2.2, I demonstrate that the model is able to detect changes in volatility due to varying underlying proportions of (in)congruency. In section 2.2.3, I propose a task design that induces trial blocks with varying volatility. Using this design, I show that the Bayesian model successfully detects different volatility levels due to various frequencies of

changes in proportion (in)congruency. The performance of this model and reinforcement learning models with various fixed learning rates are also compared using trial sequences generated under this task design.

2.2.1Time Courses of Model Variables in Simulating a Trial Sequence

In order to inspect this Bayesian model's performance, a sequence of 100 randomly generated trials (proportion incongruency = 0.8) was fed to the model. The time courses of trial-by-trial estimates of predicted conflict level and volatility derived from the model are depicted in Figure 6. The first 10 trials were used as burn-in period and were thus discarded. As can be seen in Figure 6a, the predicted conflict level in general approaches the underlying proportion incongruency, suggesting that the model estimates incorporate long-term information. In addition, after a (rare) congruent trial, the predicted conflict level is reduced in the subsequent trial, indicating that the estimates also incorporate short-term information. The time course of volatility (Figure 6b) displays an expected pattern: volatility drops in the absence of rare congruent trials, suggesting a stable environment. After encountering a congruent trial, the model raises its estimate of volatility.

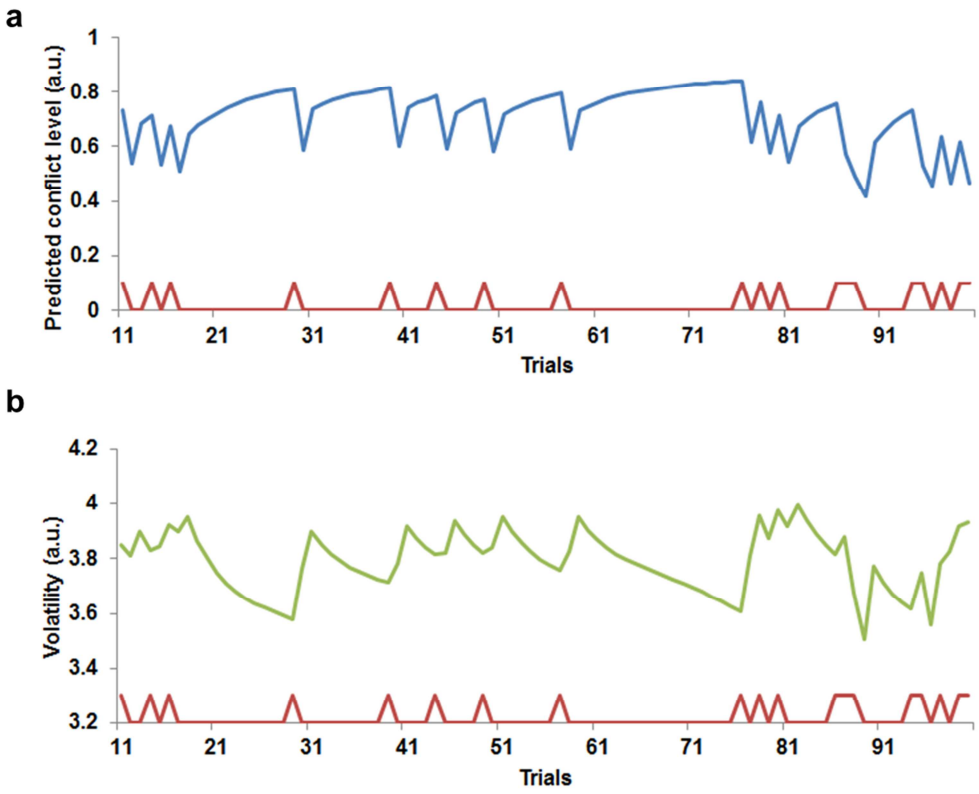


Figure 6: Time courses of predicted conflict level (a, blue line) and volatility (b, green line) in learning a randomly generated trial sequence that has a proportion incongruency of 0.8. The trial sequence is plotted in red line. Spikes in the red line indicate onsets of congruent trials.

In the Bayesian model, the purpose of the volatility variable is to provide a learning-rate modulator: higher volatility should result in a larger learning rate. To test this hypothesis, I estimated the trial-by-trial learning rate based on the classic reinforcement learning algorithm. Specifically, according to the temporal difference (TD(0)) reinforcement learning algorithm (Sutton, 1988), the update of predicted conflict level can be expressed as follows:

$$f_{i+1} = (1 - \alpha)f_i + \alpha o_i \qquad \text{(Eq 11)}$$

Thus, given the estimates of predicted conflict level $f_{i+1}$and f_i, and the congruency o_i, the learning rate α can be estimated as:

$$\alpha = \frac{f_{i+1} - f_i}{o_i - f_i} \quad \text{(Eq 12)}$$

The estimated learning rate exhibits a highly significant positive correlation with volatility (Figure 7, $r = 0.55$, $P < 0.0001$), indicating that volatility does in fact regulate the updating of predicted conflict level. Specifically, because o_i represents short-term (most-recent and only contains a single trial) information, the α indeed represents the dependence on short-term information in predicting future conflict level. Accordingly, the positive correlation between learning rate and volatility suggests that the Bayesian model augments its reliance on short-term information in predicting conflict level when the environment becomes more volatile.

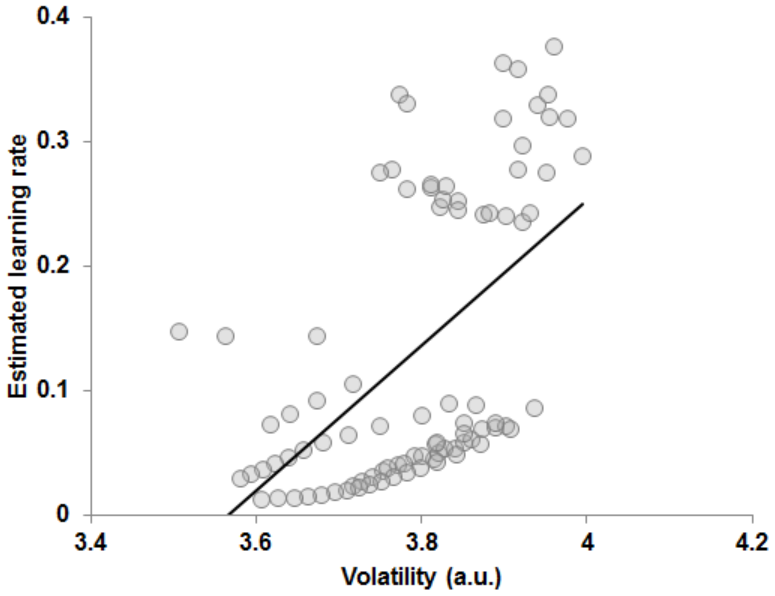


Figure 7: Estimated learning rate, plotted as a function of volatility. The trend line shows a significant positive correlation between these 2 estimates at the trial level.

2.2.2 Modeling Proportion Incongruency Induced Volatility

A crucial validation for this Bayesian model is to test its sensitivity to the differences in volatility between various experimental settings. To this end, I first tested whether the model could distinguish volatility caused by proportion incongruency. Recall that one way of understanding volatility is to treat is as the certainty of prediction (Yu & Dayan, 2005). From this perspective, a trial sequence with more extreme proportion incongruency (e.g., 0.9 or 0.1) should be less volatile compared to a trial sequence with a proportion incongruency of 0.5, because in the former the prediction of the incongruency is more certain (i.e., the SD of congruency is smaller). Similarly, the proportion incongruency of 0.9 and 0.1 should have similar volatility. If the Bayesian

model operates as expected, its estimates of volatility should reflect these varying levels of volatility across the proportion incongruency conditions.

For each of three proportion incongruency conditions (0.1, 0.5 and 0.9), fifty trial sequences were randomly generated, each containing 100 trials. The trial sequences were then processed by the Bayesian model. The resulting estimates of predicted conflict level and volatility after the burn-in period (the first 10 trials) were averaged. The condition mean predicted conflict level and volatility were then compared across conditions using 1-way ANOVA and post-hoc 2-sample t-tests.

As a sanity check, the Bayesian model successfully distinguished the condition-mean predicted conflict level between the 3 conditions (Figure 1Figure 8a; ANOVA: $F_{2,147}$= 62416.24, P < 0.0001; Post-hoc t-tests: proportion incongruency of 0.9 > proportion incongruency of 0.5, t_{98} = 182.23, P < 0.0001; proportion incongruency of 0.5 > proportion incongruency of 0.1, t_{98} = 171.39, P < 0.0001; proportion incongruency of 0.9 > proportion incongruency of 0.1, t_{98} = 353.39, P < 0.0001). The comparison of volatility also matches the predictions: a significant main effect was found in the ANOVA (Figure 8b; ANOVA: $F_{2,147}$= 8.22, P < 0.001). Post-hoc t-tests showed that the volatility in the condition with 0.5 proportion incongruency was significantly higher than those in the conditions with 0.9 (t_{98} = 3.63, P < 0.001) and 0.1 (t_{98} = 3.62, P < 0.001) proportion incongruency. No significant difference of volatility was found between the conditions with 0.9 and 0.1

proportion incongruency (t_{98} = 0.40, P > 0.69). Thus, as expected, the Bayesian model successfully distinguishes volatility between different experimental settings.

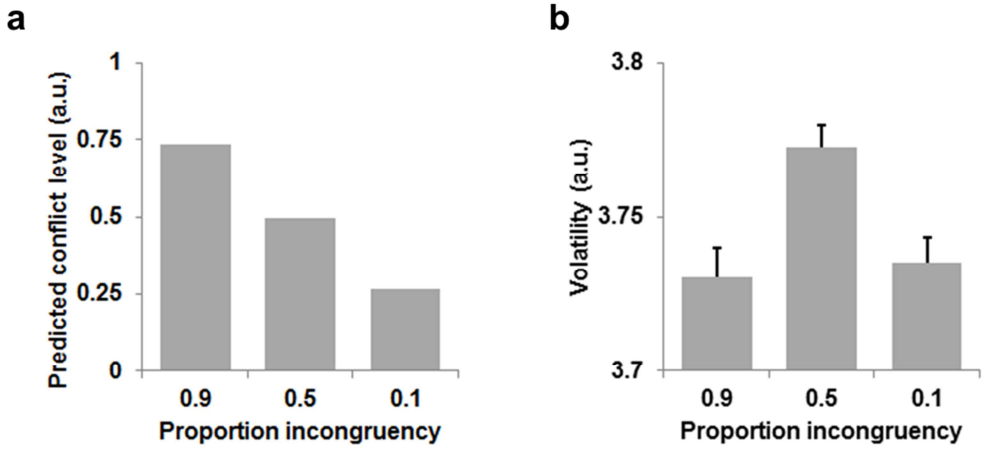


Figure 8: Group mean predicted conflict level (a) and volatility (b) and their mean standard error (MSE), plotted as a function of the underlying proportion incongruency. Note that in (a) the error bars are too short to become visible.

2.2.3 Modeling Volatility Induced by the Frequency of Alternating Proportion Incongruency

Another strategy of creating experimental conditions with various volatility is to manipulate the frequency of alternating two levels of pre-selected proportion incongruency (Behrens et al., 2007). A higher frequency of alternation leads to a faster-changing environment and in turn makes for condition with higher volatility. Thus, distinguishing volatility induced by different frequencies of alternating underlying proportions of incongruency can serve as another test for the Bayesian model.

Following this rationale, two experimental conditions (the "stable" condition and the "volatile" condition) were designed. In each condition, a hundred trial sequences

were randomly generated. A trial sequence consisted of 100 trials, the first 20 of which had a proportion incongruency of 0.5 and served as the burn-in block. The proportion incongruency in remaining 80 trials stayed at 0.8 in the stable condition and alternated between 0.2 and 0.8 every 20 trials (0.05 Hz) in the volatile condition. The trial sequences were then learned by the Bayesian model. The trials in the burn-in block were removed from further analysis. The condition-specific time course of predicted conflict level was averaged across trial sequences and visualized for sanity check. The sequence mean volatility was compared between the 2 conditions to test if the model can successfully detect changes of volatility.

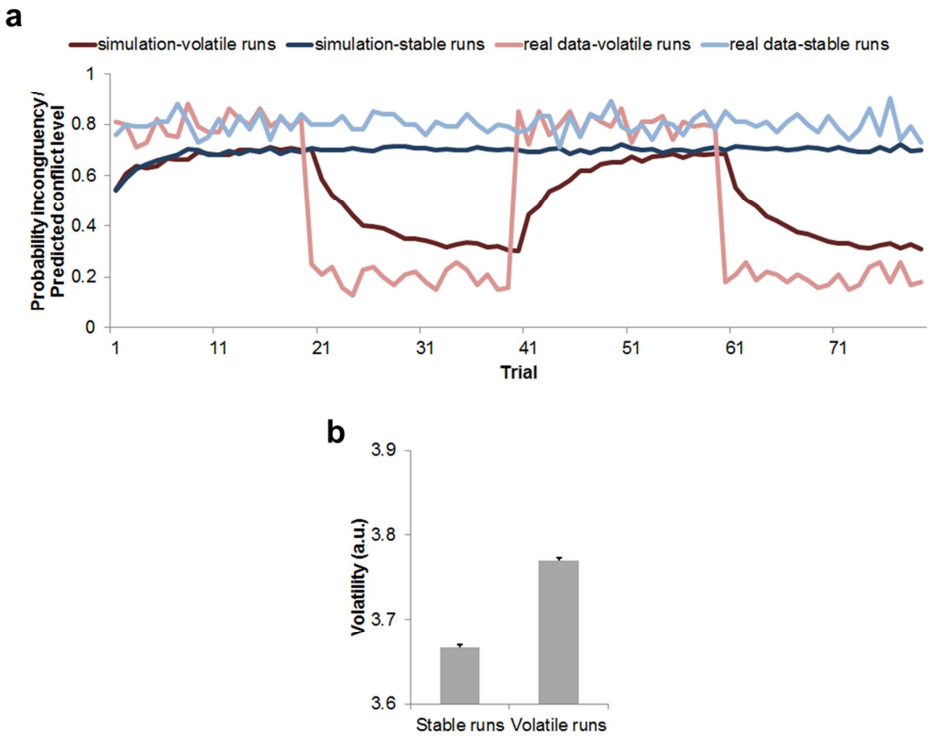


Figure 9: Simulation results. (a) Group mean time courses of proportion incongruency and predicted conflict level of actual trial sequences and model

predictions, respectively. (b) Group mean volatility and MSE, plotted as a function of run conditions.

As depicted in Figure 9a, in both conditions, the predicted conflict level tracks (the change of) the underlying proportion incongruency, suggesting that the model correctly updated its belief about the task sequences. In particular, in the volatile condition, the predicted conflict level adjusted swiftly to reflect the most-recent proportion incongruency. More importantly, the volatility estimates in the volatile condition were significantly higher than in the stable condition (t_{198} = 17.84, P < 0.0001). This result strongly supports that the claims the Bayesian model is able to detect differences in volatility caused by varying frequencies of alternating underlying proportions of incongruency.

I argued above that a fundamental limitation of reinforcement learning models with fixed learning rates is that they are unable to achieve optimal performance across experimental settings with various volatility settings. To validate this hypothesis, ninety-nine reinforcement learners with (fixed) learning rates from 0.01 to 0.99 (step size = 0.01) were trained using the trial sequences described above. To quantify performance, the square of prediction error (SPE, defined as the difference between the predicted conflict level and observed congruency) was computed for each of the reinforcement learners and the Bayesian model for each trial sequence. Two-sample t-tests of SPE between the Bayesian model and each of the reinforcement learners were conducted

within each condition, as well as across the 2 conditions (trial sequences from the 2 conditions were randomly paired).

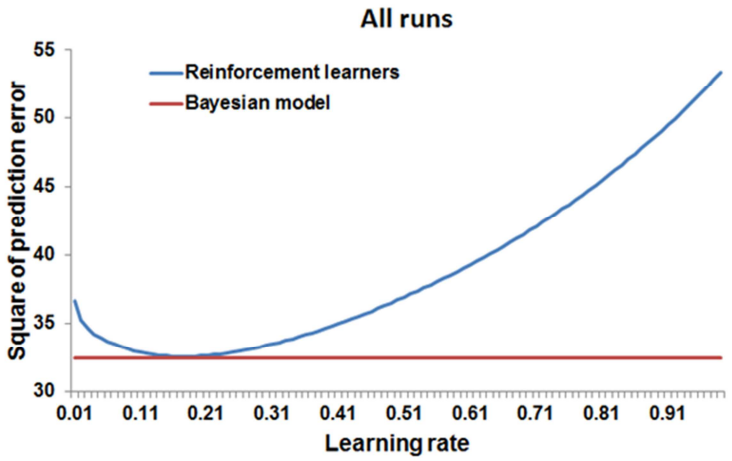
The results are depicted in Figure 10. Two patterns are revealed here. First, as expected, the learning rate achieving optimal performance in stable runs (learning rate = 0.08, SPE = 14.05) is lower than the optimal learning rate for volatile runs (learning rate = 0.25, SPE = 17.86), because stable runs require less contribution from the most-recent information in making predictions. These results replicated the ones found in Botvinick et al (2001), where a larger learning rate was selected to best account for the conflict adaptation effect than the proportion incongruency effect. It is also not surprising to find that the optimal learning rate for both conditions combined falls between these two values (learning rate = 0.18, SPE = 32.55) .

