Computational Psychiatry: Combining multiple levels of analysis to understand brain disorders.

by

Thomas Viktor Wiecki

Diploma, University of Tübingen, January 2010

M. Sc., Brown University, May 2012

A Dissertation submitted in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy
in the Cognitive, Linguistic & Psychological Sciences at Brown University

Providence, Rhode Island

May 2015

This dissertation by Thomas Viktor Wiecki is accepted in its present form by the Cognitive, Linguistic & Psychological Sciences as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date 9/23/2014

Michael J. Frank, Director

Recommended to the Graduate Council

Date 9/24/2014

Thomas Serre, Reader

Date 9/24/2014 Erik B Sudderth

Erik Sudderth, Reader

Date 9-25-14

Benjamin Greenberg, Reader

Approved by the Graduate Council

Date ______

Peter M. Weber, Dean of the Graduate School

Acknowledgements

This dissertation would not have been possible without the support of many people and institutions throughout the last several years. First and foremost I am grateful to my supervisor and friend Michael J Frank who contributed to my development as a researcher in a major way. I am also thankful to the other members of my thesis committee – Thomas Serre, Erik Sudderth and Ben Greenberg – for thoughtful feedback. Daniel Dillon, as well as Sara Tabrizi, Chrystalina Antoniades, Chris Kennard, Beth Borowsky, Monica Lewis, and Mina Creathorn deserve my gratitude for generously sharing their clinical data with me and helpful discussions. In addition, I am grateful to the members of the Frank Lab and fellow grad students and faculty at Brown. Specifically, Imri Sofer and Christopher Chatham have contributed in a major way through many thoughtful discussions and their friendship.

I would like to thank my friends and family for their continuous support. Finally, I am grateful to my wife, Emiri, for moving across the Atlantic to join me, and for being the source of inspiration for this research. To her I dedicate these pages.

Abstract

The premise of the emerging field of computational psychiatry is to use models from computational cognitive neuroscience to gain deeper insights into mental illness. In this thesis my goal is to provide an overview of this endeavor and advance it by developing new software as well as quantitative methods. To demonstrate their usefulness I will apply these methods to real-world data sets. A central theme will be the bridging of multiple levels of analysis of the brain ranging from neuroscience and cognition to behavior. In chapter 1 I describe the current crisis in research and treatment of mental illness and argue that computational psychiatry provides the tools to solve some long-standing issues that hindered progress in this area. I describe these tools by reviewing the current literature on computational psychiatry and demonstrate their usefulness on two real-world data sets. To provide a coherent scope, I will focus on response inhibition as it provides a rich literature in each of the different levels of analysis with clear links to psychopathology. In chapter 2 I first establish a neuronal basis by presenting a biologically plausible neural network model of key areas involved in response inhibition. Capturing the high-level computations of this fairly complex model requires more abstract cognitive process models. Towards this goal we developed software (chapter 3) to estimate a decision making model in a hierarchical Bayesian manner which improves parameter recovery in a simulation study. In chapter 4 I then bridge the neuronal and cognitive level by fitting a psychological process model to the simulated behavioral output of the neural network model under certain biological manipulations. By analyzing which biological manipulation is best captured by changes in certain high-level computational parameters I start to link both levels of analysis. I then apply this same psychological process model to two data sets from selective response inhibition tasks administered to patients suffering from Huntington's disease (chapter 5) and depression (chapter 6). Having identified neurobiological correlates of certain model parameters allows to then formulate theories not only about cognitive processes impacted by these disorders but also which neuronal mechanism are likely to be involved. In addition, I demonstrate that the description of subjects' performance by computational model parameters can lead to better classification accuracy of disease state when compared to traditionally used summary statistics.

Contents

List of Tables

List of Figures

Chapter 1

Model-based cognitive neuroscience approaches to computational psychiatry: clustering and classification

This chapter has been published and reflects contributions of other authors:

Wiecki, T.V., Poland, J.S. & Frank, M.J. (in press). Model-based cognitive neuroscience approaches to computational psychiatry: clustering and classification. *Clinical Psychological Science.*

1.1 Abstract

Psychiatric research is in crisis. We highlight efforts to overcome current challenges, focusing on the emerging field of computational psychiatry, which might enable us to move from a symptom-based description of mental illness to descriptors based on objective computational multidimensional functional variables. We survey recent efforts

towards this goal, and describe a set of methods that together form a toolbox to aid this research program. We identify four levels in computational psychiatry: (i) Behavioral tasks indexing various psychological processes; (ii) computational models that identify the generative psychological processes; (iii) parameter estimation methods concerned with quantitatively fitting these models to subject behavior, focusing on hierarchical Bayesian estimation as a rich framework with many desirable properties; and (iv) machine learning clustering methods which identify clinically significant conditions and sub-groups of individuals. As a proof of principle we apply these methods to two different data sets. Finally, we highlight challenges for future research.

1.2 Motivation

Imagine going to a doctor because of chest-pain that has been bothering you for a couple of weeks. The doctor would sit down with you, listen carefully to your description of symptoms and prescribe medication to lower blood pressure in case you have a heart condition. After a couple of weeks your pain has not subsided. The doctor now prescribes medication against reflux, which finally seems to help. In this scenario not a single medical analysis (e.g. EKG, blood work or a gastroscopy) was performed and medication with potentially severe side-effects prescribed on a trial-and-error basis. While highly unlikely to occur if you walked into a primary care unit with these symptoms today, this scenario resembles much of contemporary psychiatry diagnosis and treatment.

There are several reasons for this discrepancy in sophistication between psychiatry and other fields of medicine. First and foremost, mental illness affects the brain the most complex biological system yet encountered. Compared to the level of scientific understanding achieved on other organs of the human body such as the heart, our understanding of the normally functioning brain is still, arguably, in its infancy.

Despite this complexity, concerted efforts in the brain sciences have led to an explosion of knowledge and understanding about the healthy and diseased brain in the last decades. The discovery of highly effective psychoactive drugs in the 50s and 60s raised expectations that psychiatry would progress in a similar fashion. Unfortunately, in retrospect it appears that these discoveries were serendipitous in nature as little progress has been made since (e.g. Insel et al., 2010; Hyman, 2012b). This lack of progress also caused many major pharmaceuticals companies like AstraZeneca and GlaxoSmithKline to withdraw from psychiatric drug development and to close large research centers (Nutt and Goodwin, 2011; Cressey, 2011). While treatment of psychiatric disorders is arguably quite effective, research on mental illness, based on conventional psychiatric diagnostic categories and practices (as reflected in the DSM), has been widely viewed as disappointing, and the DSM system of classification itself has been viewed as an impediment to more productive research. As a consequence, psychiatry is a field in crisis (Poland et al., 1994; Insel et al., 2010; Hyman, 2012b; Sahakian et al., 2010). As outlined in more detail below, a central issue is a lack of sufficiently powerful theoretical and methodological resources for managing the features of mental illness (e.g., a lack of measurable quantitative descriptors). This lacuna prevents effective management of the multidimensional hierarchical complexity, dynamic interactivity, causal ambiguity, and heterogeneity of mental illness. And, it leads to an explanatory gap of how basic neurobiological processes and other causes result in complex disorders of the mind (Montague et al., 2011; Hyman, 2012a).