Second, in the comparison involving only the volatile runs, the Bayesian model (SPE = 17.56) outperformed all 99 reinforcement learners (Figure 10b, $P < 0.0001$ for all paired t-tests). This is attributed to the fact that the Bayesian model adaptively changes its dependence on short-term information (e.g., larger dependence when the proportion incongruency alternated, versus lower dependence when the model learned the new proportion incongruency). Although the Bayesian model was not the best model in stable runs, its performance was still better than 78 out of the 99 reinforcement learners (Figure 10c, SPE = 14.88, $P < 0.05$ for each of the 78 paired t-tests). Crucially, when both conditions were combined together to form an environment that requires flexibility in

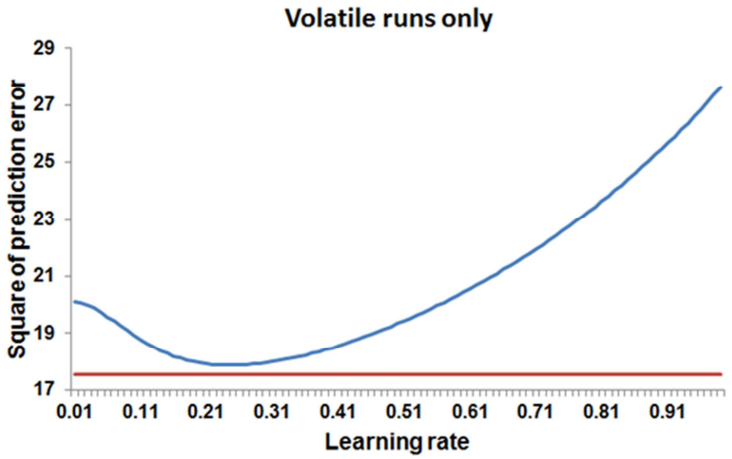
the adjustment of cognitive control, the Bayesian model performed significantly better than all 99 reinforcement learners (Figure 10a, SPE = 32.45, $P < 0.05$ for all paired t-tests). When viewing the choice of an optimal learner from all reinforcement learners as a model selection process, one can see that even with the unrealistic advantage of selecting the best-performing fixed learning rate in a post-hoc manner, the reinforcement learning algorithm still fails to outperform the Bayesian model, which predicted forthcoming congruency on-the-fly. Moreover, the reinforcement learning algorithm selected 3 distinct learning rates to best account for 3 experimental settings, while the Bayesian model worked in exactly the same way for all settings. Taken together, in this section it has been demonstrated that the Bayesian model does not only detect the difference in volatility changes between different experimental environments, the detected changes also successfully inform the model's prediction of forthcoming congruency.

The results also support the notion that the proposed experimental design is effective in constructing experimental environments to test the flexible adjustment of cognitive control. Compared to the manipulation using proportion incongruency in section 2.2.2, this approach is more effective (i.e., more statistically significant in the comparison of volatility between conditions). Thus, in the next chapters, this design is used as the main approach for manipulating volatility.

a


All runs
Reinforcement learners
Bayesian model
Square of prediction error
55
50
45
40
35
30
0.01
0.11
0.21
0.31
0.41
0.51
0.61
0.71
0.81
0.91
Learning rate

b


Volatile runs only
Square of prediction error
29
27
25
23
21
19
17
0.01
0.11
0.21
0.31
0.41
0.51
0.61
0.71
0.81
0.91
Learning rate

c

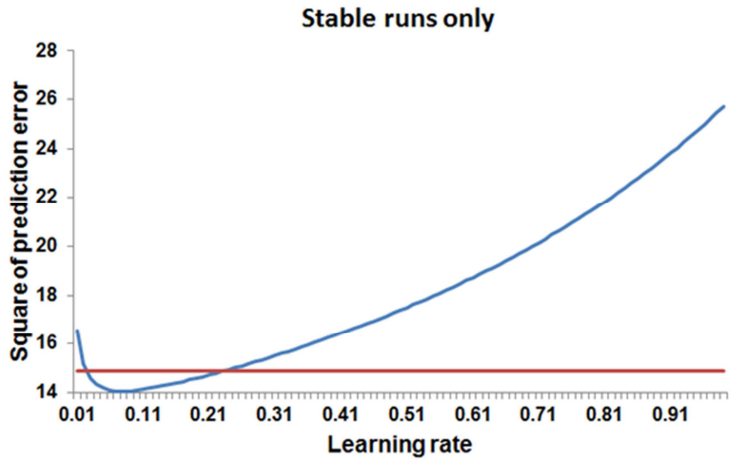

Stable runs only
Square of prediction error
28
26
24
22
20
18
16
14
0.01
0.11
0.21
0.31
0.41
0.51
0.61
0.71
0.81
0.91
Learning rate

Figure 10: Comparison of the SPE in predicting forthcoming congruency between the Bayesian model and reinforced learning models with various learning rates. (a), (b) and (c) depict comparison results based on all runs, only volatile runs and only stable runs, respectively. The SPE of the Bayesian model (red) is plotted as a baseline to facilitate visual inspection of the results.

3. The Computational Mechanisms of Cognitive Control

The aforementioned proof-of-principle validations have demonstrated that the Bayesian model is sensitive to changes in volatility. In this chapter, I describe 3 behavioral studies that further test whether this model can account for behavioral patterns observed from tasks involving flexible cognitive control.

3.1 Simulating the Short-term and Long-term Trial History Effects of Cognitive Control

In this section, I first confirm that short-term and long-term trial history effects occur simultaneously in a single empirical dataset (e.g., Torres-Quesada et al., 2013). Then I demonstrate that both effects can be simulated simultaneously using the Bayesian model. Importantly, no post-hoc optimization of parameters was necessary in the simulation.

3.1.1 Subjects

Fifty-six healthy volunteers (mean age = 26.1, 30 females) gave informed consent in accordance with institutional guidelines. All subjects were native or highly proficient English speakers and had normal or corrected-to-normal vision.

3.1.2 Stimuli and Procedure

Stimulus delivery and behavioral data collection were carried out using Presentation software (http://www.neurobs.com/). Stimuli were presented on a 19 inch LCD screen with a refresh rate of 60 Hz. Stimuli consisted of a collection of 24 black and

white photographs of male and female faces (12 each) of neutral expression that were overlaid with red gender word labels ("male" and "female"), which could be printed in lower or upper case lettering. On each trial, one face-word compound stimulus (subtending approximately 3° of horizontal and 4° of vertical visual angle) was presented against a gray background in the center of the screen. Stimuli were presented for 500 ms, followed by a jittered inter-stimulus interval (ISI) ranging from 2 to 3 s in uniformly distributed steps of 500 ms, during which a fixation cross remained on screen. Subjects performed a speeded button response that categorized the gender of the face stimulus with either index finger (for example, left-hand response to male faces, right-hand response to female faces, counterbalanced across subjects), while trying to ignore the task-irrelevant gender labels and stimulus locations. Face stimuli never repeated across adjacent trials, and the lettering alternated between lower- and upper-case across trial. A practice run was conducted before the main task to ensure subjects comprehended the task requirements.

3.1.3 Experimental Design

This task consisted of 7 runs of 4 blocks each. Each block contained 41 trials with pseudo-randomized congruency. Across all blocks, the proportion of congruent trials followed the order of approximately (deviated by 1 trial, or ~2.4%) 15%, 35%, 65%, 85%, 75%, 50% and 25%, repeated for 4 times over the 7 runs. Within each block, the proportion incongruency remained constant. The starting proportion incongruency and

the order of the sequence were counter-balanced across subjects. To model both the conflict adaptation and proportion incongruency effects, RT data was analyzed using a 7 (proportion incongruency) × 2 (previous congruency) × 2 (current congruency) factorial design.

3.1.4 Data Analysis

For the behavioral data, the mean RT was computed in each subject for each of the experimental cells, excluding incorrect and post-error trials, as well as RT values that deviated >2 SDs from an individual subject's grand mean. The trimmed RT values were then averaged across subjects and entered into repeated measures 3-way analyses of variance (ANOVAs) with the factors described above. For the simulation data, the trial sequences observed by the subjects were fed to the model to produce trial-by-trial estimates of volatility and predicted conflict level. Then, for each experimental cell, the mean volatility and predicted conflict level were computed, excluding trials that were excluded in empirical data analysis. Finally, to link model predictions to the empirical data, a general linear model (GLM) was constructed using group means of the parameter estimates (28 conditions), which were then fit to the group mean of RTs. Specifically, this GLM contained 6 regressors (i.e. free parameters), namely the volatility, the predicted conflict level, and the grand mean, separately for congruent and incongruent trials. Note that this fitting procedure is not geared at finding the optimal parameters for the model. Rather, the purpose of this fitting was similar to within-trial

simulation, or in other words, to quantify how predictions made prior to a trial influence the information processing during that trial, as reflected in the RTs.

3.1.5 Results and Discussion

The subjects performed the task with high accuracy (mean = 92.9%). The 3-way ANOVA on empirical RTs showed a significant effect of current congruency ($F_{1,55}$ = 57.07, $P < 0.001$), due to longer RTs in incongruent trials (592 ± 11 ms) than in congruent trials (569 ± 10 ms). An interaction between proportion incongruency and current congruency was also found (the proportion congruent effect, $F_{1,55} = 2.73$, $P < 0.03$), driven by a decrease in interference effects as the proportion of incongruent trials increased (Figure 11a). There was also an interaction between previous and current trial congruency (the conflict adaptation effect, $F_{1,55} = 5.03$, $P < 0.03$), driven by a larger interference effect (26 ± 3 ms) in post-congruent trials than in post-congruent trials (16 ± 3 ms; Figure 11b). Thus, the behavioral results replicated a large literature on the proportion incongruency (for review, see Bugg & Crump, 2012) and the conflict adaptation effect (for review, see Egner, 2007), as well as previous findings of these two effects occurring simultaneously in the same data set (Torres-Quesada et al., 2013). As can be seen in Figure 11c, both effects were successfully simulated using theBayesian model. Specifically, the model predicted an interference effect (congruent trials: 569 ms; incongruent trials: 592 ms), a decreased interference effect as the proportion of incongruent trials increased (Figure 11d), and a higher interference effect in post-

congruent trials (27 ms) than in post-incongruent trials (18 ms). These simulation results suggest that the Bayesian model is able to simultaneously account for both long-term and short-term effects of cognitive control. These fits were achieved using a single control mechanism with a flexible learning rate rather than the dual mechanism structure of De Piasapia & Braver (2006) or the separate fits with different learning rates as applied by Botvinick and colleagues (2001). Moreover, the data fits were derived from on-the-fly simulations and not based post-hoc setting of learning rate parameters.

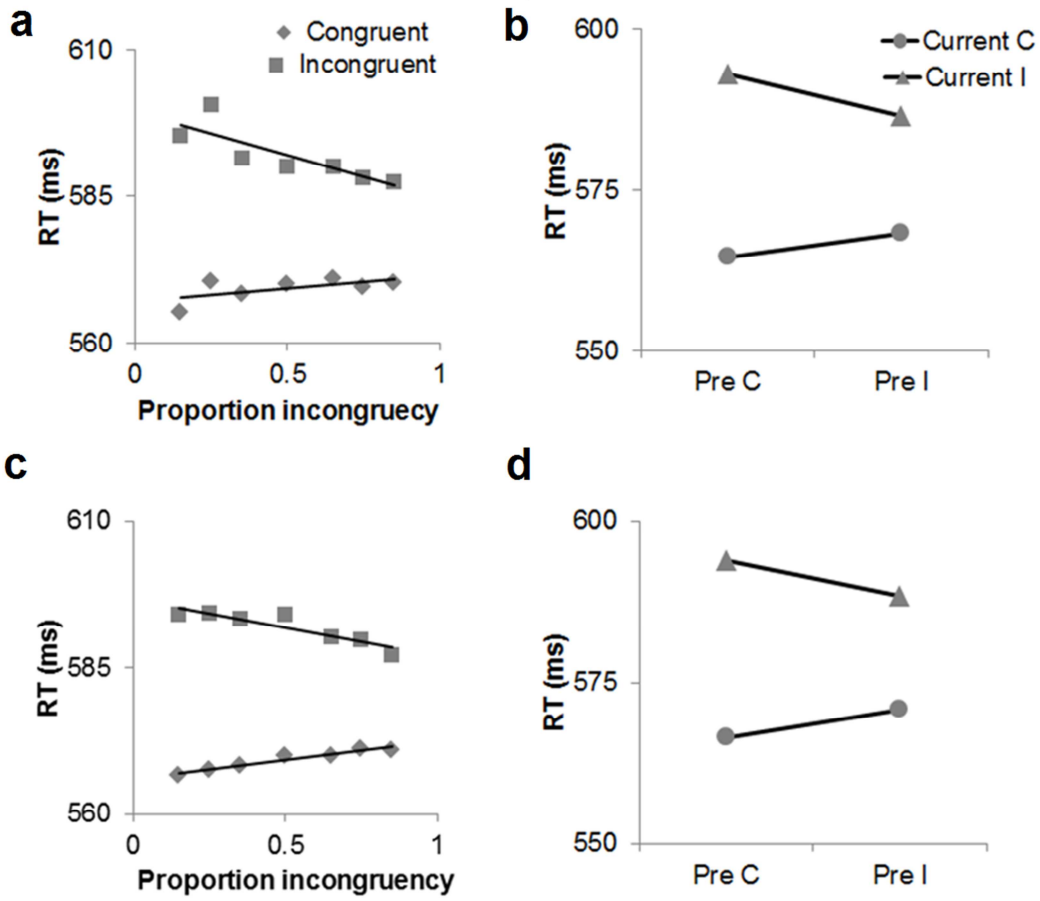


Figure 11: Empirical and simulated effects of congruency, proportion incongruency, and conflict adaptation. (a) Empirical proportion incongruency effect,

with RT plotted as a function current trial congruency and the block-wise proportion of incongruent trials. (b) Empirical conflict adaptation effect, with RT plotted as a function of current and previous trial congruency. (c) Simulated proportion incongruency effect, plotted in the same way as in (a). (d) Simulated conflict adaptation effect, plotted in the same way as in (b). Pre C/Pre I = Preceded by a congruent/incongruent trial; Current C/Current I = current trial is congruent/incongruent.

3.2 Simulating the Flexibility of Conflict-control

To further demonstrate the model's ability to simulate the flexibility of cognitive control, a second experiment was conducted, in which I created two environments with different dependence on short-term and long-term information. I show that the model can successfully simulate the behavioral patterns observed in the empirical data.

3.2.1 Subjects

Forty-six healthy volunteers (mean age = 19.9, 33 females) gave informed consent in accordance with institutional guidelines. All subjects were native or highly proficient English speakers and had normal or corrected-to-normal vision.

3.2.2 Stimuli and Procedure

The same stimuli and basic task procedure was used as the one described above in section 3.2.

3.2.3 Experimental Design

This task consisted of 4 runs of 9 blocks each. Each block contained 20 trials with pseudo-randomized congruency. The first block had 50% congruent trials and served as

a burn-in block to bring the predictions to the same baseline at the beginning of each run. To create experimental environments that differ in their dependence on long-term and short-term trial history, a run could be either volatile (the proportion incongruency altered between 20% and 80% every block) or stable (the proportion incongruency remained either 20% or 80% for all 8 post-reset blocks). The order of volatile and stable runs was counter-balanced across subjects. This manipulation resulted in a 2 (volatile / stable) × 2 (proportion incongruency) × 2 (current trial congruency) factorial design.

3.2.4 Data Analysis

The same analyses were applied as described in section 3.1.4. Note that trials in burn-in blocks were also given to the model, so as to also generate a reset of trial-by-trial estimates of volatility and predicted conflict level in the model at the beginning of each run. However, reset block trials were excluded from further analyses.

3.2.5 Results and Discussion

Participants performed the task with high accuracy (mean = 92.8%) in this task. The 3-way ANOVA on empirical RTs again revealed a significant effect of current trial congruency ($F_{1,45}$ = 49.3, P < 0.001), due to longer RTs in incongruent trials (549 ± 13 ms) than in congruent trials (522 ± 11 ms). The proportion incongruency effect was also found, reflected in a significant interaction between proportion incongruency and current trial congruency ($F_{1,45}$ = 10.5, P = 0.002). This effect was driven by a larger interference effect in 80% congruency blocks (33 ± 5 ms) than in 20% congruency blocks

(22 ± 4 ms). Importantly, asignificant main effect of volatility was observed ($F_{1,45}$ = 4.5, P = 0.04), due to longer RTs in volatile runs (540 ± 12 ms) than in stable runs (531 ± 12 ms). Note that this main effect was not driven by "outliers" in experimental cells, because no interactions involving volatility and any of the other factors were found. Furthermore, a trend for longer RTs in volatile compared to stable environments can be observed in all 4 current trial congruency × proportion incongruency conditions (Figure 12a).

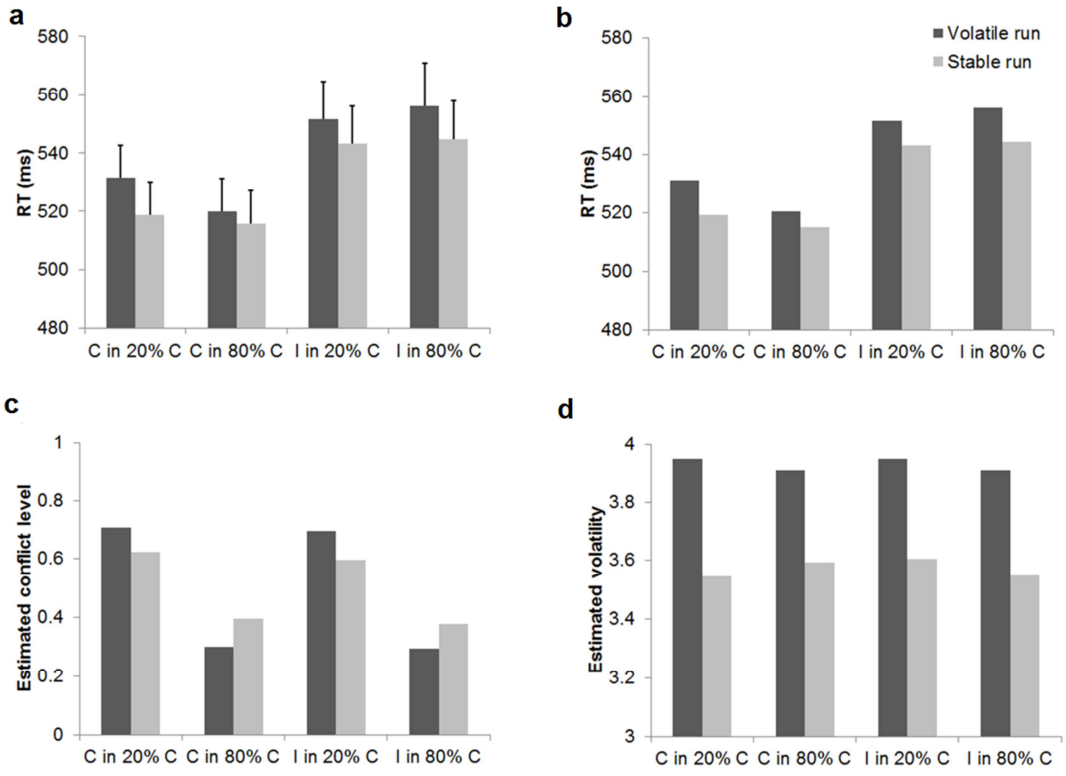


Figure 12: Empirical and simulated effects of congruency, proportion incongruency in an environment of changing volatility. All quantities are plotted as a function of volatility, proportion incongruency and congruency at the current trial. (a) Empirical reaction times and their standard errors. (b) Simulated reaction time. (c) Estimated predicted conflict level from the Bayesian model. (d) Estimated volatility from the Bayesian model in arbitrary units. C/I in 20%/80% C = congruent/incongruent trials in a block of 20%/80% congruent trials.

This pattern of RTs was again successfully simulated using the model (Figure 12b): firstly, the model recapitulated the main effect of current congruency (congruent trials: 522 ms; incongruent trials: 549 ms); secondly, it simulated the proportion incongruency effect (inference effect in high proportion incongruency blocks: 33 ms; inference effect in low proportion incongruency blocks: 22 ms); and lastly (and most importantly), it predicted longer RTs in volatile runs (540 ms) than in stable runs (531 ms). I contend that this "volatility cost" performance pattern can in fact only be accounted for by using a model that estimates the volatility of the environment. Consider a model with no information about volatility (e.g. having a fixed learning rate): in such a model, each trial's contribution to the prediction of predicted conflict level is a constant. Furthermore, the trial's contribution to the prediction of all forthcoming trials is also a nearly constant, because its influence decays continuously with a constant discounting rate after each trial, and approaches zero in a relatively short period of time (except for extremely low learning rates, which are unrealistic given the commonly observed short-term effects of cognitive control). As a consequence, using a fixed learning rate, two sequences of trials with the same proportion incongruency and number of trials will produce the same sum of conflict-level estimates across all trials, regardless of the volatility of those sequences. Indeed, even as shown in the model, estimates of predicted conflict level displayed an interaction between volatility and

proportion incongruency (Figure 12c). Thus, it is impossible to account for the pattern of slower RTs in volatile compared to stable runs in the empirical dataset when using only predicted conflict level estimates; they would have to be combined with estimates of volatility (Figure 12d) to simulate the empirical data pattern. To further illustrate this point, I performed an additional analysis to simulate the results in this experiment using a reinforcement learning algorithm. Specifically, I created 3 learners with different learning rates: 0.05 (high dependence on long-term information), 0.5 (balanced dependence on long-term and short-term information) and 0.95 (high dependence on short-term information). The learning was conducted using the following equation:

$$f_{i+1} = f_i \times (1 - \alpha) + o_i \times \alpha \quad \text{(Eq 13)}$$

Where f_{i+1} is the predicted conflict level at trial i+1, α is the learning rate, and o_i is the observed congruency. If any of these learners were able to account for the behavioral pattern observed in experiment 2, there should be a significant difference in f_{i+1} between the 2 volatility levels, because f_{i+1} is the only output of these learners. However, in none of the 3 learners was such significant a difference (all Ps > 0.2) observed. Thus, in the absence of a (volatility-modulated) flexible learning rate, the model is unable to account for the behavioral patterns obtained across different task settings.

Note that the predicted conflict levels display the inverse pattern of empirical RTs, e.g., for incongruent trials higher estimated (or predicted) conflict levels are

predictive of faster RTs. This pattern essentially corresponds to the classic, intuitive explanation of the empirical proportion incongruency effect, namely, that control is higher in conditions where conflict is frequently encountered (e.g., Carter et al., 1998).

A few additional points should be noted in the interpretation of these data. First, both conflict adaptation and proportion incongruency effects have at times been argued to exclusively reflect associative processes related to the particular stimulus and response features of the task (e.g., Mayr et al., 2003; Hommel et al., 2004; Schmidt and Besner, 2008). By contrast, the present results show that both of these effects can be faithfully captured by a model that does not consider specific stimulus or response features at all – it only learns about the incidence of congruent and incongruent stimuli. This documents that, at least in principle, learning of specific physical stimulus and response properties is not a necessary precondition for producing these effects. Second, while this main effect of volatility can be quantitatively accounted for by the Bayesian model (the manipulation of volatility was captured by the volatility variable, as can be seen in Figure 12c), the model architecture itself does not necessitate such an effect. In other words, the model could equally well fit behavioral data in the absence of a main effect of volatility (as observed in other, unpublished observations).

Nevertheless, this does of course not mean that the empirical data themselves were not potentially subject to such lower-level learning effects. It is unlikely that such processes contributed in a substantial manner to the present results, however, for the

following reasons. First, in order to prevent trial-by-trial priming effects at the level of physical stimulus features (Hommel, Proctor, & Vu, 2004; Mayr, Awh, & Laurey, 2003), face stimuli in the present experiments never repeated across successive trials, and the lettering of the distracter labels alternated between lower- and upper-case across trials. Second, in order to minimize the possibility that proportion incongruency effects in the protocol would be mediated by subjects associating specific face stimuli with a particular response (e.g., the gender-congruent response in high proportion incongruency blocks), the stimuli included a large number (24) of unique facial identities (cf. (Bugg & Hutchison, 2013)). Nonetheless, this leaves the possibility that subjects may use the contingency between the distracter (in this case, the gender word) and the response to guide their action selection (Bugg, 2012; Bugg & Chanani, 2011; Schmidt & Besner, 2008). For example, in an environment of high proportion incongruency, the gender word is highly predictive of the correct response. However, note that in volatile runs, this contingency changes every 10 occurrences for each word (on average), leading to a less predictive word-response association than in stable runs, where the contingency remains unchanged. Therefore, if distracter-response contingency were a major contributing factor to the proportion incongruency effect in the data, this effect should be modulated by the contingency's predictive power (i.e., volatility), resulting in a 3-way interaction between volatility, proportion incongruency and current-trial congruency. However, such an interaction was not observed ($F_{1,45} = 2.9$, n.s.). More specifically, the

contingency account predicts that contingency with higher predictive power (i.e., the stable runs) should evoke larger proportion incongruency effects than low-contingency conditions (i.e., the volatile runs). However, numerically, the opposite is true for the present data (volatile runs: proportion incongruency effect = 16 ± 5 ms; stable runs: proportion incongruency effect = 4 ± 4 ms; $t_{45} = 1.7$, n.s.). Thus, contingency learning seems highly unlikely to have contributed to the empirical proportion incongruency effects that the model simulated.

In sum, I showed that a Bayesian model that learns to predict control demand using a flexible, volatility-driven learning rate, can account for simultaneously occurring conflict adaptation, proportion incongruency, and volatility effects, without the need for multiple controllers or post-hoc fit-derived learning rate parameters. I conclude that this model represents a promising new application of a Bayesian approach to exploring computational mechanisms of cognitive control, in particular with respect to simulating the flexibility that is required of control processes wielded in a changing environment.

3.3 Simulating the Flexibility of Conflict-control Using Model-based Analysis

The two studies above have established the validity of the Bayesian model in accounting for behavioral data from tasks that requires flexible cognitive control. The study presented in this section extends those studies in two ways: (1) replicating the experiment in section 3.2 with a different timing setting by adopting jittered ISIs that is more suitable for event-related fMRI studies, and (2) by introducing a model-based,

trial-based analysis method that overcomes several limitations in the conventional, condition-based ANOVA I used above.

3.3.1 Subjects

Fifty-five subjects gave informed consent online in accordance with institutional guidelines and participated in this study through Amazon Mechanical Turk (AMT), which is an online cloud-sourcing platform that can be adapted for fast and efficient data acquisition for behavioral tasks (see Appendix A for details). Two subjects were excluded from analysis due to chance-level accuracy. The remaining 53 subjects self-reported demographic information online (36 females; mean age, 32).

3.3.2 Stimuli and Procedure

The stimuli used in this study were the same as the ones described above in section 3.2.2, except that two more distracter word labels ("man" and "woman") were added, to further reduce the efficacy of employing a contingency learning strategy during this task. Stimulus delivery and response recording were carried out by JavaScript. The size of face images were fixed at 216 × 270 pixels and always presented in the center of the browser window against a grey background. Lower-case and upper-case distracter words were presented in font sizes of 80 and 60 points respectively to match the space they covered on the screen. Stimuli were separated by exponentially jittered ISIs (range = 4-6s, step size = 1s).

Fifty-five online assignments of this task were posted simultaneously on AMT. Potential participants interested in this task were able to first read the online instructions and complete a brief practice session consisting of 24 trials. After the practice session, they had the option to formally participate in this study (i.e., to take one of the 55 assignments from AMT). The subjects that agreed to participate in this study then gave informed consent and filled out a demographic survey. They then proceeded to the task, which was hosted on Dropbox (https://www.dropbox.com/). After the subjects completed the task, the JavaScript submitted their responses along with stimulus delivery information to AMT. Finally, the subjects were compensated at the rate of ~$3 per hour through the AMT payment system. In accordance with AMT policy and institutional guidelines, a payment was made if (1) a subject finished the task and submitted their responses and stimulus delivery information and (2) the performance met a pre-specified requirement.

3.3.3 Experimental Design

The experimental design was identical to the one described above in section 3.2.3, except that the experiment consisted of 8 runs, each of which had five blocks, including a burn-in block. Both of the first and the last four runs contained two volatile and two stable runs (one for each proportion incongruency). The order of volatile and stable runs was counter-balanced within and across subjects.

3.3.4 Data Analysis

First, the analysis described above in section 3.2.4 was here repeated as a comparison to the previous study. Incorrect trials, post-error trials, outlier trials (RT values that deviated >2.5 SDs from an individual subject's grand mean), and post-outlier trials were excluded from further analysis. In addition, a model-based, trial-based analysis was performed. Specifically, the sequence of trial congruency (concatenated across runs) experienced for each subject was processed by the Bayesian model to generate trial-based estimates of volatility and predicted conflict level. Variable estimates and congruency for excluded trials were discarded. The remaining estimates were grouped into a chronological vector for each model variable. These vectors were then normalized and multiplied to form 7 variable vectors (volatility, predicted conflict level, congruency, and their 2-way and 3-way interactions). Subsequently, these vectors were grouped, along with a constant vector, to form a general linear model (GLM). To test the effect of each of the 7 variable vectors while ensuring that effects were not confounded by shared variance with any of the other variables, the "test variable" vector was first regressed against the other variable vectors. The residual resulting from this regression (after removal of all shared variance with the other variables) was then fit as a predictor to the RT vector, along with other six variable vectors in the GLM as nuisance effects. RTs were normalized within subjects prior to regression to remove potential biases due to individual variance in RT magnitude. The resulting fitting coefficient for

the test variable vector was then tested against 0 using a one-sample t-test across subjects.

3.3.5 Results and Discussion

Participants performed the task with high accuracy (mean = 96.5%). The 3-way ANOVA (condition-based analysis) on empirical RTs again revealed a significant effect of current trial congruency ($F_{1,52}$ = 89.0, P < 0.001), due to longer RTs in incongruent trials (550 ± 19 ms) than in congruent trials (511 ± 18 ms). The proportion incongruency effect was also found, reflected in a significant interaction between proportion incongruency and current trial congruency ($F_{1,52}$ = 4.79, P = 0.033). This effect was driven by a larger interference effect in 80% congruency blocks (43 ± 5 ms) than in 20% congruency blocks (34 ± 4 ms).

Similarly, the trial-based analysis also revealed a significant effect of current trial congruency (t_{20} = 9.1, P < 0.0001) and a significant negative interaction between predicted conflict level and current trial congruency (t_{20} = 2.1, P = 0.04). The negative direction of the interaction conforms to the condition-based analysis, suggesting reduced interference effect in trials with high predicted conflict level.

Both analyses revealed significant effects of interference and adjustment of cognitive control (proportion incongruency × congruency interaction and predicted conflict level × congruency interaction in the condition-based and trial-based analyses,

respectively). Thus even with the longer ISIs needed for fMRI scanning, this design is still able to elicit these classic behavioral phenomena of cognitive control.

Despite of the similarity of results from the two analyses, it should be noted that the nature of these analyses differs fundamentally. Specifically, compared to the conventional condition-based approach, this approach has three advantages: (1) unlike the condition-based approach, the trial-based nature of the present approach takes the trial-by-trial variance into consideration, thus making it more sensitive in theory; (2) the residual analysis ensures that any variance explained can only be attributed to the specific variable / interaction tested, because any variance shared with other variable vectors has been removed. Thus the trial-based analysis is a more stringent test than the condition-based approach; (3) most importantly, one crucial goal in this dissertation is to unify different forms of adjustment of cognitive control that rely on different integration of short-term (e.g., conflict adaptation effect) and long-term (e.g., proportion incongruency effect) information. However, the condition-based approach fails to integrate these, because the previous trial congruency and the proportion incongruency must be modeled as separate factors in a factorial design. From the perspective of the Bayesian model, this integration is natural because the predicted conflict level is already an integration of both short-term and long-term information. Its interaction with congruency is in turn a joint effect of both the conflict adaptation effect and the proportion incongruency effect.

Moreover, despite adding two more distracter labels to further discourage the use of a contingency learning strategy, the still-significant behavioral effects in both analyses suggests that these effects were not due to learning based on features of the stimuli but reflected genuine adjustments of cognitive control. To further test this claim, I compared the effect of proportion incongruency × congruency interaction between the data from section 3.2 and this section. The contingency learning argument would predict that this interaction should be higher in the data from section 3.2 than the data from this study. However, there was no significant difference of the proportion incongruency × congruency interaction between these two datasets (t_{99} = -0.28, n.s.), again suggesting that the effects were mainly driven by the adjustment based on conflict or trial congruency.

Another interesting point to note in the present data set is that the main effect of volatility was not significant, as confirmed by both the condition-based and trial-based analyses. Because a main effect of volatility was observed in two independent datasets (see section 3.2 and the appendix), it is unlikely that this main effect represents a type II error. Why did we not observe this main effect of volatility in the present data set then? Compared to those two datasets, the present task has two key differences: there were two additional distracter labels and longer ISIs. The discussion above has discounted the possibility that the former difference induced any change in behavioral patterns. Thus, I speculate that the longer ISIs might have caused the absence of the main effect of

volatility. Short ISIs may encourage subjects to make faster, premature responses (e.g., select an action before sufficient evidence has been accumulated from processing the stimulus). Low volatility also indicates better precision of predicted congruency (Behrens et al., 2007), thus allowing for less adjustment of cognitive control *within* a trial, which may result in faster RT. Nevertheless, future studies are necessary to examine this speculation.

Taken together, this study has demonstrated that the experimental design is suitable for an fMRI study and validated the trial-based analysis.

4. The Neural Mechanisms of Flexible Cognitive Control

Although the neural mechanisms underlying the adjustment of cognitive control have been extensively inspected, few studies have targeted the *flexibility* of such adjustment. Most of these studies mainly focused on either the adjustment depending on short-term information (e.g., the conflict adaptation effect, (M. Botvinick et al., 1999; Durston et al., 2003; Egner & Hirsch, 2005a, 2005b; Kerns, 2006; Kerns et al., 2004)), or the adjustment depending on long-term information (e.g., the proportion incongruency effect, (Grandjean et al., 2012; Krug & Carter, 2012; Sohn, Ursu, Anderson, Stenger, & Carter, 2000; Wilk, Ezekiel, & Morton, 2012)). The only neuroimaging study that investigated both effects modeled these effects separately due to the limitation of condition-based analysis and did not find any brain regions significantly displaying either effect (Torres-Quesada, Franziska, Funes, Lupianez, & Egner, 2014). Nevertheless, many of the studies above found adjustment-related activation in the ACC and dlPFC, regardless of whether they investigated the short-term or long-term effects. Thus there may be a unified neural mechanism supporting both effects. Or more broadly, this mechanism may support flexible adjustment of cognitive control by integrating information sampled from various temporal resolutions.

The Bayesian model provides a natural integration of long-term and short-term information and thus represents a novel tool for the quest of exploring the neural mechanism of flexible adjustment of cognitive control. Furthermore, the mechanisms of

the Bayesian model generates a few specific predictions regarding how this integration is carried out: (1) the updating of predicted conflict level is modulated by volatility; (2) cognitive control is regulated by the predicted conflict level; (3) the updating of volatility is guided by the (unsigned) prediction error of conflict. In the following sections of this chapter, I present an fMRI study that investigated the flexible adjustment of cognitive control using the Bayesian model and an experimental manipulation that created two environments with different reliance on long-term and short-term information (based on the task design described above under section 3.3.3).