Below we will review current challenges in psychiatry and recent efforts to overcome them. Several examples from the domain of decision making show the promise of moving away from symptom-based description of mental illness and instead formulating objective, quantifiable computational biomarkers as a basis for further psychiatric research. We then introduce a computational cognitive toolbox that is suited to construct these computational biomarkers. We focus on sequential sampling models of decision making, which serve as a case study for how computational models, when fit to behavior, have successfully been used to identify and quantify latent neurocognitive processes in healthy humans. Bayesian methods provide a resourceful framework to fit these models to behavior and establish individualized descriptors of neurocognitive function. After establishing the validity of these models to provide neurocognitive descriptors of individuals, we will review how clustering techniques can be used to construct a map of individual differences based on these neurocognitive descriptors.

To demonstrate the viability and potential of these methods we reanalyze two data-sets, providing a proof of principle before discussing future challenges in application to psychiatric populations. The first data-set consists of a group of young and old subjects performing three different decision making tasks (Ratcliff et al., 2010). After fitting participant's choices and response time distributions with the drift diffusion model using hierarchical Bayesian parameter estimation, each participant's parameter estimates are provided as inputs to an unsupervised clustering algorithm. We show that the clustering is sensitive to age after regressing out nuisance variables, and that this clustering shows consistently better recovery of the age groups than when using behavioral summary statistics (e.g. mean RT and accuracy) alone. Moreover, factor analysis on the computational parameters extracts meaningful latent variables that describe cognitive ability. For this dataset, no identified brain-based mechanism was analyzed. In contrast, for the second data-set we rely on a hypothesis-driven approach that suggested a mechanism for how a specific decision parameter -- the decision threshold -- varies as a function of activity communicated between frontal cortex and the subthalamic nucleus (STN). A previous study showed that STN deep brain stimulation disrupted decision threshold regulation across a group of patients with Parkinson's disease (Cavanagh et al., 2011). Below we show that we can classify individual patients' brain stimulation status (off or on) with relatively high accuracy given model parameters, and better than that achieved based on brain-behavior correlations alone.

1.2.1 Current challenges in psychiatry

The current crisis in psychiatry has complex causes that are deeply rooted in existing classification systems (e.g., DSM, ICD). In this section we identify some of the problems these systems introduce, and provide indications of the sorts of resources required for more productive research programs. In the subsequent section we review recent attempts to meet these challenges and the sorts of resources that have been introduced for this purpose. As others before us have done, we proceed to suggest an approach to research of mental disorders which aims to link cognitive and pure neuroscience to mental illness without the restrictions of prior classification schemes (Poland and Von Eckardt, 2013; Cuthbert and Insel, 2010; Robbins et al., 2012).

Diagnostic and Statistical Manual of Mental Disorders and Research

For decades the Diagnostic and Statistical Manual of Mental Disorders (DSM) has been the basis of clinical diagnosis, treatment and research of mental illness. At its core, the DSM defines distinct disorder categories like schizophrenia (SZ) and depression in a way that is atheoretical (i.e., with no reference to specific causal hypotheses) and focused on clinical phenomenology. Thus, these categories are mainly derived from translating subjective experience to objective symptomatology (Nordgaard et al., 2012) while assuming unspecified biological, psychological, or behavioral dysfunctions (Poland et al., 1994).

While primarily intended to be of value to clinicians, the DSM has also played a substantial role as a classification system for scientific research with the goals of validating the diagnostic categories and translating research results directly into clinical practice. While these research goals are commendable, decisions regarding systematic classification are more often based on perceptions of clinical utility rather than scientific merit (Poland and Von Eckardt, 2013). As a consequence, DSM-based research programs have failed to deliver consistent, replicable and specific results, and it has been widely observed that the validation of DSM categories has been limited, that DSM categories do not provide well defined phenotypes, and that they have limited research utility (Kendell and Jablensky, 2003; Andreasen, 2007; Regier and Narrow, 2009; Kendler et al., 2009; Cuthbert and Insel, 2010; Hyman, 2010).

Heterogeneity and Comorbidity

One major problem of contemporary psychiatric classification is the heterogeneity of individuals receiving identical diagnoses. With respect to symptomatology, one striking example of this is Schizophrenia where one must exhibit at least 2 out of 5 symptoms to receive a diagnosis (Heinrichs, 2001). It is thus possible to have patients with completely different symptomatology being diagnosed as schizophrenic. It is important, however, that problems of heterogeneity concern more than just symptoms; there is probably heterogeneity at all levels of analysis including heterogeneity of causal processes (Poland, et al 1994). And, as we shall see below, such heterogeneity is not just a feature of clinical populations, but may be a feature of the general population. As a consequence, heterogeneity poses a serious challenge for research (e.g., it introduces uncontrolled sources of variance, it limits the generalizability of results) and points to the necessity of developing techniques for its management.

Comorbidity is widely believed to constitute a second major problem for psychiatric classification. Defined as the co-occurrence of multiple disorders in one individual, it has been widely documented (Markon, 2010; Krueger and Markon, 2006) that comorbidity between mental disorders is the rule rather than the exception, invading nearly all canonical diagnostic boundaries. (Buckholtz and Meyer-Lindenberg, 2012). It is important to differentiate between two relevant types of comorbidity: (i) True comorbidity is a result of independent disorders co-occurring; (ii) artificial comorbidity is a result of separately classifying disorders that have overlapping symptom criteria, have a common cause, or share a pathogenic cascade. This distinction points to a more general problem concerning the management of causal ambiguity that is found at the level of symptoms, but also at other levels of analysis. Specifically, the problem is one of identifying which causal structures and processes produce a given clinical presentation or a given pattern of functioning at some other level; because clinical presentations and patterns of functioning can be produced by different causal structures and processes, the challenge for researchers is to develop techniques for identifying and managing such causal ambiguity.