4.1 Methods

4.1.1 Materials

Twenty-one healthy, right-handed volunteers (8 females, mean age = 26 years) gave informed consent in accordance with institutional guidelines. All subjects were native or highly proficient English-speakers and had normal or corrected-to-normal vision.

4.1.2 Apparatus and Stimuli

The stimuli used were identical to those used in section 3.3.2. Stimulus delivery and behavioral data collection were carried out using Presentation (http://www.neurobs.com/). Visual stimuli were presented on a back projection screen viewed via a mirror attached to the scanner headcoil, and responses were collected

using an MRI-compatible button box. The stimuli subtended approximately 3° of horizontal and 4° of vertical visual angle.

4.1.3 Procedure and Task Design

The procedure and task design were the same as the ones used in section 3.3.3, except that the length of burn-in blocks were reduced to 16 trials.

4.1.4 Image Acquisition and Preprocessing

Images were acquired parallel to the AC-PC line on a 3T GE scanner (Milwaukee, WI). Structural images were scanned using a T1-weighted SPGR axial scan sequence (146 slices, slice thickness = 1mm, TR = 8.124ms, FoV = 256mm * 256mm, in-plane resolution = 1mm * 1mm). Functional images were scanned using a T2*-weighted single-shot gradient EPI sequence of 39 contiguous axial slices (slice thickness = 3mm, TR = 2s, TE = 28ms, flip angle = 90 °, FoV = 192mm * 192mm, in-plane resolution = 3mm * 3mm). Functional data were acquired in 8 runs of 240 images each. Preprocessing was done using SPM8 (http://www.fil.ion.ucl.ac.uk/spm/). After discarding the first five scans of each run, the remaining images were realigned to their mean image and corrected for differences in slice-time acquisition. Each subject's structural image was co-registered to the mean functional image and normalized to the Montreal Neurological Institute (MNI) template brain. The transformation parameters of the structural image normalization were then applied to the functional images. Normalized functional images were kept in their native resolution.

To gauge the trial-wise activation in the fMRI data, a task model was built for each run. Similar to the behavioral studies reported above, error trials, post-error trials, outliers (i.e., trials with RTs greater than 2.5 standard deviations from the mean), post-outlier trials, and burn-in trials were excluded from further analyses. A task model consisted of regressors representing the onset of each non-excluded trial, along with 5 nuisance regressors representing the onsets of each type of excluded trials and 2 other nuisance regressors separately encoding onsets of left and right button-presses. This task model was then convolved with SPM 8's canonical hemodynamic response function. The convolved tasked model was appended by regressors representing head motion parameters and the grand mean of the run to form a design matrix, against which the normalized functional images were regressed. The resulting activation maps were then concatenated across runs. As the final output of preprocessing, each grey matter (GM) voxel obtained an activation vector that chronologically represented the activation level at each trial. These activation vectors were used for fMRI analyses below.

4.1.5 Data Analysis

The trial sequences of congruency exposed to each subject was processed by the Bayesian model to generate trial-by-trial estimates of volatility and predicted conflict level. These estimates first underwent a sanity check (see next section), and then entered the analyses for behavioral data and fMRI data. Estimates corresponding to excluded trials were also discarded from further analyses.

4.1.5.1 Sanity Check for Model Estimates

Similar to the analysis in section 2.2.3, to examine whether the Bayesian model is sensitive to the manipulation of volatility in this task, the mean volatility in volatile runs was compared to the mean volatility in stable runs using a paired t-test across subjects.

4.1.5.2 Behavioral Analysis

The behavioral analysis was conducted using both the conventional 3-way ANOVA (volatility × proportion incongruency × current trial congruency) described in section 3.2.5 and the trial-based analysis using residual of variable vectors as described in section 3.3.4.

4.1.5.3 Searchlight-based Analyses Investigating Neural Representation of Model Variables

For determining the encoding of model variables and (their interactions), we examined two possible coding schemes: a "homogeneous", univariate scheme, where information is encoded by local voxel populations with similar response properties (e.g., a group of voxels whose fMRI signal magnitudes all scale positively with predicted conflict), and a "heterogeneous", distributed scheme, where information is encoded in multivariate activation patterns over local voxel populations(Turk-Browne, 2013). To account for these schemes, a univariate analysis and a multi-variate pattern analysis (MVPA) were conducted in parallel (Clithero, Carter, & Huettel, 2009). Both analyses were carried out with a searchlight approach (Kriegeskorte, Goebel, & Bandettini, 2006) that scanned through small clusters (radius: 2 voxels) of GM voxels . Within each

searchlight, the univariate analysis assessed homogeneous encoding by amplitude by fitting the variable vector to the searchlight mean activation vector. The univariate analysis was conducted via a two-fold averaging approach between the first and last four scanning runs. Each half contained two volatile runs and two stables runs (one for each proportion of incongruent trials) and was tested separately. The results were averaged across the two halves to reduce the impact of outliers. The MVPA quantified distributed encoding by the amount of signal variance in a model variable vector that could be explained by the mean-centered activation vectors from all voxels in a given searchlight through linear regression. The removal of searchlight mean signals renders the MVPA independent from the univariate analysis. Over-fitting was controlled for by a two-fold cross-validation scheme between the first and last four scanning runs. To ensure the unique attribution of fMRI signal to a given variable, all other variable and interaction vectors were used as nuisance variables in both analyses.

For each analysis, the quantification of encoding was mapped to the center voxel of each searchlight to form a spatial map of information content. The map was then smoothed using a Gaussian kernel of 6mm (2 voxels) radius. One-sample t-tests were then conducted on the maps across individual maps to test for group-level effects of homogeneous encoding and distributed encoding.

Statistical results were corrected for multiple comparisons at $P < 0.05$ for combined searchlight classification accuracy and cluster extent thresholds, using the

AFNI ClusterSim algorithm. 10,000 Monte Carlo simulations determined that an uncorrected voxelwise p value threshold of < 0.01 in combination with a searchlight cluster size 30 to 35 searchlights (depending on the specific contrast) ensured a false discovery rate of < 0.05.

4.1.5.4 Inspecting the Modulation of Volatility on the Learning Rates of Predicted Conflict Level

The function of volatility in the Bayesian model suggests that the volatility belief modulates the updating (or learning) of the neural coding of predicted conflict level. To test this hypothesis, I first extracted the neural coding of predicted conflict level with no assumption from the Bayesian model. For each searchlight in the caudate ROI (see below), its GM voxels' activation vectors were applied to fit the variable vector of congruency using linear regression, and the fitted activation vector (i.e., the regression coefficient-weighted sum of activation vectors) was used as the model-free activation vector of predicted conflict level. This is because that the linear regression minimized the sum of square error between the observed congruency and the predicted conflict level, and hence the fitted activation vector can be considered as the best approximation of congruency (or prediction of conflict level) based on activation vectors in the searchlight. Consistent with the Bayesian model, in the linear regression, congruent and incongruent trials were represented as 0 and 1 respectively. To constrain the predicted conflict level between the range of 0 and 1, off-limit values (< 0 or > 1) were set to their closest limits in the fitted activation vector.. Finally, the ROI mean model-free activation vector of

predicted conflict level averaged was used to estimate learning rates using the reinforcement learning algorithm.

The prediction error at each trial in the fitted activation vector was then calculated as $o_i - \widetilde{f_i}$, where $\widetilde{f_i}$ is the activation in the fitted activation vector at trial i. The updating was calculated as $\widetilde{f_{i+1}} - \widetilde{f_i}$. The learning rate α can then be estimated across trials by a linear regression of:

$$\widetilde{f_{i+1}} - \widetilde{f_i} = a(o_i - \widetilde{f_i}) \quad \text{(Eq 14)}$$

To compare the learning rates between trials with high volatility and trials with low volatility, the ROI mean model-free activation vector of predicted conflict level was divided into 2 vectors based on a mean-split of the volatility estimates. For each vector, the learning rate was estimated using the method described above. The estimated learning rates were then compared between the two vectors across subjects via a paired t-test.

To further test the modulation of volatility on learning rates, an updated linear model was applied to the neural coding of predicted conflict level:

$$\widetilde{f_{i+1}} - \widetilde{f_i} = a\left(o_i - \widetilde{f_i}\right)$$
$$\alpha = 1 + \beta v_{i+1} \quad \text{(Eq 15)}$$

This new linear model defines a linear correlation between the learning rate α and the volatility v_{i+1}. Importantly, the sign of modulation index β is predicted to be

positive by the Bayesian model, indicating a larger learning rate in a more volatile environment. This linear model was then applied to estimating β, which was further tested against 0 using a one-sample t-test across subjects.

4.1.5.5 Using the Interaction between Predicted Conflict Level and Congruency as a Measure of Prediction Error of Congruency

After normalization, the congruent and incongruent trials were represented as -1 and 1, respectively. The normalized predicted conflict level f_i represents the (belief of the) probability of encountering an incongruent trial. Assuming f_i is centered as re-scaled between -1 and 1, the unsigned prediction error of congruency can then be quantified using the negative of the interaction term between predicted conflict level and congruency, $-f_i \times o_i$.The unsigned prediction error of congruency, formally formulated as:

$$\begin{cases} 1 - f_i, if\ o_i = 1 \\ 1 + f_i, if\ o_i = -1 \end{cases} \quad \text{(Eq 16)}$$

can be further re-formulated to $1 - f_i \times o_i$. In the context of a regression analysis, the constant 1 can furthermore be discarded without affecting the results. In fact, although the normalized f_i was not re-scaled between -1 and 1 in the analyses, $-f_i \times o_i$ can still be used as prediction error multiplied by the re-scaling factor.

4.2 Results

In the present study, the model was employed to generate trial-by-trial variable parameter estimates that allow following analyses to delineate the neural substrates

mediating these putative control computations. Specifically, I gauged neural dynamics relating to three core hypothesized mechanisms inherent in the model: (1) updating of predicted conflict level should be modulated by beliefs about volatility (2) predicted conflict level drives cognitive control, and therefore modulates the effect of congruency on neural processing; (3) congruency prediction error drives the adjustment of volatility estimates.

4.2.1 The Bayesian Model Captures Behavioral Patterns in Flexible Adjustment of Cognitive Control

Subjects performed the task with high accuracy (mean accuracy = 94.2%). A three-way ANOVA (volatility × proportion of incongruent trials × current trial congruency) revealed a main effect of proportion of incongruent trials ($F_{1,20}$ = 4.77, P = 0.041) due to better accuracy in blocks of 20% incongruent trials (94.8% ± 1.0) than blocks of 80% incongruent trials (93.2% ± 1.2), and a main effect of current trial congruency ($F_{1,20}$ = 9.15, P = 0.007), caused by better accuracy in congruent (94.9% ± 1.2) than incongruent (93.1% ± 1.0) trials. In RT, a three-way ANOVA detected a significant main effect of current trial congruency ($F_{1,20}$ = 20.4, P < 0.0001), driven by a slower RT in incongruent trials (456ms) than congruent trials (416ms). The Bayesian model was then used to simulate behavioral data on a trial-by-trial basis (Figure 13b). In brief, the individual subjects' trial sequences were processed by the model to produce trial-by-trial estimates of volatility and predicted conflict level. The estimates of predicted conflict level tracked the time course of the underlying proportion congruency very closely

(Figure 13c), and the model estimate of volatility was higher for the volatile than the stable runs (Figure 13d, paired t-test, F_{20} = 16.38, P < 0.0001). These results indicate that the model beliefs successfully tracked the experimental manipulations. I next employed the three model variables (trial-wise estimates of volatility and predicted conflict level, and the observed congruency) and their interactions (i.e., a total of seven variables, resembling a three-way ANOVA) to account for the variance in RT using the regression analysis described above in sections 3.3.4 and 4.1.5.2. The trial-based analysis showed co-variance between RTs and observed congruency, with slower responses associated with incongruent trials (t_{20} = 5.75, P < 0.0001, Figure 13e). In addition, this analysis found a significant positive correlation between prediction error of congruency (quantified as $-f_i \times o_i$) and RT (t_{20} = -3.28, P < 0.005, Figure 13f), showing that superior conflict level prediction was associated with faster responses.

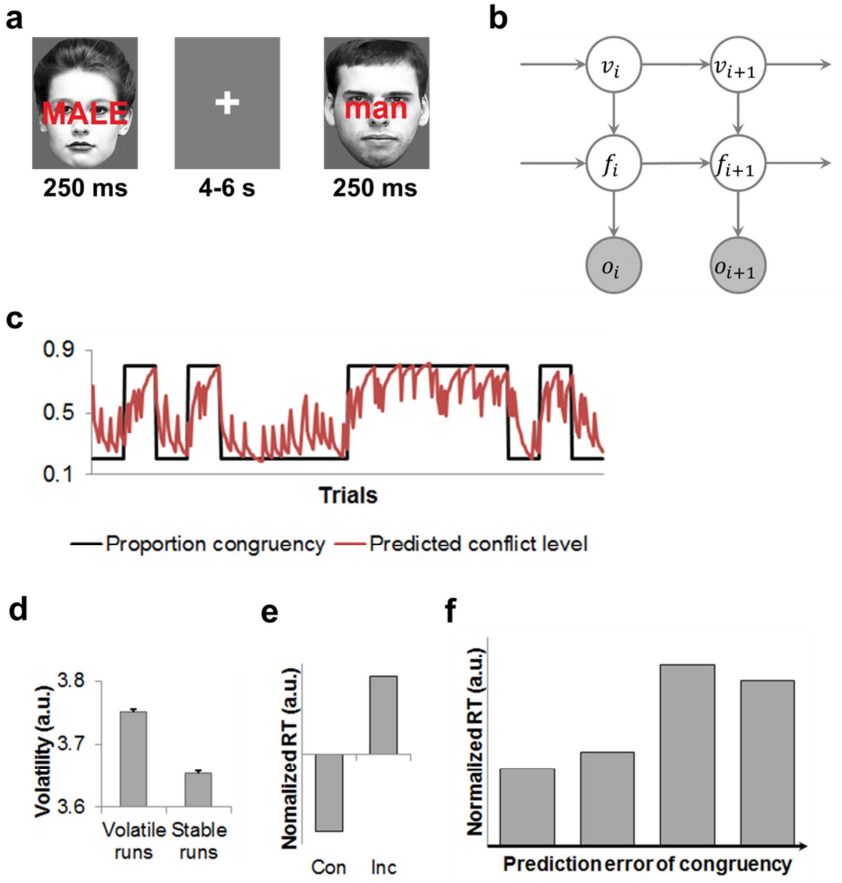


Figure 13: Experimental task, Bayesian model, and simulation and behavioral results. (a) Example stimuli and timing of the present task. This example depicts an incongruent trial, followed by a congruent trial. (b) The graphical representation of the Bayesian model of flexible conflict-control. The model uses 3 variables, volatility (v), conflict (f), and observation (o, shown in grey indicating this variable is observable) for each trial. The directed edges indicate the information flow. (c) Time course of the estimated predicted conflict level (in red) and the underlying proportion congruency level (in black) in an example session. (d) Group mean model belief of volatility and its mean standard error (MSE), plotted as a function of run type. (e) Group mean normalized RT, plotted as a function of congruency (Cog = congruent trials; Inc = incongruent trials). (f) Group mean of normalized RT, plotted as a function of prediction error of congruency.

4.2.2 Encoding of Model Variables in the Brain

The most basic assumption of our model is that the three model variables (volatility, predicted conflict, congruency) are represented in the brain. Thus, before validating the hypotheses that concern specific dynamic interactions between model components, I first tested this fundamental hypothesis using the searchlight-based analyses described above in section 4.1.5.3.

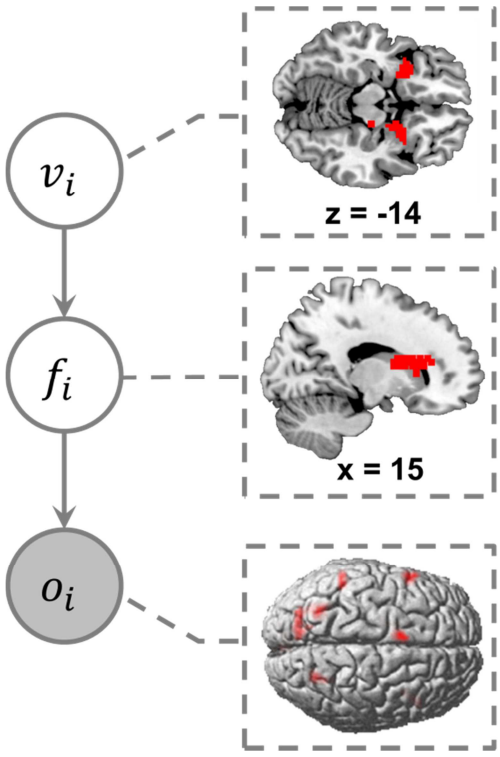


Figure 14: A graphical representation of the hierarchy of information processing in a single trial, between model variables (left) and brain areas showing significant co-variation (P < 0.05, corrected) between fMRI activation and these model variables (right, from top to bottom: volatility, predicted conflict level, and congruency). Select brain regions encoding the model variables are shown in the box linked to the corresponding variable.

The test for neural representation of each of the model variables found the estimated volatility of control-demand to be tracked by activity in the bilateral insula and adjacent inferior frontal gyri, amygdala, putamen, and right parahippocampal gyrus and precuneus (Figure 14; $P < 0.05$, corrected) via a homogenous (univariate) coding scheme. Predicted conflict levels were encoded in the right caudate (Figure 14; $P < 0.05$, corrected) via a heterogeneous (multivariate) coding scheme. Additionally, predicted conflict levels were encoded in a homogeneous fashion in the left inferior parietal lobule, right paracentral lobule and superior frontal gyrus. Finally, observed congruency was tracked by signals in medial frontal cortex, in particular the supplementary motor area (SMA), as well as in bilateral inferior frontal gyri and superior parietal lobule, left inferior parietal lobule (IPL) and precuneus (Figure 14; $P < 0.05$, corrected; univariate coding). The latter results broadly replicate previous findings from studies of Stroop-type conflict effects (Jiang & Egner, 2013).