In addition to challenges of heterogeneity and comorbidity, there are several other features of the domain of mental illness that pose challenges to research and require sophisticated tools and techniques for their effective management. These include, hierarchical organization of the brain and various sorts of inter-level relationship and coordination (e.g., ”the explanatory gap), dynamic interactivity, multidimensional complexity, context sensitivity, identification of norms of functioning, and identification of meaningful groupings of individuals. As we shall see below, each of these features creates problems that contribute to an understanding of why the current crisis in research exists and of the sorts of resources and strategies required for more productive research programs.

1.3 Potential Solutions

As outlined above, the short-comings of the current DSM classification system and the problems they pose for research are well documented. In the following we will outline some current efforts to address these challenges.

1.3.1 Research Domain Criteria Project (RDoC) and A Roadmap for Mental Health Research in Europe (ROAMER)

The Research Domain Criteria Project (RDoC) is an initiative by the National Institute for Mental Health (NIMH) (Insel et al., 2010). RDoC improves on previous research efforts based on the DSM in the following ways. First, as the name implies it is conceptualized as a research framework only and is thus clearly separated from clinical practice. Second, RDoC is completely agnostic about DSM categories. Instead of a top-down approach which aims at identifying neural correlates of psychiatric disorders, RDoC suggests a bottom-up approach that builds on the current understanding of neurobiological underpinnings of different cognitive processes and links those to clinical phenomena. Third, the RDoC research program integrates data from different levels of analysis like imaging, behavior and self-reports.

At its core, RDoC is structured into a matrix with columns representing different units of analysis and rows for research domains. The units of analysis include genes, molecules, cells, circuits, physiology, behavior, and self-reports. Research domains are clustered into negative and positive valence systems, cognitive systems, systems for social processes and arousal/regulatory systems. Each of these domains is further subdivided into distinct processes; for example, cognitive systems include attention, perception, working memory, declarative memory, language behavior and executive control.

Despite clear improvements over previous DSM-based research programs, the RDoC initiative currently lacks explicit consideration of computational descriptors (Poland and Von Eckardt, 2013). As outlined below, computational methods show great promise to help link different levels of analysis, elucidate clinical symptoms and identify sub-groups of healthy and patient populations.

More recently, the European Commission started the Roadmap for Mental Health

Research in Europe (ROAMER) initiative with the goal of better integrating biomedicine, psychology, and public health insights to further research into mental illnesses (Schumann et al, 2014).

1.3.2 Neurocognitive phenotyping

In a recent review article, Robbins et al. (2012) suggest the use of neurocognitive endophenotypes to study mental illness: Neurocognitive endophenotypes would furnish more quantitative measures of deficits by avoiding the exclusive use of clinical rating scales, and thereby provide more accurate descriptions of phenotypes for psychiatric genetics or for assessing the efficacy of novel treatments. (pg. 82)

Of particular interest are three studies that use such neurocognitive endophenotypes by constructing multi-dimensional profiles (MPs) from behavioral summary statistics across a battery of various neuropsychological tasks used to identify subtypes of ADHD (Durston et al., 2008; Sonuga-Barke, 2005; Fair et al., 2012).

Durston et al. (2008) argues that there are distinct pathogenic cascades within at least three different brain circuits that can lead to symptomatology involved in ADHD. Specifically, abnormalities in dorsal frontostriatal, orbito-frontostriatal, or fronto-cerebellar circuits can lead to impairments of cognitive control, reward processing and timing, respectively. Core deficits in one or multiple of these brain networks can thus result in a clinical diagnosis of ADHD and provides a compelling explanation for the heterogeneity of the ADHD patient population. Preliminary evidence for this hypothesis is provided by Sonuga-Barke (2005) who used principal component analysis (PCA) on multi-dimensional profiles (based on a neuropsychological task battery) of ADHD patients and identified 3 distinct sub-types co-varying on timing, cognitive control, and reward.

A similar approach of identifying clusters in the ADHD population using MPs was taken by Fair et al. (2012). The authors applied graph theory to identify individual behavioral functional clusters not only within the ADHD patient population but also within healthy controls (HC). Interestingly, the authors found that HC and ADHD is not the predominant dimension along which clusters form. Instead, the authors uncovered different functional profiles (e.g. one cluster might show differences in response inhibition while another one shows differences in RT variability), each of which contained both healthy and patient sub-groups. Nevertheless, and critically, a classifier trained to predict diagnostic category achieved better performance when classifying within each functional profile than a classifier trained on the aggregated data. In other words, this implies that the overall population clusters into different cognitive profiles, and ADHD affects individuals differently based on which cognitive profile they exhibit. Importantly, this study suggests that the source of heterogeneity may not only be distinct pathogenic cascades being labeled as the same disorder but may actually be a result of the inherent heterogeneity present in the overall population healthy and disordered.

The above studies all exemplify the danger of lumping subjects at the level of symptoms and treating them as one homogeneous category with a single, identifiable pathological cascade. Instead, these studies use MPs to find an alternative characterization of subjects independent of their DSM classification that is (i) quantitatively measurable, (ii) a closer approximation to the underlying neurocircuitry (Robbins et al., 2012), and (iii) cognizant of heterogeneity in the general population.

Nevertheless, this approach still has problems. First, although there is less reliance on DSM categories, these studies still use the diagnostic label for recruiting subjects, selecting tasks, framing and testing hypotheses, and drawing inferences. It could be imagined, for example, that patients with compulsive disorders like OCD or Tourettes have abnormalities in similar brain circuits, and consequently pathologies, deficits and impairments may crosscut these (and other) diagnostic categories. Thus, if only ADHD patients are recruited, a critical part of the picture might be missed. Second, the cognitive task battery only covers certain aspects of cognitive function. Other tasks that for example measure working memory or reinforcement learning, both of which involve fronto-striatal function, would be a useful addition to help resolve causal ambiguity. More specifically, performance on each individual task is assessed by an aggregate performance score. Recent behavioral and neuropsychological findings, however, suggest that executive control (for example) in a single task may instead be more accurately characterized as a collection of related but separable abilities (Baddeley, 1966; Collette et al., 2005), a pattern referred to as the unity and diversity of executive functions (Duncan et al., 1997; Miyake et al., 2000). Further, most cognitive tasks rely on a concerted and often intricate interaction of various neural networks and cognitive processes (see e.g. Collins and Frank, 2012). This task impurity problem (Burgess, 1997; Phillips, 1997) complicates identification of separate functional impairments and brain circuits based solely on MPs.

In sum, while cognitive phenotypes provide a useful framework for measuring brain function there is still ambiguity when using behavioral scores that provide an aggregate measure of various brain networks. The idea that a neural circuit can contribute to different cognitive functions helps explain why diverse mental illnesses can exhibit similar symptoms (comorbidity)(Buckholtz and Meyer-Lindenberg, 2012). Disentangling these transdiagnostic patterns of psychiatric symptoms thus requires identification and measurement of underlying brain circuits and functions. While Buckholtz and Meyer-Lindenberg propose the use of functional imaging studies and genetic analysis we will discuss how computational modeling can contribute to disambiguate the multiple pathways leading to behavioral features.

1.3.3 Computational psychiatry

Computational models at different levels of abstraction have had tremendous impact on the field of cognitive neuroscience. The aim is to construct models based on integrated evidence from neuroscience and psychology to explain neural activity as well as cognitive processes and behavior. While more detailed biologically inspired models such as biophysical and neural network models are generally more constrained by neurobiology, they often have many parameters which make them less suitable to fit them directly to human behavior. More abstract, algorithmic models on the other hand often have fewer parameters that allow them to be fit directly to data at the cost of being less detailed about the neurobiology. Normal linking of one level of analysis to another is useful to identify plausible neural mechanisms that can be tested with quantitative tools (for review; Frank, in press). Critically, all of these models allow for increased specificity in the identification of different neuronal and psychological processes that are often lumped together when analyzing task behavior based on summary statistics.

The nascent field of computational psychiatry uses computational models to infer dysfunctional latent processes in the brain. Montague et al. (2011) define the goal for computational psychiatry as extract[ing] normative computational accounts of healthy and pathological cognition useful for building predictive models of individuals. [...] . Achieving this goal will require new types of phenotyping approaches, in which computational parameters are estimated (neurally and behaviorally) from human subjects and used to inform the models. (pg. 75). More generally, the tools and techniques of computational cognitive neuroscience (e.g., modeling at multiple levels of analysis, parameter estimation, classification algorithms) are especially well suited for representing and managing the various features of mental illness identified above (e.g., hierarchical and multi-dimensional organization, non-linear dynamic interactivity, context sensitivity, heterogeneity and individual variation, etc.) Thus, computational psychiatry holds out considerable promise as a research program directed at mental illness.

Based on this approach, Maia and Frank (2011) identify computational models as a valuable tool in taming [the complex pathological cascades of mental illness] as they foster a mechanistic understanding that can span multiple levels of analysis and can explain how changes to one component of the system (for example, increases in striatal D2 receptor density) can produce systems-level changes that translate to changes in behavior (pg. 154). Moreover, three concrete strategies for how computational models can be used to study brain dysfunction were defined:

- Deductive approach: Established neuronal or neural circuit models can be tested for how pathophysiologically plausible alterations in neuronal state, e.g. connectivity or neurotransmitter levels (for example, dopamine is known to be reduced in Parkinsons disease), affect system level activations and behavior. This is essentially a bottom-up approach as it involves the study of how known or hypothesized neuronal changes affect higher-level functioning.
- Abductive approach: Computational models can be used to infer neurobiological causes from known behavioral differences. In essence, this is a top-down approach which tries to link behavioral consequences back to underlying latent causes.
- Quantitative abductive approach: Parameters of a computational model are fit to a subjects behavior on a suitable task or task battery. Different parameter values point to differences in underlying neurocircuitry of the associated subject or subject group. These parameters can either be used comparatively to study group differences (e.g. healthy and diseased) or as a regressor with e.g. symptom severity. This approach is more common with abstract models than with neural network models as the former typically have fewer parameters and thus can be more easily fit to data.

Case studies in the domain of decision making

One key area in which computational models have had tremendous success is in elucidating how the different cognitive and neurobiological gears work together in the domain of decision making. Many mental illnesses can be characterized by aberrant decision making of one sort or another (Maia and Frank, 2011; Wiecki and Frank, 2010; Montague et al., 2011). In the following we review recent cases where computational models of decision making have been used to better understand brain disorders.

Computational models of reinforcement learning

Parkinsons Disease

Our first case study concerns Parkinsons disease (PD). Its most visible symptoms affect the motor system as manifest in hypokinesia, bradykinesia, akinesia, rigidity, tremor and progressive motor degeneration. However, recently, cognitive symptoms have received increased attention (e.g., Cools, 2005; Frank, 2005; Moustafa et al., 2008; Cunha et al., 2009). PD is an intriguing neuropsychiatric disorder because its core pathology is well identified to be the cell death of midbrain dopaminergic neurons in the substantia nigra pars compacta (SNc) (Kish et al., 1988). Neural network models of the basal ganglia (BG) (Frank, 2005, 2006) interpret this brain network as an adaptive action selection device that conditionally gates internal or external actions based on their previous reward history, which is learned via dopaminergic signals (Ljungberg et al., 1992; Montague et al., 1996; Schultz, 1998; Waelti et al., 2001; Pan et al., 2005; Bayer et al., 2007; Roesch et al., 2007; Sutton and Barto, 1990; Barto, 1995; Schultz et al., 1997). Behavioral reinforcement learning tasks show that the chronic low levels of DA in PD patients result in a bias towards learning from negative reward prediction errors at the cost of learning from positive reward prediction errors (Frank et al., 2004; Collins & Frank, in press for review). In extension, we have argued that PD is not a motor disorder per se but rather an action selection disorder in which the progressive decline of motor and cognitive function can be interpreted in terms of aberrant learning not to select actions (Wiecki and Frank, 2010; Wiecki et al., 2009; Beeler et al., 2012).

In this case study, an existing biological model of healthy brain function was paired with a known and well localized neuronal dysfunction to extend our understanding of the symptomatology of a brain disorder and to reconceive the nature of the dysfunctions involved. Note, however, that the model was not fit to data quantitatively, nor were multi-dimensional profiles provided to resolve residual causal ambiguity associated with the task impurity problem. In the terminology established by Maia and Frank (2011), this is an example of the deductive approach in which the model provides a mechanistic bridge that explains how abnormal behavior can result from neurocircuit dysfunctions.

Schizophrenia

Despite schizophrenia (SZ) being the focus of intense research over the last decades, no single theory of its underlying neural causes has been able to explain the diverse set of symptoms that lead to a SZ diagnosis. Current psychiatric practices view the symptomatology of SZ as structured in terms of positive symptoms like psychosis, negative symptoms like anhedonia which refers to the inability to experience pleasure from activities usually found enjoyable such as social interaction, and cognitive deficits (Elvevå g and Goldberg, 2000).