In sum, the initial analyses portray a distributed network of frontal, parietal and subcortical regions that encode volatility of control demand and observed congruency, while the caudate nucleus appears to play a central role in the prediction of conflict (control-demand). In the following, specific model-derived hypotheses concerning dynamic interactions between model nodes are tested.

4.2.3 Volatility-modulated Updating of Predicted Conflict Level in the Caudate

Prior to stimulus presentation in a given trial, the Bayesian model updates predicted conflict level by integrating the most recently observed congruency, modulated by the model's belief of volatility (Figure 15a). As described above, the model assumes that high volatility drives up the learning rate to allow for a stronger weighting of short-term information in predicting forthcoming congruency. To test this hypothesis, I performed a learning rate estimation analysis on the caudate searchlights that have been shown to encode the predicted conflict level (see section 4.1.5.4). For each searchlight in each subject, I mean-split the data into low-volatility versus high-volatility trials, and estimated the learning rate for each half employing a traditional RL algorithm (Sutton, 1988). Across subjects, it was found that the ROI mean estimated learning rate was significantly higher in trials with higher volatility (Figure 15b, t_{20} = 4.51, P = 0.0002). I further tested the modulation of volatility on learning rate in the activation vectors in the caudate. A significant positive modulation (i.e., learning rate increases with volatility) was observed (Figure 15c, t_{20} = 6.31, P < 0.0001). These results provide strong evidence that volatility modulates the updating of neural representations in the caudate that encode the anticipated demand on cognitive control.

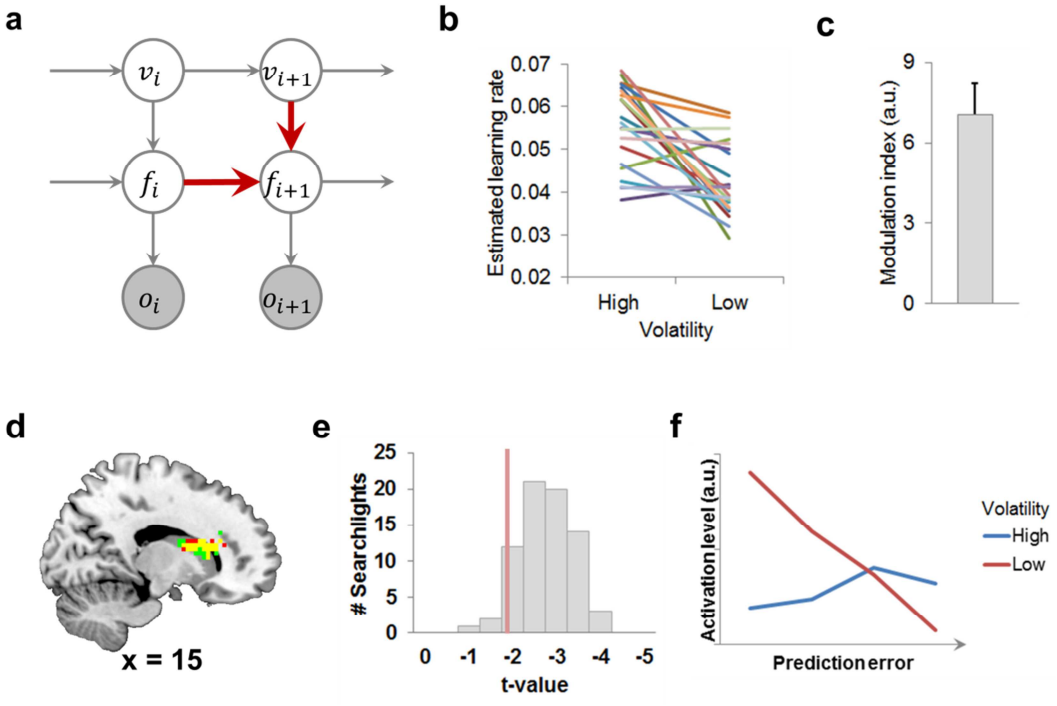


Figure 15: Modulation of volatility on brain activity encoding predicted conflict level. (a) A graphical representation of the Bayesian model, highlighting in red the information processing mechanisms related to the modulation of volatility on predicted conflict level. (b) Comparison of caudate activity-derived learning rates between high and low volatility trials. Each line represents a participant. (c) Group mean modulation of volatility on caudate activity-derived learning rate and its mean standard error (MSE). (d) Visualization of caudate searchlights showing encoding of predicted conflict level (red, P < 0.05 corrected) and caudate searchlights showing interaction between volatility and prediction error of congruency (green, P < 0.05 corrected) and their overlap (yellow). (e) Histogram of t-values measuring group level univariate effect of volatility × prediction error of congruency interaction. The t-values were calculated from searchlights in the caudate ROI. The red vertical line denotes the threshold for statistical significance (P < 0.05). (f) Activation in the caudate ROI, plotted as a function of volatility and prediction error of congruency.

If the belief of volatility regulates the updating of predicted conflict level, as suggested by the present and previous behavioral and modeling results, then it can be predicted that volatility also modulates the neural representation of prediction error of

congruency, which is another driving factor in the updating of predicted conflict level. Compared to a volatile condition, a stable condition should down-regulate the representation of short-term trial history information (in this case, the prediction error of congruency) to reduce its weight in predicting conflict level, resulting in a negative modulation. At trial i, this modulation can be formulated as a three-way interaction vector $-v_i \times f_i \times o_i$, which is the product of volatility and the prediction error of congruency $-f_i \times o_i$ (see section 4.1.5.5). Using the univariate analysis described above, I found negative three-way interaction in 63 ($P < 0.05$) out of the 73 caudate searchlights that displayed distributed encoding of predicted conflict level (Figure 15d,e). Accordingly, across these caudate searchlights, the mean effect of the three-way interaction was significantly negative ($t_{20} = -4.63$, $P = 0.0002$). The pattern of this interaction confirmed the prediction of increased suppression of prediction error in more stable conditions (Figure 15f). Importantly, the distinct coding schemes between volatility's modulation on prediction error (homogeneous encoding) and predicted conflict level (distributed encoding) suggest that the results are unlikely to be caused by intrinsic correlation between the variable and the 3-way interaction vectors.

4.2.4 Predicted Conflict Level-mediated Cognitive Control in the PFC

In our model, after predicted conflict level is updated following the most-recent observation, conflict prediction guides the application of cognitive control (Figure 16a), i.e., titrating attentional selectivity based on the anticipated conflict level. This process

should be reflected in a negative two-way interaction between predicted conflict level and observed congruency, due to higher activation when observed conflict was higher than predicted. In other words, the more unexpected an incongruent trial is (e.g., incongruent trials following a congruent trial (M. Botvinick et al., 1999), incongruent trials in a block of mostly congruent trials (Sohn et al., 2000)), the more control effort should be required for resolving conflict. Thus, we hypothesize that within unexpected incongruent trials (where predicted conflict level, or the estimated probability that the forthcoming trial was incongruent, was less than 0.5), the predicted conflict level should mediate cognitive control in a manner that trials with higher predicted conflict levels (i.e., a better match between predicted and observed conflict) should elicit less activation than trials with lower predicted conflict levels (i.e., a poor match between anticipated and observed conflict), resulting in a negative co-variation between predicted conflict level and activation level. We observed this signature of cognitive control in the right dorsal ACC and left caudal dlPFC (Figure 16b, c, $P < 0.05$, corrected). Of note, these results are unlikely to be caused by encoding of the predicted conflict level per se, because no encoding of predicted conflict level was found in the reported regions in the previous analyses (see above), and the direction of the negative co-variance is opposite to that of the activation patterns reported above, where higher activation was associated with higher predicted conflict level and higher conflict (Niendam et al., 2012) (e.g., incongruent trials).

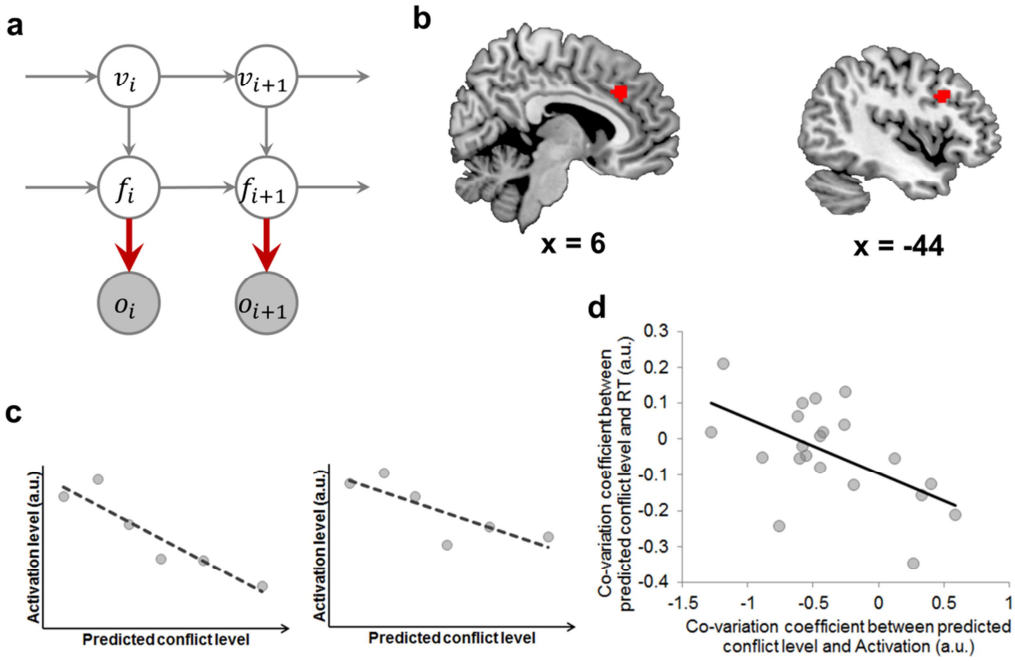


Figure 16: Modulation of predicted conflict level on cognitive control. (a) A graphical representation of the Bayesian model, highlighting in red the information processing mechanisms related to the mediation of predicted conflict level on cognitive control. (b) Centers of searchlights in the ACC (left) and dlPFC (right) regions showing significant (P < 0.05, corrected) negative co-variation between predicted conflict level and activation elicited by unexpected incongruent trials. (c) Activation in the ACC ROI (left) and the dlPFC ROI (right), plotted as a function of predicted conflict level. The dotted lines show the linear trend lines. (d) The individual co-variation coefficient between predicted conflict level and RT, plotted as a function of individual co-variation coefficient between predicted conflict level and activation level in the dlPFC cluster in (b).

It also follows from our model that the modulation of predictive conflict level on cognitive control should be reflected in behavioral performance data: participants who compensate better for the underestimated conflict level in unexpected incongruent trials in neural terms (i.e., showing a stronger negative co-variation between predicted conflict level and neural "control" activation level) should also show less behavioral conflict in

unexpected incongruent trials (i.e., less negative co-variation between predicted conflict level and RT). This predicts a data pattern where the co-variation coefficient between predicted conflict level and neural activation level is negatively correlated with the co-variation coefficient between predicted conflict level and RT across subjects. Consistent with this hypothesis, we observed such a negative correlation in the left caudal dlPFC cluster identified above (Fig. 4d, r = -0.57, P = 0.007) . This correlation was not significant in the ACC cluster (r = 0.06, n.s.), however.

4.2.5 Prediction Error-driven Updating of Volatility in the Insula

After current-trial congruency is observed, the model's volatility estimate needs to be updated to further guide the adjustment of predicted conflict level (Figure 17a). The model updates volatility based on the prediction error of congruency. Hence, according to our model, brain regions encoding volatility should also be expected to represent congruency prediction error. As noted above, the representation of the prediction error of congruency was tested for using the interaction vector of predicted conflict level and observed congruency $-\mathrm{f}_i \times o_i$. Compatible with the above prediction, a cluster of searchlights showing significant predicted conflict level × congruency effect was found in the left insula (Figure 17b, corrected). This cluster overlaps the insula cluster of searchlights encoding volatility (Figure 14), within which 24 out of 73 searchlights also showed encoding of prediction error of congruency (Fig. 5c, P < 0.05). Accordingly, across the searchlights in the left insula ROI encoding belief of volatility

(Figure 14), the mean representation of congruency prediction error was significant (t_{20} = -2.13, P = 0.045).

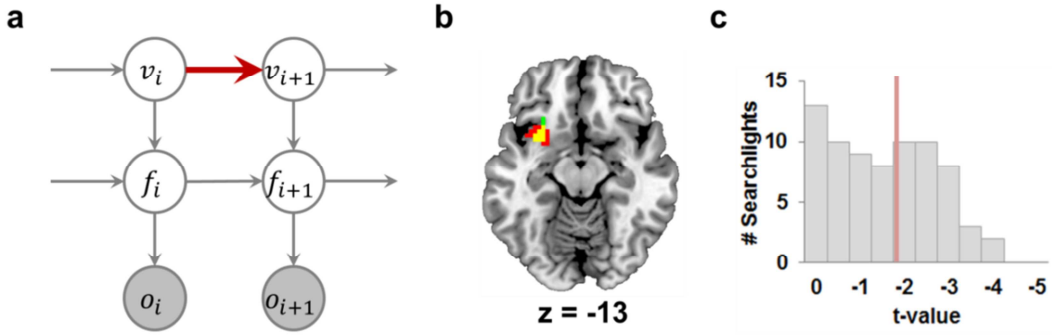


Figure 17: Congruency prediction error drives the updating of volatility. (a) A graphical representation of the Bayesian model, highlighting in red the information processing mechanisms related to the updating of volatility. (b) Visualization of searchlights showing encoding of volatility (red, P < 0.05 corrected) and searchlights showing interaction between predicted conflict level and congruency (green, P < 0.05 corrected) and their overlap (yellow). (c) Histogram of t-values measuring group level univariate effect of volatility × predicted error of congruency interaction. The t-values were calculated from searchlights in the volatility-encoding cluster shown in (b). The red vertical line denotes the threshold for statistical significance (P < 0.05).

4.3 Discussion

In the present fMRI study, the Bayesian model was employed to guide the investigation of the neural substrates underlying the flexible adjustment of cognitive control. Specifically, the neural encoding of the three model variables (volatility, predicted conflict level and congruency) were inspected, then three model predictions (volatility modulates the updating of predicted conflict level; predicted conflict level mediates cognitive control, and the prediction error of congruency drives the updating of volatility) were examined using the estimates of model variables and fMRI data.

The conflict monitoring theory claims that the brain tracks the conflict level via reinforcement learning with a fixed learning rate (M. M. Botvinick et al., 2001). As shown above, reinforcement learning is unable to account for the flexible adjustment of cognitive control. Facilitated by the Bayesian model, I found that the model estimates of predicted conflict level were represented in the caudate. Crucially, the predicted conflict level was encoded by a distributed coding scheme, which suggests that the predicted conflict level is represented by its statistical distribution across neuronal groups, with each group's activation encoding the current belief of the predicted conflict level at a specific level. This finding opens the possibility that the predicted conflict level is monitored by a more sophisticated statistical approach (i.e., by its statistical distribution) compared to the claim that the conflict level is encoded using only one scalar value. The parallel circuits in the caudate provide the structural and functional foundations of estimating and updating a representation of statistical distribution (G. E. Alexander & Crutcher, 1990; Frank & O'Reilly, 2006). Moreover, the results support the hypothesis that volatility modulates the updating of predicted conflict level in the caudate. Specifically, volatility belief regulates the dependence on short-term information (e.g., the learning rate in a reinforcement learning algorithm) and modulates the representation of short-term information (e.g., the prediction error of congruency). These findings shed light on the mechanisms of flexibly predicting the conflict level (or control-demand), which is then employed to adaptively guide cognitive control.

Previous studies defined the adjustment of cognitive control as a function of change in a specific context, such as the previous trial congruency (M. Botvinick et al., 1999; Durston et al., 2003; Egner & Hirsch, 2005a, 2005b; Kerns, 2006; Kerns et al., 2004) and the proportion incongruency (Grandjean et al., 2012; Krug & Carter, 2012; Sohn et al., 2000; Wilk et al., 2012). The present study defines the adjustment of cognitive control as a function of change in predicted conflict level, which is an integration of both contexts, and more broadly, of both short-term and long-term information regarding demands on cognitive control. The predicted conflict level was found to mediate the interference effect in ACC and dlPFC. Similar to the conflict adaptation and proportion incongruency effects, the activation profiles of these regions suggest augmented cognitive control (reflected by reduced interference effect) as the predicted conflict level increases. As key regions in monitoring conflict and adaptively applying top-down biasing to posterior regions, trial history-based modulation on interference effects have commonly been reported in ACC (Barch et al., 2001; M. Botvinick et al., 1999; Carter et al., 1998; Kerns et al., 2004; MacDonald et al., 2000; MacLeod & MacDonald, 2000) and dlPFC (Egner & Hirsch, 2005a; Kerns et al., 2004; MacDonald et al., 2000). The model-based analysis further suggests that the source of the modulation in ACC and dlPFC might be a prediction of conflict level signal, which originates in the caudate and is estimated adaptively across changing contexts via a sophisticated statistical learning mechanism.