Recent progress has been made by the application of RL models to understand individual symptoms or a single symptom category (e.g. negative symptoms) rather than SZ as a whole (Waltz et al., 2011; Gold et al., 2008, 2012; Strauss et al., 2011b).

Using a RL task, Waltz et al. (2007) found that SZ patients show reduced performance in selecting previously rewarded stimuli compared with HCs, and that this performance deficit was most pronounced in patients with severe negative symptoms. Notably, SZ and HC did not differ in their ability to avoid actions leading to negative outcomes. However, due to the task impurity problem, this behavioral analysis did not allow researchers to differentiate whether SZ patients were impaired at learning from positive outcomes or from a failure in representation of the prospective reward values during decision making. The following is a strategy for resolving this problem.

This dichotomy in learning vs representation is also present in two types of RL models actor-critic and Q-learning models (Sutton and Barto, 1998). An actor-critic model consists of two modules: an actor and a critic. The critic learns the expected rewards of states and trains the actor to perform actions that lead to better-than expected outcomes. The actor itself only learns action propensities, in essence stimulus response links. Q-learning models on the other hand learn to associate actions with their reward values in each state. Thus, while a Q-learning model has an explicit representation of which action is most valued in each state, the actor-critic will choose actions based on those that have previously yielded positive prediction errors regardless of whether those arose from an unexpected reward or the absence of an expected loss. Thus, the differences between these two models can be exploited to attempt to resolve the causal ambiguity exhibited by the results above.

In a follow-up study, Gold et al. (2012) administered a new task that paired a neutral stimulus in one context with a positive stimulus and in another context with a negative stimulus. While the neutral stimulus has the same value of zero in both contexts, it is known that DA signals reward prediction errors (RPE) that drive learning in the BG and code outcomes relative to the expected reward (Montague et al., 1996; Schultz et al., 1997). Thus, in the negative context, receiving nothing is better than expected and will result in a positive RPE, driving learning in the BG to select this action in the future (V., 2010). In a test period in which no rewards were presented, participants had to choose between an action that had been rewarding and one that had simply avoided a loss. Both actions should have been associated with better-than-expected outcomes. An actor-critic model should thus show a tendency to select the neutral stimulus while a Q-learning model with representation of the reward contingencies should mainly select the one with a higher reward. Intriguingly, when both of these models were fit to participant data, the actor-critic model produced a better fit for SZ patients with high degree of negative symptoms while HC and SZ with low negative symptoms were better fit by a Q-learning model. In other words, patients with negative symptoms largely based decisions on learned stimulus-response associations instead of expected reward values. Notably, HC and the low negative symptom group did not differ significantly in their RL behavior. This study demonstrates how computational analyses can differentiate between alternative mechanisms that can explain deficiencies in reward-based choice. Many RL tasks can be solved by learning either stimulus-response contingencies or expected reward values (or both), but the model and appropriate task manipulation allows one to extract to which degree these processes are operative, and hence helps to resolve the task impurity problem.

In a related line of work, Strauss et al. (2011a) tested HC and SZ patients on a reinforcement learning task that allowed subjects to either adopt a safe strategy and exploit the rewards of actions with previously experienced rewards, or, to explore new actions with perhaps even higher payoffs. Frank et al. (2009) develop a computational model that can recover how individual subjects balance this exploration-exploitation trade-off. Intriguingly, applying this model to SZ patients, Strauss et al. (2011a) found that patients with high anhedonia ratings were less willing to explore their environment and uncover potentially better actions. This result suggests a reinterpretation of the computational cognitive process underlying lack of social engagement associated with anhedonia. For example, one might assume that the lack of engagement of social activities of anhedonistic patients results from an inability to experience pleasure and as a consequence, a failure to learn the positive value of social interaction. Instead, this study suggests that lack of social engagement associated with anhedonia is a result of an inability to consider the prospective benefit of doing something that might lead to better outcomes. It also leads to the prediction that patients with SZ would not, for example, seek out new social interactions (due to the low value placed on exploration) but could still enjoy social interactions once established. Again, computational strategies allow for a reconceptualization and disambiguation of clinical phenomena.

In sum, Gold et al. (2012) and Strauss et al. (2011a) used a quantitative abductive approach to infer aberrant computational cognitive processes in RL in a subgroup of SZ patients. By grouping subjects according to symptom type and severity instead of diagnosis the authors identified more refined research targets and addressed the problem of heterogeneity. By combining models and strategically designing task demands, Gold, et al pursued an innovative strategy for resolving problems of interpretation resulting from task impurity.

Another relevant line of work includes that of Brodersen et al. (2013) who use dynamic causal modeling (DCM; Friston et al., 2003) – a Bayesian framework for inferring network connectivity between brain areas from fMRI data on healthy and SZ patients performing a numerical n-back working-memory task. Supervised learning methods demonstrated a clear benefit (71% accuracy) of using DCM compared to more traditional methods like functional connectivity (62%). Moreover, clustering methods were sensitive to various SZ subtypes showing the potential of this approach to identify clinically meaningful groups in an unsupervised manner.

Finally, we refer to Huys et al. (2012a) for an example of how a computational psychiatry analysis can be used to relate depressive symptom severity to a specific cognitive process involved in planning multiple future actions.

Computational models of response inhibition

Besides RL, response inhibition is another widely studied phenomenon in cognitive neuroscience of relevance to mental illness. Response inhibition is required when actions in the planning or execution stage are no longer appropriate and must be suppressed. The antisaccade task is one such task that is often used in a psychiatric setting (e.g. Aichert et al., 2012; Fukumoto-Motoshita et al., 2009). It requires subjects to inhibit a prepotent response to a salient stimulus and instead saccade to the opposite side (Hallett, 1979). A wealth of literature has demonstrated reduced performance of psychiatric patients with disorders including attention deficit/hyperactivity disorder (ADHD) (Nigg, 2001; Oosterlaan et al., 1998; Schachar and Logan, 1990), obsessive compulsive disorder (OCD) (Chamberlain et al., 2006; Menzies et al., 2007; Penadés et al., 2007; Morein-Zamir et al., 2009), schizophrenia (SZ) (Huddy et al., 2009; Bellgrove et al., 2006; Badcock et al., 2002), Parkinson's disease (PD) (van Koningsbruggen et al., 2009) and substance abuse disorders (Monterosso et al., 2005; Nigg et al., 2006). However, as demonstrated by Wiecki and Frank (2013), even a supposedly simple behavioral task such as the antisaccade task requires a finely orchestrated interplay between various brain regions including frontal cortex and basal ganglia. It thus can not be said that decreased accuracy in this task is evidence of response inhibition deficits per se as the source of this performance impairment can be manifold (i.e., the antisaccade task exhibits the task impurity problem).