The encoding of volatility was found in bilateral anterior insula cortices and adjacent inferior frontal gyri. A previous study has shown that the anterior insula cortex displayed higher activation in decision-making involving ambiguity (unknown probabilistic distribution of reward) than that involving known probabilistic distribution of reward (Huettel, Stowe, Gordon, Warner, & Platt, 2006). Ambiguity (or uncertainty) of underlying probabilistic distribution is similar to volatility in that volatility quantifies the likelihood the underlying probabilistic distribution varies (in other words, become uncertain). The model further predicts that the updating of volatility relies on the predicted error of congruency, which was also found encoded in the left anterior insula cortex in this study. Prediction-error related activation in the anterior insula was also reported in another study of risky decision making (Preuschoff, Quartz, & Bossaerts, 2008) in the anterior insula. Moreover, a recent theory of the function of anterior insula cortex also supports the encoding of volatility and its updating: Craig (2009) claims that a generic function of anterior insular is to integrate salience across various factors (e.g., the structure of the task) to form a representation of awareness at a given moment. According to that model, the salience of any factor is determined by its contribution to the homeostasis of factors. Thus, volatility and prediction error of congruency, as measures reflecting potential violation of homeostasis of one's belief of the probabilistic distribution of conflict level, would feasibly be encoded in the anterior insula according to this model. Moreover, Craig (2009) proposes that a representation of awareness in a

"finite present" (a present-centered period of time) is generated from the representations of awareness via temporal integration, which can potentially support the temporal updating of volatility.

In sum, facilitated by the Bayesian model, the present fMRI study extends the classic conflict monitoring model (M. M. Botvinick et al., 2001; M. M. Botvinick et al., 2004) to account for the flexible adjustment of cognitive control. Instead of estimating a fixed learning strategy to predict conflict level, the present model-based analyses revealed the involvement of anterior insula in encoding control-demand volatility, which modulates the updating of predicted conflict level, and the involvement of the caudate, which provides a flexible learning system to predict conflict level. These findings depict an orchestra of various systems that adaptively adjust cognitive control to maintain optimal behavior.

5. Summary

In this section, I first summarize the major findings of the present dissertation in section 5.1, organized according to the relationship between 4 factors: the experimental design, the Bayesian model, the behavioral data and the fMRI data. Related works and their links to this dissertation are also discussed. I then point out some limitations of this dissertation and some open topics that future studies might tackle in section 5.2.

5.1 General Discussion

In this dissertation, I investigated potential mechanisms supporting the flexibility of adjustments in cognitive control, allowing us to adapt to environments with varying patterns of conflict (and thus, varying control-demand). I argue that this flexibility can be explained by a mechanism that gauges volatility, which determines how information sampled from various temporal scales should be integrated to optimally predict the demand of cognitive control. This mechanism was then formulated in a Bayesian model that simulated control processes and task performance on a trial-by-trial basis. Prior to the onset of the stimulus at each trial, volatility regulates the prediction of conflict level, which is in turn involved in determining the demand for cognitive control. After the stimulus is presented, the perceived congruency of the stimulus updates the belief of volatility and predicted conflict level. This prediction-update loop repeats for each trial and produces trial-by-trial estimates of volatility and predicted conflict level.

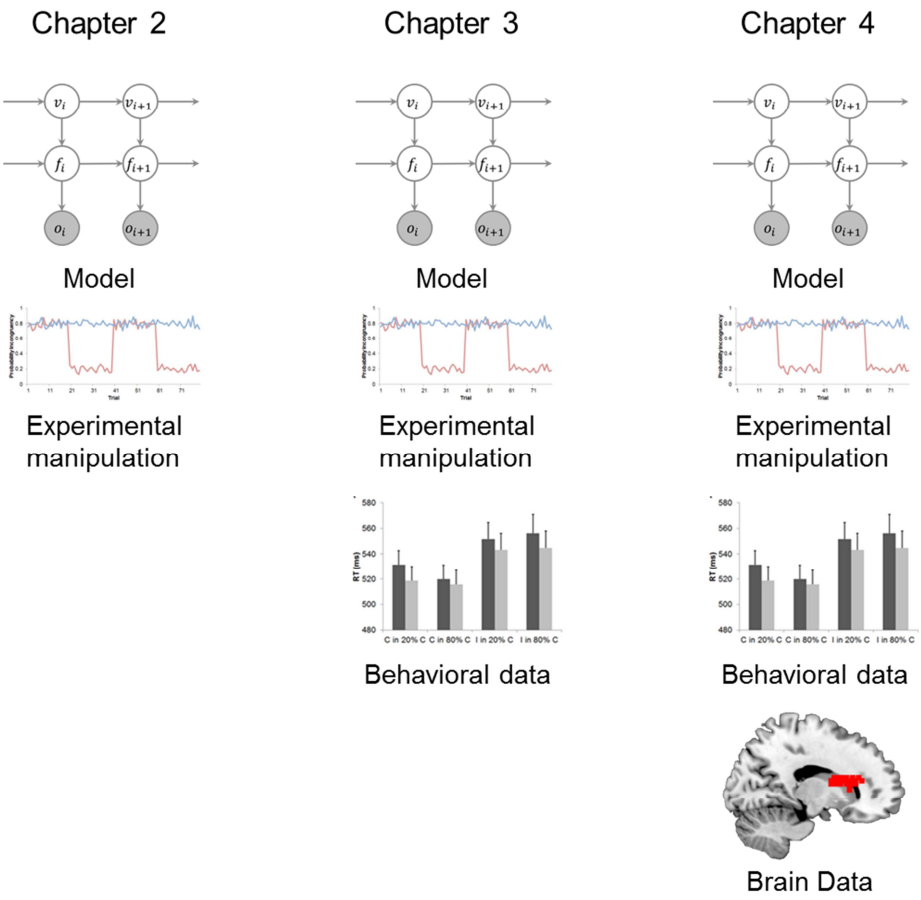


Figure 18: An illustration showing the order of different "components" being integrated into this dissertation to investigate flexible adjustment of cognitive control.

Starting with this model, I conducted a series of studies to investigate the mechanism of flexible adjustment of cognitive control by successively including additional "components" in each chapter (Figure 18). These added components validate the model and/or provide new evidence of the mechanisms underlying the flexibility of the adjustment of cognitive control. The model was first validated using an experimental design that manipulated the reliability of short-term and long-term information by

varying the frequency of alternating two underlying proportion incongruency distributions (section 2.2.3). Then I demonstrated that the proposed model simultaneously accounts for two classic behavioral phenomena in cognitive control that are believed to have different reliance on short-term and long-term trial-history (section 3.1). In particular, the behavioral studies in sections 3.2 and 3.3 indicated that (1) the experimental design was able to elicit behavioral change related to flexible adjustment of cognitive control and (2) that the behavioral changes could be accounted for by the Bayesian model. They also lay the foundations for an fMRI study (chapter 4), which employed the Bayesian model and the same experimental design to guide the exploration of neural substrates of flexible cognitive control. Compared to the conflict monitoring model and the dual mechanisms model of cognitive control, the Bayesian model is more parsimonious (e.g., using a single mechanism to account for different behavioral phenomena) and more appropriate (e.g., using one-the-fly simulation rather than post-hoc fitting). Taken together, the success of the Bayesian model provides evidence to support the existence of a unified system for flexible adjustment of cognitive control. In the following sections, I extend these findings and discuss their potential relation to other works.

5.1.1 How "Bayesian" is Cognitive Control?

According to Bowers & Davis (2012), there are three levels at which one can use Bayesian methods in modeling cognitive processes: as computational tools, for

generating "optimal" benchmarks for cognitive processes, and for modeling the actual neural computations carried out by the brain. The Bayesian models reviewed above in this dissertation, along with the proposed Bayesian model, all operate at the second level: these Bayesian models were treated as "optimal observers", and produced optimal predictions, which in turn were used to account for behavior and neuroimaging data. Using the estimates generated by the "optimal observer", behavioral data and neural activity involved in the flexible adjustment of cognitive control can be accounted for. For example, volatility and predicted conflict level were represented in the dlPFC and caudate, respectively (chapter 4). The neuroimaging results further revealed a chain of modulation from volatility to predicted conflict level, and then to impact of observed congruency. These results support the argument that the brain adjusts cognitive control through the information flow as envisaged in the proposed Bayesian model.

If the brain processes information related to cognitive control in a Bayesian (or more broadly, a predictive statistical) approach, two more premises must be met. First, the variables (e.g., volatility, predicted conflict level) should be represented as probabilistic variables. In other words, the encoding of these variables should be in the form of probabilistic distributions rather than a scalar value. Second, the updating of variable states should be computed on probabilistic distributions. The first premise requires neurons encoding the variables to either perform distributed encoding (e.g., each neuron represents the probabilistic density at a given value) or to have the ability of

swiftly change tuning curves to accommodate the updating of the probabilistic distribution. Regarding the second premise, the updating on probabilistic distribution requires at least two variables (e.g., a variable to be updated, and other variable(s) providing the information to update) and hence is performed on a joint probabilistic distribution that has a dimension of at least two (one for each variable). Because homogeneous encoding can only represent one-dimensional information, the second premise predicts distributed encoding.

Following this logic, the distributed representation of predicted conflict level found in the fMRI study supports both of these premises, and thus the proposal of Bayesian implementation of cognitive control in the brain. Yet, fMRI findings of homogeneous (univariate) encoding may not represent unambiguous support for homogeneous coding due to the coarse spatial resolution of fMRI. For instance, it is possible that a voxel covers a population of neurons that encodes a statistical variable and in a distributed fashion, and thus the fMRI signal of that voxel exhibits a pattern of homogeneous encoding. Neuronal recording or other techniques with higher spatial resolution than fMRI are needed for a more accurate test of these premises.

Previous studies are promising in this regard. At the neuronal level, it has been shown that Bayesian models can account for neural firing rate data in decision making (Beck et al., 2008) and attention (Rao, 2005) tasks. Thus the brain might adopt the neural implementation of those Bayesian models to support flexible cognitive control.

5.1.2 Actor-critic Models

Several prior computational models have employed an "actor-critic" architecture to simulate neural responses in the ACC (W. H. Alexander & Brown, 2011; Silvetti, Alexander, Verguts, & Brown, 2013; Silvetti, Seurinck, & Verguts, 2011, 2013). The "actor-critic" architecture consists of two components, an "actor" component and a "critic" component. During the simulation in a time-step, these models first make prediction of the stimulus/outcome using the actor component. The discrepancy between the prediction and the actual stimulus/outcome is processed by the critic component to update the actor component's prediction in the next time-step. Using the reinforcement learning algorithm driven by unsigned prediction error, these models are able to account for ACC activity in a variety of tasks.

The Actor-critic models and the proposed Bayesian model are similar in a few ways: Firstly, they share the two-step algorithm that first predicts the stimulus/outcome and then updates the prediction based on prediction error. Secondly, they all assume that this update (or learning) should be flexible. Specifically, the predicted response-outcome (PRO) model (W. H. Alexander & Brown, 2011) adjusts the learning rate based on prediction error. The reward value prediction model (Silvetti et al., 2011; Silvetti, Seurinck, et al., 2013), although adopting a fixed learning rate, argues that the adjustment of learning rate may occur in the locus coeruleus. The proposed Bayesian model regulates learning via volatility. Lastly, all of the models predict larger prediction

error-related activity in the ACC in more volatile conditions. This prediction was validated by the finding of significant positive 3-way interaction in the ACC reported in the fMRI study above.

Despite these similarities, the actor-critic models and the proposed Bayesian model also differ in some essential aspects. One key difference is the explicit consideration of volatility. The actor-critic models argue that a dedicated variable of volatility is unnecessary, because learning can be modulated directly via prediction error. On the contrary, Bayesian model explicitly includes a variable of volatility to modulate the update of predicted conflict level. In support of the contention that the brain computes the volatility of control-demand, the fMRI study presented in this dissertation found an anterior insula/IFG region that represented the time course of volatility estimates produced by the Bayesian model, as well as an interaction between volatility estimate and prediction error of congruency. Additionally, the PRO model predicts that learning rate increases as prediction error increases, whereas in the proposed Bayesian model the opposite is true. Future studies are needed to reconcile this discrepancy.

5.1.3 Predictive Coding

The predict-update approach used in the Bayesian model also relates it to the predictive coding theory of information processing in the brain (Friston, 2005; Mumford, 1992; Rao & Ballard, 1999), which states that information processing is facilitated by top-

down modulation based on prediction, and that prediction error is transmitted bottom-up to update the prediction to improve its accuracy. The updating of prediction is a crucial part in the predictive coding theory, because the efficiency of the update directly affects the performance of future information processing. It has been shown that this update can be modulated by attention, which boosts prediction error to enhance the efficiency of learning (Jiang, Summerfield, & Egner, 2013). The Bayesian model, with the variable of volatility, may suggest an additional mechanism that could improve learning in the predictive coding framework. Specifically, volatility reflects the likelihood that prediction error is due to a fundamental change in the environment and it thus determines the influence of prediction error on the updating. In the implementation of the Bayesian model, volatility regulates the update from prediction error by modulating the likelihood of the new prediction deviating from its previous state, which has a similar effect as modulating learning rates in a reinforcement learning model, as demonstrated both in the validation (section 2.2.1) and neural data (section 4.3). Thus, in addition to attention amplifying prediction error, the effect of volatility demonstrates a complementary mechanism that can facilitate the updating from prediction error by modulating its utility.

5.1.4 Volatility vs. Other 2nd-order Measures

In the proposed Bayesian model, the function of volatility is to determine how likely the new predicted conflict level is to deviate from its previous state. Thus it can be

considered as a quantity of second-order uncertainty (e.g. the deviation of probability, see (Yu & Dayan, 2005)) that reflects the variance or SD of the distribution of predicted conflict level (although strictly speaking, volatility is not equivalent to SD in the Bayesian model because the SD of $p(f_{i+1}|f_i, v_{i+1})$ also depends on the previous state of predicted conflict level). Deviation of reward from its mean has been found to be encoded in posterior cingulate cortex neurons (McCoy & Platt, 2005). Another study has shown that in the primate anterodorsal septum, neural firing rates display an "inverted-U" shape as a function of the probability of reward, peaking at 50% (Monosov & Hikosaka, 2013). Given that the SD of reward increases starting from 0 reward probability, then peaks at 50%, and finally drops until the probability of reward reaches 100%, this inverted-U pattern of neural firing may also represent potential neural substrates of deviation and/or volatility. A similar neural firing pattern was also found in the midbrain (Fiorillo, Tobler, & Schultz, 2003) and the ventral striatum (Preuschoff, Bossaerts, & Quartz, 2006) in monkeys.

Volatility is also associated with confidence (or uncertainty). For example, in a stable environment, prediction of conflict is usually more confident because the expectation is less likely to be violated. Previous studies have shown that confidence of decision is encoded in the lateral intraparietal cortex (Kiani & Shadlen, 2009) as well as ventral medial and rostrolateral prefrontal cortices (De Martino, Fleming, Garrett, & Dolan, 2013).

Finally, high volatility also arguably renders decision-making more ambiguous; hence neural substrates of ambiguity, as found in orbital frontal cortex (Hsu, Bhatt, Adolphs, Tranel, & Camerer, 2005) and the lateral PFC (Huettel et al., 2006) may also contribute to encoding volatility.

Although the brain regions supporting volatility-related encoding and modulation found in this dissertation (e.g., encoding of volatility in the anterior insula, and modulation involving volatility in the caudate) do not correspond closely to these previous findings, this dissertation extends those prior works in two ways: first, compared to previous studies that usually predicted the outcome of a particular stimulus-response ensemble, the current model was employed to predict the demand of cognitive control and thus extends the previous findings. Second, the trial-based analysis employed in this dissertation extends the time-resolution of the aforementioned studies from block- or condition-level to trial level. Thus the findings in this dissertation may reveal neural substrates underlying learning (of the predicted conflict level) at a finer temporal resolution. Nevertheless, given the conceptually similarity between volatility and other 2nd-order statistical measures, these types of estimates might be supported by a generic mechanism that operates independent from specific task-settings. This dissertation provides new materials for the exploration of this generic mechanism.

5.2 Limitation and Future Directions

Although this Bayesian model has shown great potential in accounting for classic behavioral phenomena and exploring the neural substrates of flexible cognitive control, there are a few caveats that can be addressed in future studies. There are also some novel routes that are potentially interesting to take as follow-up investigations. In the following sections I discuss them in details.

5.2.1 Within-trial Simulation

In the studies reported in this dissertation, the Bayesian model only simulated cognitive control in an across-trial manner. Thus the dynamics of cognitive control within a trial remain unclear. It would be interesting to examine how the predicted conflict level modulates the resolution of conflict, and how the prediction error accumulates and guides the update of predicted conflict level and volatility within a trial, in particular when the prediction deviates from the actual congruency. Several considerations should be taken into account prior to embarking on such a project, however. Firstly, from the perspective of the Bayesian model, there may not be any change of the environment within a trial. Thus, a simplified model (e.g., without the variable of volatility or replacing the variable with a constant for volatility, assuming it stays the same during a trial) could feasibly be used for simulation. However, this is not to say that volatility has no effect on within-trial dynamics – only that this effect may not change within a single trial. Indeed, one could extract quantitative characteristics from

within-trial dynamics (e.g., speed of information accumulation, starting bias, and so on), and analyze them across trials as a function of volatility.

Secondly, from the perspective of behavioral data analysis, within-trial modeling usually depends on distributions of RTs, which in turn requires large number of trials to ensure that the sample accurately reflects the underlying distribution (Brown & Heathcote, 2005; Ratcliff, Van Zandt, & McKoon, 1999). This would increase the difficulty of data acquisition. Fortunately, data acquisition could be facilitated by AMT (see Appendix), as documented in section 3.3.