In sum, the use of computational models that allow mapping of behavior to psychological processes could thus be categorized as the computational abductive approach. However, in addition to managing the task impurity problem just mentioned, ambiguity of how psychological processes relate to the underlying neurocircuitry still has to be resolved. By combining different levels of modeling these ambiguities can be better identified and studied (Frank, in press). Ultimately, this might allow development of tasks that use specific conditions (e.g. speed-accuracy trade-off, reward modulations and conflict) to disambiguate the mapping of psychological processes to their neurocircuitry. Using biological process models to test different hypotheses about the behavioral and cognitive effects of neurocircuit modulations would correspond to the deductive approach. In other words, by combining the research approaches outlined by V. and J. (2011) we can use our understanding of the different levels of processing to inform and validate how these levels interact in the healthy and dysfunctional brain.

Thus, there are a few example studies which have applied established computational models to identify model parameters (which aim to describe specific cognitive functions) and relate them to the severity of a specific clinical symptom or use them to identify measureable cognitive impairments. Such targets (viz., specific symptoms, measureable impairments) represent more refined research targets than DSM diagnostic categories. In addition, through the use of strategically designed task batteries and multidimensional s, problems of heterogeneity and task impurity can be managed. And, the combination of various research approaches (e.g., multiple modeling strategies, task batteries and MPs, task manipulations, novel approaches to sampling) can provide a strategic framework for studying relations between neural and computational levels of analysis in mental illness.

1.4 Levels of Computational Psychiatry

The above review has identified a variety of challenges to research concerning mental illness, and it has identified various strategies that have been employed to meet those challenges. Special attention was given to computational psychiatry as an especially promising research program. In all cases, promise for effectively meeting the research challenges depends upon the availability of conceptual and representational resources and associated strategies and techniques that are sufficiently powerful given the features of the domain of mental illness and the problems it poses for research.

In this section, we provide an overview of a four level approach to the computational analysis of cognitive function and dysfunction, focusing on decision making, and sequential sampling models as a concrete example. Such models provide a versatile tool to model cognitive function, but fitting such models to data presents significant technical challenges as well. In the following, we identify four levels of the analysis:

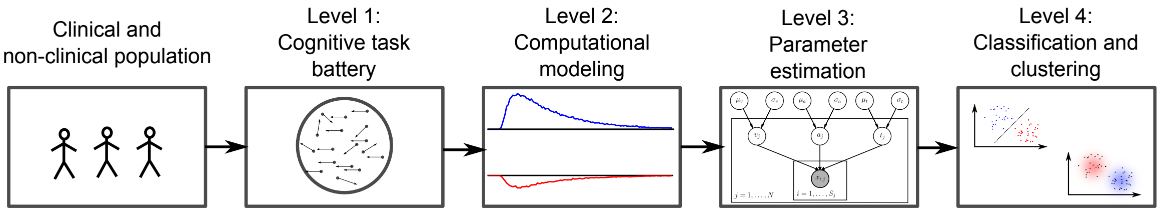
Clinical and non-clinical population
Level 1: Cognitive task battery
Level 2: Computational modeling
Level 3: Parameter estimation
Level 4: Classification and clustering

Figure 1.1: Illustration of the 4 levels of computational psychiatry. Clinical and non-clinical populations are tested on a battery of cognitive tasks. Computational models can relate raw task performance (e.g. RT and accuracy) to psychological and/or neurocognitive processes. These models can be estimated via various methods (depicted is a simplified graphical model of the hierarchical HDDM). Finally, based on the resulting computational multi-dimensional profile we can train supervised and unsupervised learning algorithms to either predict disease state, uncover groups and subgroups in clinical and healthy populations or relate model parameters to clincal symptom severity.

L1 -- strategic identification of cognitive tasks to be employed for the collection of performance data; L2 -- the fitting of computational models to the performance data; L3 -- parameter estimation; and L4 -- identification of clusters and relation to clinical symptom severity (see figure 1.1 for an overview). We show how hierarchical Bayesian modeling and Bayesian mixture models can be deployed to engage a variety of challenges at the various levels of the analysis. Subsequently, we demonstrate the use of these methods on two data sets as a ”proof of concept”. The methods identified in this section have direct applicability to the analysis of cognitive functions in mental illness.

Terminology

- Psychological process model: A computational model that tries to parameterize the cognitive processes underlying behavior. This class of models is not primarily concerned with neural implementations of these processes. Often these models have a parsimonious parameterization which allows them to be fit to behavior.

- Drift-Diffusion Model: An evidence accumulation model used in decision making research.

- Reinforcement learning: Learning to adapt behavior to maximize rewards and minimize punishment.
- Parameter estimation/fitting: The process of finding parameters that best capture the behavior on a certain task.
- Bayesian modeling: A parameter estimation method that allows for great flexibility in defining structure and prior information about a certain domain.
- Comorbidity: The co-occurrence of multiple disorders in one individual.
- Heterogeneity: The fact that there is systematic variation between subjects diagnosed with the same mental illness.
- Task-impurity problem: The fact that no single cognitive task measures just one construct but that task performance is a mixture of distinct cognitive processes.
- Multi-dimensional (MP): A multi-dimensional descriptor of a subject's cognitive abilities as measured by summary statistics (e.g. accuracy) of cognitive tasks spanning multiple cognitive domains.
- Computational multi-dimensional profile (CMP): A MP that includes parameters estimated from a psychological process model that (i) more directly relates to cognitive ability, and (ii) deconstructs different cognitive processes contributing to individual task performance (i.e. task impurity problem).

1.4.1 Level 1: Cognitive tasks

Cognition spans many mental processes that include attention, social cognition, memory, emotion, decision making, and reasoning, to name a few. Various sub-fields devoted to each of these have developed a range of cognitive tasks that purport to reveal the underlying mechanisms. Research in computational psychiatry can draw on these tasks to create task batteries for the collection of performance data usable for the analysis of cognitive function; both the sensitivity and the specificity of tasks to cognitive functions are important characteristics, although the task impurity problem complicates the analysis of data and their use in isolating and specifying cognitive functions. Rather than provide a list of tasks used (see the case-studies above for some examples) we discuss desirable properties that cognitive tasks should exhibit. Ideally, a single cognitive task used in computational psychiatry should be tuned to assess a specific cognitive function, separable from others; this is enabled by:

- a task analysis that identifies what functions are engaged and how they are engaged;
- parsimony in relying on as few cognitive processes as possible;
- stress on cognitive processing in some way to reveal break-off points and allow a sensitive measure of the target function;
- an established theory regarding the neural correlates of the target functions; and
- an established computational model that links behavior to psychological process parameters.

Given the task impurity problem and other forms of causal ambiguity, ideally task batteries should be strategically constructed to measure a range of relevant cognitive functions and other variables to aid in the interpretation of task performance and the isolation of specific functions and dysfunctions. This can be achieved by including co-varying factors (i.e. conditions) in individual tasks that only affect one mental function, which can then be identified. For example, Collins and Frank (2012) were able to separately estimate the contributions of working memory and reinforcement learning in a single task by testing multiple conditions that increased load on working memory alone. Because working memory contributions can contaminate the estimation of the RL component, this manipulation enabled a model to not only capture the WM component, but to better estimate the RL component.

1.4.2 Level 2: Computational models

Computational models in cognitive neuroscience exist on various levels of abstraction, ranging from biophysical neuronal models to abstract psychological process models. While each of these is informative in their own regard in elucidating mental function and dysfunction, we focus here on psychological process models. This class of model has the unique advantage of being simple enough so that they can be fit directly to behavior; that is they are preferred from a statistical analysis point of view given the level of data collected (see Dayan and Abbott, 2005; Frank, in press). The fitted parameters often quantify cognitive ability in terms of psychological process variables rather than behavioral summary statistics. For example, in a simple detection task we might consider the RT speed as a good measure of task performance. However, by adjusting the speed-accuracy trade-off, mean RT can easily be shortened just by increasing the false-alarm rate. Obviously this would not indicate an individuals superior processing abilities. A sequential sampling model analysis, however, would be able to disentangle response caution (i.e. decision threshold) and processing abilities (i.e. drift-rate): these are generative parameters that produce the joint distribution of accuracy and RT. Intuitively, an increase in decision threshold would lead to more accurate but slower responses while an increase in drift-rate would also lead to higher accuracy but also faster responses (Ratcliff and McKoon, 2008). Below we present a simulation experiment that shows how two groups can be clearly separated in their DDM parameters but strongly overlap when described in terms of RT and accuracy summary statistics.

Sequential Sampling models

As outlined above, RL models have already proven to be a valuable tool in explaining neuropsychological disorders and their symptoms. A computational psychiatric framework that aims to explain the multi-faceted domain of mental illness must include computational cognitive neuroscience models that cover a broad range of cognitive processes (see e.g. O'Reilly et al. (2012) for a broad coverage of such models). We will focus on sequential sampling models as an illustrative example of how these models have been applied to study normal and aberrant neurocognitive phenomena, how they can be fit to data using Bayesian estimation, and how subgroups of similar subjects can be inferred using mixture models.

Sequential sampling models (e.g. Townsend and Ashby, 1983a) like the Drift Diffusion Model (DDM) have established themselves as the de-facto standard for modeling data from simple decision making tasks (e.g. Smith and Ratcliff, 2004). Each decision is modeled as a sequential extraction and accumulation of information from the environment and/or internal representations. Once the accumulated evidence crosses a threshold, a corresponding response is executed. This simple assumption about the underlying psychological process has the important property of reproducing not only choice probability and mean RT, but the entire distribution of RTs separately for accurate and erroneous choices in simple two-choice decision making tasks. Interestingly, this evolution of the decision signal in SSMs can also be interpreted as a Bayesian update process (e.g. Gold and Shadlen, 2002; Huang and Rao, 2013; Deneve, 2008; Bitzer et al., 2014). This may be useful because it would place SSMs under a more axiomatic framework and prevent the impression that SSMs are merely convenient heuristics.

The DDM models decision making in two-choice tasks. Each choice is represented as an upper and lower boundary. A drift-process accumulates evidence over time until it crosses one of the two boundaries and initiates the corresponding response (Ratcliff and Rouder, 1998; Smith and Ratcliff, 2004). The speed with which the accumulation process approaches one of the two boundaries is called the drift rate and represents the relative evidence for or against a particular response. Because there is noise in the drift process, the time of the boundary crossing and the selected response will vary between trials. The distance between the two boundaries (i.e. threshold) influences how much evidence must be accumulated until a response is executed. A lower threshold makes responding faster in general but increases the influence of noise on decision making while a higher threshold leads to more cautious responding. Reaction time, however, is not solely comprised of the decision making process perception, movement initiation and execution all take time and are summarized into one variable called non-decision time. The starting point of the drift process relative to the two boundaries can influence if one response has a prepotent bias. This pattern gives rise to the reaction time distributions of both choices (see figure 1.2; mathematical details can be found in the appendix).

Relationship to cognitive neuroscience

SSMs were originally developed from a pure information processing point of view and primarily used in psychology as a high-level approximation of the decision process. More recent efforts in cognitive neuroscience have simultaneously (i) validated core assumptions of the model by showing that neurons indeed integrate evidence probabilistically during decision making (Smith and Ratcliff, 2004; Gold and Shadlen, 2007) and (ii) applied this model to describe and understand neural correlates of cognitive processes (e.g. Forstmann et al., 2010a; Cavanagh et al., 2011).

Further, multiple routes to decision threshold modulation have been identified, thereby demonstrating the value of this modeling approach for managing problems of the context sensitivity of cognitive function, causal ambiguity, and the TIP. On the one hand, decision threshold in the speed-accuracy trade-off is modulated by