Finally, from the perspective of the imaging methodology, the fMRI approach employed in this dissertation would be sub-optimal for studying the dynamics of within-trial brain activity at high temporal resolution compared to electroencephalography (EEG), magnetoencephalography (MEG) and invasive neural recordings. Future studies may therefore consider using these techniques in conjunction with the new Bayesian model to study flexible within-trial cognitive control at millisecond resolution.

5.2.2 Meta-volatility?

In the Bayesian model proposed here, the transition distribution of volatility depends on a parameter σ_v, which stays constant throughout the experiment. One way of comprehending the function of σ_v is that it controls how fast volatility changes temporally. For example, volatility can remain high for a long time (low σ_v, e.g., remain

at 50% proportion incongruency for 200 trials) or volatility can vary between different values at low levels (high σ_v, e.g., alternate between 80% and 90% proportion incongruency every 20 trials). Thus σ_v could also be considered an index of "meta-volatility". The proposed Bayesian model simply assumes that σ_v is a constant within an experiment. However, in theory, to account for possible changes of speed of volatility, the model could include another variable of σ_v that would depend on yet another variable to control the temporal change of σ_v, and so forth, *ad infinitum*. Consequently, this approach opens the door to a potential infinite regress, rendering a "complete" modeling the flexibility of cognitive control impossible.

The key to this issue lies in how flexible cognitive control needs to be. The brain makes decision promptly, thus sometimes it sacrifices flexibility for efficiency (cf., the habitual S-R mappings described in chapter 1). Thus, although in theory the variables of volatility, meta-volatility and meta-meta-volatility (etc.) can be expanded to infinite levels, the brain does not need to implement all of these levels. To empirically test to which level the brain encodes information about control, a model comparison between the present Bayesian model and a Bayesian model that incorporates σ_v as a variable could be conducted on data acquired using a task that manipulates the speed of the shifting of volatility. If adding the variable of meta-volatility cannot significantly improve the performance of the model in accounting for human data, it may be the case that the brain does not encode σ_v as a variable.

5.2.3 Accounting for the Cost and Benefit of Cognitive Control

The present Bayesian model simply assumes that the amount of control applied is linearly correlated with the predicted conflict level. Nevertheless, Shenhav, Botvinick & Cohen (2013) propose that the amount of control applied also depends on considerations the cost and benefit brought by engaging cognitive control. For example, in the gender Stroop task, one could apply little control in incongruent trials if the cost of applying control overweighs the gain of correctly responding to a stimulus. Given the high accuracy in the present studies, I assume that the subjects considered the cost-benefit function to justify applying cognitive control at a sufficient level for performing near-perfectly in this relatively simple task. However, future studies could expand the Bayesian model by incorporating an optimization problem involving the predicted conflict level, as well as estimated costs and benefits of engaging cognitive control to test the relationship between predicted conflict level and applied control level under different difficulty/motivation settings.

This route of future research can also help explain performance variability in cognitive control that exists between motivational states (Locke & Braver, 2008), between emotional states (Braem et al., 2013; Egner et al., 2008; Etkin, Egner, Peraza, Kandel, & Hirsch, 2006; Reeck & Egner, 2011) and across individuals (Burgess & Braver, 2010). The current Bayesian model succeeded in modeling the *mean* behavior and brain activity in a group of subjects. Nevertheless, it is unable to account for the *variance* above because its

estimated states of variables are only based on the congruency variable. By introducing the costs and benefits of applying cognitive control, and thus dissociate the assessment of cognitive control from the prediction of conflict level, the extended model would have the potential to explain intra- and inter- individual variance. Similar to estimating the states of volatility and conflict level, the Bayesian model could in theory infer the states of the added variables, and inferred states can be applied to explore the mechanisms underlying behavioral and brain data.

I speculate that, in a healthy population, most of the performance variance arises from the difference in the mapping from predicted conflict level to the amount of cognitive control applied. In other words, the predicted conflict level may be highly consistent across task-settings and subjects (e.g., people have similar estimates of the predicted conflict level), and the variance is more likely due to different cost functions (e.g., people differ in making decision of how much control to exert from the same predicted conflict level).

5.2.4 Assessing the Causal Roles of Model Variables Using Transcranial Magnetic Stimulation (TMS)

One common limitation in neuroimaging study is the lack of a measure of the causal involvement of a given brain region in bringing about the subjects' behavior. For example, from the fMRI study reported in chapter 4, one could only infer that volatility is encoded in the dlPFC because of the co-variation between brain activity in the dlPFC and the vector of volatility estimates. However, co-variation does not guarantee

causality. One way of showing causality is through TMS, which can temporarily inhibit a brain's activation by applying a transient magnetic field over that region. Thus a possible future study could aim at impairing computations in the dlPFC using TMS and inspecting if this intervention compromise subjects' performance in flexibly adjusting cognitive control.

5.3 Conclusions

In this dissertation, I propose a formal Bayesian model to account for the flexible adjustment of cognitive control in conflict tasks. Endowed by a volatility variable that adaptively regulates the influence of long-term and short-term trial history on the prediction of conflict level, this model successfully simulates several classic behavioral phenomena observed from the cognitive control literature, and furthermore facilitates a principled, model-guided investigation of the neural substrates underlying the flexible adjustment of cognitive control. Extended versions of the Bayesian model may also facilitate future investigations to reveal more details of the neuro-computational architecture of cognitive control, such as its within-trial dynamics, the relationship between prediction of conflict level and the implementation of cognitive control, and so on. Thus, I conclude that the Bayesian model provides a feasible solution to explain the mechanisms underlying the flexible adjustment of cognitive control.

Appendix A

A.1 An Introduction to Amazon Mechanical Turk (AMT)

One major limitation in conducting behavioral studies in the lab is the long data acquisition time. This can be attributed to three reasons: First, the size of subject pool is typically relatively small (e.g., a few hundreds of potential participants) and subject to fluctuation (e.g., student subject pools are often unavailable during breaks). Second, the capacity of simultaneously testing several participants is often limited. Third, the hours during a day one can run behavioral studies are limited. As a result, a researcher often needs weeks or even months to recruit enough subjects for a single study.

This inefficient data acquisition problem can be greatly mitigated by using AMT (https://www.mturk.com), an online cloud-sourcing platform. In the context of conducting a behavioral study, AMT serves as a "subject pool" where experimenters (or "requesters", as AMT define them) can recruit volunteers to perform behavioral tasks online and compensate them via the AMT payment system. In 2011, the number of registered online subjects (or "workers", as AMT define them) reached 500,000, providing a huge and stable source for subject recruitment. The online subjects perform tasks on their own web browsing devices, thus multiple subjects can work on the same task independently and simultaneously. The subjects are located across different time zones (http://techlist.com/mturk/global-mturk-worker-map.php) and can participate in a study anytime in the absence of direct interactions with the experimenter, so the data

acquisition can occur at any time. Overall, AMT can be utilized to vastly accelerate data acquisition for behavioral tasks.

Despite of AMT's great potential in facilitating behavioral studies, several concerns have been raised by researchers. Many of these concerns stem from two core issues, either regarding the subjects or the techniques implementing the behavioral tasks.

A major concern about the subjects is how well they perform the tasks. This concern can be further decomposed into two more specific ones, namely: (1) Do the subjects understand the instructions? And (2) are they motivated enough to follow the instructions throughout a task? For (1), most AMT workers are located in the United States and India (Ipeirotis, 2010), both of whose primary languages are English. Thus most subjects should be able to understand instructions written in English. One can also add country of residence to the recruitment criteria to further filter out non-English speakers from participation. Moreover, AMT encourages requesters to provide previews of their tasks for workers to decide if they would like to sign up. The preview can also be used as a practice session to test if the subjects follow the instructions. For (2), AMT workers are generally motivated to follow the instructions and finish a task once they sign up, because they are paid by the quality of their work (i.e., they may not be compensated if they perform poorly), according to AMT's payment policy (Institutional Review Board at Duke also honors this policy). In addition, rejection of payment may

also impede a worker's qualification of performing future tasks. There have also been some studies (see the studies reviewed below, as well as the study reported in section 3.2) showing that most online subjects follow the instructions and complete behavioral tasks with high accuracy.

A second common concern is that the demographics of the subjects may differ from that of the general population. In fact, a recent study points out that the AMT online subjects population has demographics similar to that of the whole American online population (http://www.behind-the-enemy-lines.com/2009/03/turker-demographics-vs-internet.html). Compared to the student subject pools many researchers use, AMT subjects form a more diverse group in terms of age, education, races, ethics, and so on. Thus, in theory the AMT results should in fact be more generalizable.

Because the rules of most behavioral tasks are relatively simple, one may worry that some workers can use some software to perform the task rather that doing it themselves. This is a highly unlikely scenario, because most cognitive tasks (e.g., identify the gender of a face image) are still too difficult for computer programs to accomplish. Furthermore, the rules and stimulus sets vary significantly across tasks, making it almost impossible to develop a generic software package for all behavioral tasks. Even developing a computer program that solves a single task may require significantly more time than needed in completing that task in person.

Another concern regarding the subjects is how to prevent one subject from participating in the same task more than once. AMT's policy restricts that each worker account can only sign up for a task once. However, it is possible that some workers with multiple worker accounts perform some relatively profitable tasks more than once. To reduce this possibility, one can record the IP address for each participant and block multiple participants from one IP address.

Many behavioral tasks rely on precise timing of stimuli presentation and response recording, so some researchers may worry that running tasks online may compromise the precision of timing. The precision of timing depends on the implementation of the behavioral task. For example, in the implementation of the gender variant of the Stroop task (see below), subjects downloaded all stimuli and programs and run the task locally in their web browsers. Thus the timing was not affected by the speed of internet connection. Current industry standard requires modern web browsers to support timing that is "accurate to a thousandth of a millisecond" (http://www.w3.org/TR/hr-time/#sec-DOMHighResTimeStamp). Hence it is possible to achieve highly precise timing.

One last concern is the uncontrolled apparatus. For example, a task can be performed by online subjects using various screen/window sizes, and input devices with different lags. The subjects may even be distracted from the task. These uncontrolled factors may cause a larger amount of noise in data acquired via AMT than in controlled

lab settings. This caveat can be alleviated by including more subjects. More important, from a different perspective, these uncontrolled factors indeed eliminates the possibility that some significant effects are caused by nuisance factors such as the size of the stimuli, specific hardware, or even the way the experimenter gives instructions. In other words, significant effects observed from AMT data should be no less (if not more) generalizable compared to data acquired in the lab.

To sum up, it has been shown that AMT is a legitimate tool for conducting behavioral studies in general. A couple of studies have further validated AMT specifically for cognitive control tasks. Crump, McDonnell, & Gureckis (2013) examined online workers' performance on Stroop, task-switching and Flanker tasks. Each task contained 96 trials and lasted approximately 5 minutes. Sixty subjects participated in each task. Subject attrition rate ranged from 10% to 33%. Both the error rate and RT showed at least marginally significant interference effects in Stroop and Flanker tasks. In the task-switching task, there are significant switching effects (repeat < switch) in both error rate and RT. These patterns replicate their classic versions of these 3 tasks. Weissman, Jiang & Egner (submitted) investigated the conflict adaptation effect using a prime-probe task conducted on AMT. The task contained 388 trials and lasted around 20 minutes. Forty-five subjects were included. The subject attrition rate is low (4%). They also found significant interference effects. In addition, RT showed a significant conflict

adaptation effect, which is similar to another study conduct in the lab using the same design (Schmidt & Weissman, Submitted).

These two studies paved the road to investigating effects in cognitive control via AMT. To further illustrate that AMT is also suitable to inspect the flexibility of cognitive control, the next section reports a behavioral study conducted via AMT using the same design as the study reported in section 3.2.

A.2 Running the Flexible Cognitive Control Task on AMT

A.2.1 Subjects

Ninety subjects gave informed consent online in accordance with institutional guidelines and participated in this study through AMT. Six subjects were excluded from analysis due to random level accuracy. The remaining 84 subjects self-reported demographic information online (46 females; mean age, 31).

A.2.2 Stimuli and Procedure

The stimuli and procedure used in this study were used as the ones described above in section 3.3.2.

A.2.3 Experimental Design

The task and its design were used as the ones described above in section 3.2.3.

A.2.4 Data Analysis

The data analysis was as the ones described above in section 3.2.4.

A.3 Comparison between Lab Data and AMT Data

The online subjects achieved high accuracy during this task (mean accuracy = 0.948 ± 0.005), suggesting the followed the instructions. As can be seen in Table 3, both the mean accuracy and mean RT were comparable between this study and the laboratory-based study described in section 3.2.5. Statistical analysis on RT further revealed that the AMT data yielded comparable results on several effects, including significant interference effect ($F_{1,83}$ = 81.1, P < 0.001) due to longer RTs in incongruent trials (568 ± 18 ms) than in congruent trials (523 ± 16 ms), and proportion incongruency effect ($F_{1,83}$ = 18.0, P < 0.001) driven by a larger interference effect in 80% congruency blocks (57 ±7 ms) than in 20% congruency blocks (32 ± 5 ms), and marginally significant main effect of volatility in RT ($F_{1,83}$ = 3.6, P = 0.06), driven by slower RT in volatile runs (553 ± 19 ms) than in stable runs (539 ± 16 ms).

Finally, in the 3-way ANOVA on RT, non-significant effects in the data from section 3.2 were also non-significant in the AMT data. These results strongly support that AMT is capable of acquiring legitimate behavioral data for tasks on flexible cognitive control.

Table 1: Comparison of results between AMD data and lab data

	AMT data	Data from section 3.2
Number of subjects	84	46
Mean accuracy	94.8%	92.8%
Mean RT	545 ms	535 ms
RT: interference effect	44 ms (P < 0.001)	28 ms (P < 0.001)
RT: proportion incongruency effect	25 ms (P < 0.001)	10 ms (P = 0.002)

RT: main effect of volatility	6 ms (P = 0.06)	10 ms (P = 0.04)

References

Alexander, G. E., & Crutcher, M. D. (1990). Functional architecture of basal ganglia circuits: neural substrates of parallel processing. Trends Neurosci, 13(7), 266-271.

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. Nat Neurosci, 14(10), 1338-1344.

Bach, D. R., & Dolan, R. J. (2012). Knowing how much you don't know: a neural organization of uncertainty estimates. Nat Rev Neurosci, 13(8), 572-586.

Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. Trends Cogn Sci, 12(5), 193-200.

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology, 37(4-5), 407-419.

Barch, D. M., Braver, T. S., Akbudak, E., Conturo, T., Ollinger, J., & Snyder, A. (2001). Anterior cingulate cortex and response conflict: effects of response modality and processing domain. Cereb Cortex, 11(9), 837-848.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., . . . Pouget, A. (2008). Probabilistic population codes for Bayesian decision making. Neuron, 60(6), 1142-1152.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. Nat Neurosci, 10(9), 1214-1221. doi: nn1954

Blais, C., Robidoux, S., Risko, E. F., & Besner, D. (2007). Item-specific adaptation and the conflict-monitoring hypothesis: a computational model. Psychol Rev, 114(4), 1076-1086.

Botvinick, M., Nystrom, L. E., Fissell, K., Carter, C. S., & Cohen, J. D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. Nature, 402(6758), 179-181.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. Psychol Rev, 108(3), 624-652.

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. Trends Cogn Sci, 8(12), 539-546.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. Psychol Bull, 138(3), 389-414.

Braem, S., King, J. A., Korb, F. M., Krebs, R. M., Notebaert, W., & Egner, T. (2013). Affective modulation of cognitive control is determined by performance-contingency and mediated by ventromedial prefrontal and cingulate cortex. J Neurosci, 33(43), 16961-16970.

Braver, T. S. (2012). The variable nature of cognitive control: a dual mechanisms framework. Trends Cogn Sci, 16(2), 106-113.

Braver, T. S., & Barch, D. M. (2002). A theory of cognitive control, aging cognition, and neuromodulation. Neurosci Biobehav Rev, 26(7), 809-817.

Braver, T. S., Gray, J. R., & Burgess, G. C. (2007). Explaining the may varieties of working memory variation: Dual mechanisms of cognitive control. In A. Conway, C. Jarrold, M. Kane, A. Miyake & J. Towse (Eds.), Variation in Working Memory (pp. 76-106). Oxford: Oxford University Press.

Brown, S., & Heathcote, A. (2005). A ballistic model of choice response time. Psychol Rev, 112(1), 117-128.

Bugg, J. M. (2012). Dissociating Levels of Cognitive Control: The Case of Stroop Interference. Psychological Science, 21(5), 302-309.

Bugg, J. M., & Chanani, S. (2011). List-wide control is not entirely elusive: evidence from picture-word Stroop. Psychon Bull Rev, 18(5), 930-936.

Bugg, J. M., & Crump, M. J. (2012). In Support of a Distinction between Voluntary and Stimulus-Driven Control: A Review of the Literature on Proportion Congruent Effects. Front Psychol, 3, 367.

Bugg, J. M., & Hutchison, K. A. (2013). Converging evidence for control of color-word Stroop interference at the item level. J Exp Psychol Hum Percept Perform, 39(2), 433-449.

Burgess, G. C., & Braver, T. S. (2010). Neural mechanisms of interference control in working memory: effects of interference expectancy and fluid intelligence. PLoS One, 5(9), e12861.

Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. Science, 280(5364), 747-749.

Clithero, J. A., Carter, R. M., & Huettel, S. A. (2009). Local pattern classification differentiates processes of economic valuation. Neuroimage, 45(4), 1329-1338.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the Stroop effect. Psychol Rev, 97(3), 332-361.

Craig, A. D. (2009). How do you feel--now? The anterior insula and human awareness. Nat Rev Neurosci, 10(1), 59-70.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. PLoS One, 8(3), e57410.

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. Nat Neurosci, 16(1), 105-110.

De Pisapia, N., & Braver, T. S. (2006). A model of dual control mechanisms through anterior cingulate and prefrontal cortex interactions. Neurocomputing, 69(10-12), 1322-1326.

den Ouden, H. E., Daunizeau, J., Roiser, J., Friston, K. J., & Stephan, K. E. (2010). Striatal prediction error modulates cortical coupling. J Neurosci, 30(9), 3210-3219.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. Annual review of neuroscience, 18, 193-222.

Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. Nat Rev Neurosci, 2(11), 820-829.

Durston, S., Davidson, M. C., Thomas, K. M., Worden, M. S., Tottenham, N., Martinez, A., . . . Casey, B. J. (2003). Parametric manipulation of conflict and response competition using rapid mixed-trial event-related fMRI. Neuroimage, 20(4), 2135-2141.

Egner, T. (2007). Congruency sequence effects and cognitive control. Cogn Affect Behav Neurosci, 7(4), 380-390.

Egner, T., Ely, S., & Grinband, J. (2010). Going, going, gone: characterizing the time-course of congruency sequence effects. Front Psychol, 1, 154.

Egner, T., Etkin, A., Gale, S., & Hirsch, J. (2008). Dissociable neural systems resolve conflict from emotional versus nonemotional distracters. Cereb Cortex, 18(6), 1475-1484.

Egner, T., & Hirsch, J. (2005a). Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. Nat Neurosci, 8(12), 1784-1790.

Egner, T., & Hirsch, J. (2005b). The neural correlates and functional integration of cognitive control in a Stroop task. Neuroimage, 24(2), 539-547.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. Perception and Psychophysics, 25, 249-263.

Etkin, A., Egner, T., Peraza, D. M., Kandel, E. R., & Hirsch, J. (2006). Resolving emotional conflict: a role for the rostral anterior cingulate cortex in modulating activity in the amygdala. Neuron, 51(6), 871-882.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. Science, 299(5614), 1898-1902.

Frank, M. J., & O'Reilly, R. C. (2006). A mechanistic account of striatal dopamine function in human cognition: psychopharmacological studies with cabergoline and haloperidol. Behav Neurosci, 120(3), 497-517.

Friston, K. (2005). A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci, 360(1456), 815-836.

Fuster, J. M. (2008). The prefrontal cortex. London: Academic press.

Grandjean, J., D'Ostilio, K., Phillips, C., Balteau, E., Degueldre, C., Luxen, A., . . . Collette, F. (2012). Modulation of brain activity during a Stroop inhibitory task by the kind of cognitive control required. PLoS One, 7(7), e41513.

Gratton, G., Coles, M. G., & Donchin, E. (1992). Optimizing the use of information: strategic control of activation of responses. Journal of experimental psychology. General, 121(4), 480-506.

Gratton, G., Coles, M. G., Sirevaag, E. J., Eriksen, C. W., & Donchin, E. (1988). Pre- and poststimulus activation of response channels: a psychophysiological analysis. J Exp Psychol Hum Percept Perform, 14(3), 331-344.

Hommel, B., Proctor, R. W., & Vu, K. P. (2004). A feature-integration account of sequential effects in the Simon task. Psychol Res, 68(1), 1-17.

Hopfield, J. J. (1982). Neural Networks and Physical Systems with Emergent Collective Computational Abilities. Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences, 79(8), 2554-2558.

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., & Camerer, C. F. (2005). Neural systems responding to degrees of uncertainty in human decision-making. Science, 310(5754), 1680-1683.

Huettel, S. A., Stowe, C. J., Gordon, E. M., Warner, B. T., & Platt, M. L. (2006). Neural signatures of economic preferences for risk and ambiguity. Neuron, 49(5), 765-775.

Ide, J. S., Shenoy, P., Yu, A. J., & Li, C. S. (2013). Bayesian prediction and evaluation in the anterior cingulate cortex. J Neurosci, 33(5), 2039-2047.

Ipeirotis, P. G. (2010). Demographics of Mechanical Turk (Tech. Rep. No. Ce-DER-10-01).

Jiang, J., & Egner, T. (2013). Using Neural Pattern Classifiers to Quantify the Modularity of Conflict-Control Mechanisms in the Human Brain. Cereb Cortex. doi: 10.1093/cercor/bht029

Jiang, J., Schmajuk, N., & Egner, T. (2012). Explaining neural signals in human visual cortex with an associative learning model. Behav Neurosci, 126(4), 575-581.

Jiang, J., Summerfield, C., & Egner, T. (2013). Attention sharpens the distinction between expected and unexpected percepts in the visual brain. J Neurosci, 33(47), 18438-18447.

Johan Masreliez, C., & Martin, R. D. (1977). Robust Bayesian estimation for the linear model and robustifying the Kalman filter. IEEE Trans. Automatic Control, 22(3), 361-371.

Kerns, J. G. (2006). Anterior cingulate and prefrontal cortex activity in an FMRI study of trial-to-trial adjustments on the Simon task. Neuroimage, 33(1), 399-405.

Kerns, J. G., Cohen, J. D., MacDonald, A. W., 3rd, Cho, R. Y., Stenger, V. A., & Carter, C. S. (2004). Anterior cingulate conflict monitoring and adjustments in control. Science, 303(5660), 1023-1026.

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. Science, 324(5928), 759-764.

King, J. A., Korb, F. M., von Cramon, D. Y., & Ullsperger, M. (2010). Post-error behavioral adjustments are facilitated by activation and suppression of task-relevant and task-irrelevant information processing. J Neurosci, 30(38), 12759-12769.

Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. Science, 302(5648), 1181-1185.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. Proc Natl Acad Sci U S A, 103(10), 3863-3868.

Krug, M. K., & Carter, C. S. (2012). Proactive and reactive control during emotional interference and its relationship to trait anxiety. Brain Res, 1481, 13-36.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. Journal of the Optical Society of America. A, Optics, image science, and vision, 20(7), 1434-1448.

Liu, X., Banich, M. T., Jacobson, B. L., & Tanabe, J. L. (2004). Common and distinct neural substrates of attentional control in an integrated Simon and spatial Stroop task as assessed by event-related fMRI. Neuroimage, 22(3), 1097-1106.

Locke, H. S., & Braver, T. S. (2008). Motivational influences on cognitive control: behavior, brain activation, and individual differences. Cogn Affect Behav Neurosci, 8(1), 99-112.

Logan, G. D., & Zbrodoff, N. J. (1979). When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a Stroop-like task. Memory and Cognition, 7, 166-174.

MacDonald, A. W., 3rd, Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. Science, 288(5472), 1835-1838.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. Psychol Bull, 109(2), 163-203.

MacLeod, C. M., & MacDonald, P. A. (2000). Interdimensional interference in the Stroop effect: uncovering the cognitive and neural anatomy of attention. Trends Cogn Sci, 4(10), 383-391.

Mayr, U., Awh, E., & Laurey, P. (2003). Conflict adaptation effects in the absence of executive control. Nat Neurosci, 6(5), 450-452.

McCoy, A. N., & Platt, M. L. (2005). Risk-sensitive neurons in macaque posterior cingulate cortex. Nat Neurosci, 8(9), 1220-1227.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. Annu Rev Neurosci, 24, 167-202.

Monosov, I. E., & Hikosaka, O. (2013). Selective and graded coding of reward uncertainty by neurons in the primate anterodorsal septal region. Nat Neurosci.

Mozer, M. C., Colagrosso, M. D., & Huber, D. E. (2002). A rational analysis of cognitive control in a speeded discrimination task. Advances in Neural Information Processing Systems 14, Vols 1 and 2, 14, 51-57.

Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. Biol Cybern, 66(3), 241-251.

Niendam, T. A., Laird, A. R., Ray, K. L., Dean, Y. M., Glahn, D. C., & Carter, C. S. (2012). Meta-analytic evidence for a superordinate cognitive control network subserving diverse executive functions. Cogn Affect Behav Neurosci, 12(2), 241-268.

Norman, D. A., & Shallice, T. (1986). Attention to action: willed and automatic control of behavior. In G. E. Schwarz & D. Shapiro (Eds.), Consciousness and self-regulation (Vol. 4). New York: Plenum Press.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychol Rev, 87(6), 532-552.

Pearl, J. (1988). Probabilistic reasoning in intelligent systems : networks of plausible inference. San Mateo, Calif.: Morgan Kaufmann Publishers.

Preuschoff, K., Bossaerts, P., & Quartz, S. R. (2006). Neural differentiation of expected reward and risk in human subcortical structures. Neuron, 51(3), 381-390.

Preuschoff, K., Quartz, S. R., & Bossaerts, P. (2008). Human insula activation reflects risk prediction errors as well as risk. J Neurosci, 28(11), 2745-2752.

Rao, R. P. (2005). Bayesian inference and attentional modulation in the visual cortex. Neuroreport, 16(16), 1843-1848.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci, 2(1), 79-87.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. Psychol Rev, 106(2), 261-300.

Reeck, C., & Egner, T. (2011). Affective privilege: asymmetric interference by emotional distracters. Front Psychol, 2, 232.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black & W.F. Prokasy (Eds.), Classical Conditioning II (pp. 64-99). New York: Appleton-Century-Crofts.

Reynolds, J., & Mozer, M. (2009). Temporal dynamics of cognitive control. Paper presented at the Advances in neural information processing systems.

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. Science, 306(5695), 443-447.

Schmajuk, N. A., Lam, Y. W., & Gray, J. A. (1996). Latent inhibition: A neural network approach. Journal of Experimental Psychology-Animal Behavior Processes, 22(3), 321-349.

Schmidt, J. R., & Besner, D. (2008). The Stroop effect: why proportion congruent has nothing to do with congruency and everything to do with contingency. J Exp Psychol Learn Mem Cogn, 34(3), 514-523.

Schmidt, J. R., & Weissman, D. H. (Submitted). Congruency sequence effects without feature integreation or contingency learning confounds.

Servan-Schreiber, D., Bruno, R. M., Carter, C. S., & Cohen, J. D. (1998). Dopamine and the mechanisms of cognition: Part I. A neural network model predicting dopamine effects on selective attention. Biological psychiatry, 43(10), 713-722.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. Neuron, 79(2), 217-240.

Shenoy, P., Rao, R., & Yu, A. J. (2010). A rational decision making framework for inhibitory control. Paper presented at the Advances in neural information processing systems, Boston:MIT.

Shenoy, P., & Yu, A. J. (2011). Rational decision-making in inhibitory control. Front Hum Neurosci, 5, 48.

Silvetti, M., Alexander, W., Verguts, T., & Brown, J. W. (2013). From conflict management to reward-based decision making: Actors and critics in primate medial frontal cortex. Neurosci Biobehav Rev. doi: 10.1016/j.neubiorev.2013.11.003

Silvetti, M., Seurinck, R., & Verguts, T. (2011). Value and prediction error in medial frontal cortex: integrating the single-unit and systems levels of analysis. Front Hum Neurosci, 5, 75.

Silvetti, M., Seurinck, R., & Verguts, T. (2013). Value and prediction error estimation account for volatility effects in ACC: a model-based fMRI study. Cortex, 49(6), 1627-1635.

Sohn, M. H., Ursu, S., Anderson, J. R., Stenger, V. A., & Carter, C. S. (2000). The role of prefrontal cortex and posterior parietal cortex in task switching. Proc Natl Acad Sci U S A, 97(24), 13448-13453.

Summerfield, C., Behrens, T. E., & Koechlin, E. (2011). Perceptual classification in a rapidly changing environment. Neuron, 71(4), 725-736.

Sutton, R. S. (1988). Learning to predict by the method of temporal dierences. Machine Learning, 3(1), 9-44.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. Trends Cogn Sci, 10(7), 309-318.

Tenenbaum, J. B., & Xu, F. (2000). Word learning as Bayesian inference. Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society, 517-522.

Torres-Quesada, M., Franziska, K. M., Funes, M. J., Lupianez, J., & Egner, T. (2014). Comparing neural substrates of emotional vs. non-emotional conflict modulation by global control context. Front. Hum. Neurosci.

Torres-Quesada, M., Funes, M. J., & Lupianez, J. (2013). Dissociating proportion congruent and conflict adaptation effects in a Simon-Stroop procedure. Acta Psychol (Amst), 142(2), 203-210.

Turk-Browne, N. B. (2013). Functional interactions as big data in the human brain. Science, 342(6158), 580-584.

Tzelgov, J., Henik, A., & Berger, J. (1992). Controlling Stroop effects by manipulating expectations for color words. Memory & cognition, 20(6), 727-735.

Verguts, T., & Notebaert, W. (2008). Hebbian learning of cognitive control: dealing with specific and nonspecific adaptation. Psychol Rev, 115(2), 518-525.

Verguts, T., & Notebaert, W. (2009). Adaptation by binding: a learning account of cognitive control. Trends Cogn Sci, 13(6), 252-257.

Vilares, I., & Kording, K. (2011). Bayesian models: the structure of the world, uncertainty, behavior, and the brain. Ann N Y Acad Sci, 1224, 22-39.

Vossel, S., Mathys, C., Daunizeau, J., Bauer, M., Driver, J., Friston, K. J., & Stephan, K. E. (2013). Spatial Attention, Precision, and Bayesian Inference: A Study of Saccadic Response Speed.

Weissman, D. H., Jiang, J., & Egner, T. (submitted). Determinants of congruency sequence effects without learning and memory confounds.

White, C. N., Brown, S., & Ratcliff, R. (2012). A test of Bayesian observer models of processing in the Eriksen flanker task. J Exp Psychol Hum Percept Perform, 38(2), 489-497.

White, C. N., Ratcliff, R., & Starns, J. J. (2011). Diffusion models of the flanker task: discrete versus gradual attentional selection. Cogn Psychol, 63(4), 210-238.

Wilk, H. A., Ezekiel, F., & Morton, J. B. (2012). Brain regions associated with moment-to-moment adjustments in control and stable task-set maintenance. Neuroimage, 59(2), 1960-1967.

Wittfoth, M., Buck, D., Fahle, M., & Herrmann, M. (2006). Comparison of two Simon tasks: neuronal correlates of conflict resolution based on coherent motion perception. Neuroimage, 32(2), 921-929.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. Psychol Rev, 114(2), 245-272.

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. Neuron, 46(4), 681-692.

Yu, A. J., Dayan, P., & Cohen, J. D. (2009). Dynamics of attentional selection under conflict: toward a rational Bayesian account. J Exp Psychol Hum Percept Perform, 35(3), 700-717.

Biography

Jiefeng Jiang was born on August 26, 1981 in Mianyang, Sichuan province, China. He obtained his bachelor's degree in computer science and engineering at Zhejiang University, China in 2003. He then attended the Institute of Automation, Chinese Academy of Sciences, and received his master's degree in pattern recognition and intelligent systems in 2009. Since fall 2009, Jiefeng is a Ph.D student in the Cognitive Neuroscience Admitting Program and the department of Psychology and Neuroscience at Duke University. Working with his advisor, Dr. Tobias Egner, Jiefeng is focused on the mechanisms of cognitive control and predictive coding.

Publications

Jiang, J., Summerfield, C., Egner, T. (2013). Attention sharpens the distinction between expected and unexpected percepts in the visual brain. Journal of Neuroscience, 33(47):18438-18447.

Jiang, J. & Egner, T.(2013). Using neural pattern classifiers to quantify the modularity of conflict-control mechanisms in the human brain. Cerebral Cortex, Doi: 10.1093/cercor/bht029.

Jiang, J., Schmajuk, N., Egner, T. (2012). Explaining neural signals in human visual cortex with an associative learning model. Behavioral Neuroscience, 126(4), 575-581.

Jiang, J., Zhu, W., Shi, F., Liu, Y., Li, J., Qin, W., Li, K., Yu, C., Jiang, T. (2009). Thick Visual Cortex in the Early Blind. The Journal of Neuroscience 29(7): 2205–2211.

Jiang, J., Zhu, W., Shi, F., Zhang, Y., Lin, L., Jiang, T. (2008). A robust and accurate algorithm for estimating the complexity of the cortical surface. Journal of Neuroscience Methods 172: 122–130.

Jiang, X., Liu, B., Jiang, J., Zhao, H., Fan, M., Zhang, J., Fan, Z., Jiang, T. (2008). Modularity in the genetic disease-phenotype network. FEBS Letters 582(2008): 2549–2554.

Jiang, T., Liu, Y., Shi, F., Shu, N., Liu, B., Jiang, J., Zhou, J. (2008). Multimodal Magnetic Resonance Imaging for Brain Disorders: Advances and Perspectives. Brain Imaging and Behavior 2: 249–257.

Zhang, Y., Jiang, J., Lin, L., Shi, F., Zhou, Y., Yu, C., Li, K., Jiang, T. (2008). A Surface-Based Fractal Information Dimension Method for Cortical Complexity Analysis. in T. Dohi, I. Sakuma, and H. Liao (Eds.): MIAR 2008, LNCS 5128, pp. 133–141, Springer-Verlag Berlin Heidelberg.

Li, X., Jiang, J., Zhu, W., Yu, C., Sui, M., Wang, Y., Jiang, T. (2007). Asymmetry of prefrontal cortical convolution complexity in males with attention-deficit/hyperactivity disorder using fractal information dimension. Brain & Development 29: 649–655.

Shi, F., Liu, Y., Jiang, T., Zhou, Y., Zhu, W., Jiang, J., Liu, H., Liu, Z. (2007). Regional Homogeneity and Anatomical Parcellation for fMRI Image Classification: Application to Schizophrenia and Normal Controls. in N. Ayache, S. Ourselin, A. Maeder (Eds.): MICCAI 2007, Part II, LNCS 4792, pp. 136–143, Springer-Verlag Berlin Heidelberg.