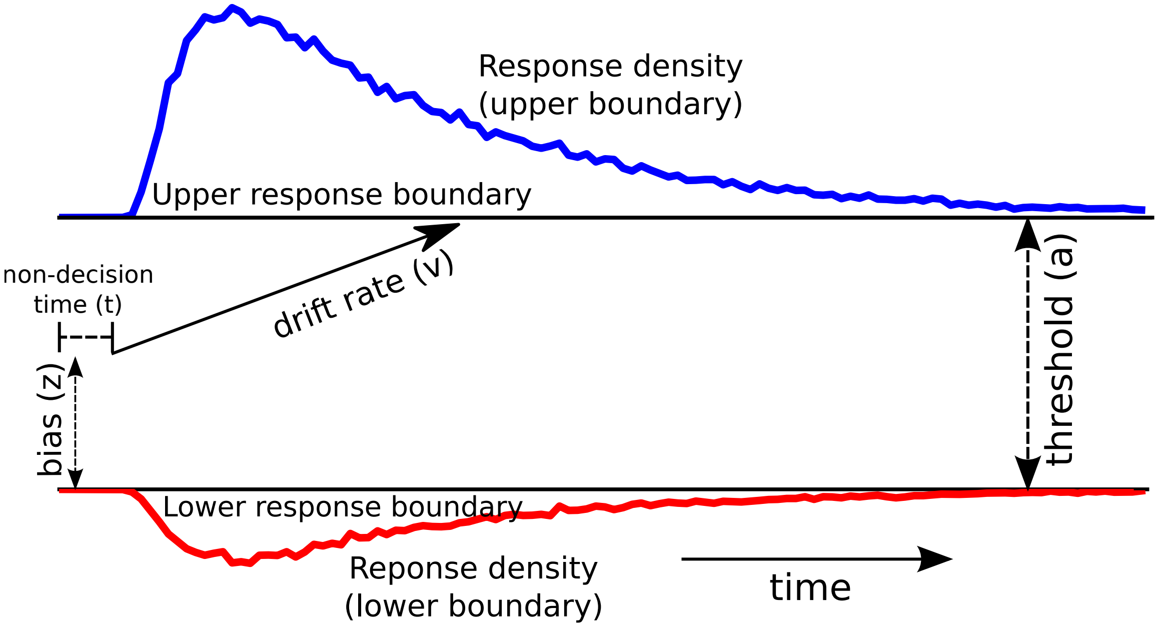

Figure 1.2: Trajectories of multiple drift-processes (blue and red lines, middle panel). Evidence is accumulated over time (x-axis) with drift-rate v until one of two boundaries (separated by threshold a) is crossed and a response is initiated. Upper (blue) and lower (red) panels contain histograms over boundary-crossing-times for two possible responses. The histogram shapes match closely to that observed in reaction time measurements of research participants.

changes in the functional connectivity between pre-SMA and striatum (Forstmann et al., 2010a). On the other hand, neural network modeling (Frank, 2006; Ratcliff and Frank, 2012) validated by studies of PD patients implanted with a deep-brain-stimulator (DBS) (Frank et al., 2007a) suggest that the subthalamic nucleus (STN) is implicated in raising the decision threshold when there is conflict between two options associated with similar rewards. This result was further corroborated by Cavanagh et al. (2011) who found that trial-to-trial variations in frontal theta power (as measured by electroencelophagraphy as a measure of response conflict (Cavanagh et al., 2012) is correlated with an increase in decision threshold during high conflict trials. As predicted, this relationship was reversed when STN function was disrupted by DBS in PD patients. When DBS stimulators were turned off, patients exhibited the same conflict-induced regulation of decision threshold as a function of cortical theta. Similarly, intraoperative recordings of STN field potentials and neuronal spiking showed that STN activity responds to conflict during decision making, and is predictive of more accurate but slower decisions, as expected due to threshold regulation (Zaghloul et al., 2012; Cavanagh et al., 2011; Zavala et al., 2013). Interestingly, these results provide a computational cognitive explanation for the clinical symptom of impulsivity observed in PD patients receiving DBS (Frank et al., 2007a; Hälbig et al., 2009; Bronstein et al., 2011).

Application to computational psychiatry

Despite its long history, the DDM has only recently been applied to the study of psychopathology. For example, threat/no-threat categorization tasks (e.g. Is this word threatening or not?) are used in anxiety research to explore biases to threat responses. Interestingly, participants with high anxiety are more likely to classify a word as threatening than low anxiety participants, although the explanation of this bias is unclear. One hypothesis assumes that this behavior results from an increased response bias towards threatening words in anxious people (Becker and Rinck, 2004; Manguno-Mire et al., 2005; Windmann and Krüger, 1998). Using DDM analysis, White et al. (2010b) showed that instead of a response bias (or a shifted starting-point in DDM terminology), anxious people actually showed a perceptual bias towards classifying threatening words as indicated by an increased DDM drift-rate.

In a recent review article, (White et al., 2010a) use this case-study to highlight the potential of the DDM to elucidate research into mental illness. Note that in this study the authors did not hypothesize about the underlying neural cause of this threat-bias. While there is some evidence that bias in decision making is correlated with activity in the parietal network (Forstmann et al., 2010b) this was not tested in respect to threatening words. Ultimately, we suggest that this research strategy should be applied to infer neural correlates of psychological DDM decision making parameters using functional methods like fMRI and employing modeling techniques at multiple levels of analysis (Frank et al., in press).

The DDM has also been successfully used to show that ADHD subjects were less able to raise their decision threshold when accuracy demands were high (Mulder et al., 2010b). Interestingly, the amount by which ADHD subjects failed to modulate their decision threshold correlated strongly with patients impulsivity/hyperactivity rating. Moreover, this correlation was specific to impulsivity and not inattentiveness. Note that in this case, the use of the DSM category (ADHD) may have obscured a more robust transdiagnostic association between decision threshold modulation and hyperactivity; and, “hyperactivity” itself may mask a variety of different causal processes.

A recent study by Pe et al. (2013a) showed that the DDM could also be used to explain previously conflicting reports on the influence of negative distractors on the emotional flanker task in depressed patients. Specifically, depression and rumination (a core symptom of depression) were associated with enhanced processing of negative information. These results further support the theory that depression is characterized by biased processing of negatively connotated information. Critically, this result could not be established by analyzing mean RT or accuracy alone, demonstrating the enhanced sensitivity to cognitive behavior of computational models.

In sum, SSMs show great promise as a tool for computational psychiatry. In helping to map out the complex interplay of cognitive processes and their neural correlates in mental illness, such models can play a role in resolving task impurity and other forms of causal ambiguity, identifying and measuring cognitive impairments, and associating such impairments with both symptoms and neural correlates. However, their applicability depends on the ability to accurately estimate them to construct individual computational, multi-dimensional profiles (CMPs). Such CMPs are parameter profiles that represent an individual's functioning as measured by the specific parameters making up the profile and derived from fitting the model to task performance data. In the next section, we will review different (L3) parameter estimation techniques, with a special focus on Bayesian methods that are usable for estimating parameters in the DDM and for generating individual CMPs. Finally, once SSMs can be fit accurately, we will identify (L4) clustering methods that can be used in a Bayesian framework to identify meaningful clusters of individuals, given their cognitive profiles (CMPs).

1.4.3 Level 3: Parameter estimation

To identify computational parameters in a variable clinical population with the DDM it is critical to have robust and sensitive estimation methods. In the following we describe traditional parameter estimation methods and their pitfalls. We then explain how Bayesian estimation provides a complete framework that avoids these pitfalls.

Random vs Fixed Parameters Across Groups of Subjects

Traditionally, fitting of computational models is treated as an optimization problem in which an objective function is minimized. Psychological experiments often test multiple subjects on the same behavioral task. Models are then either fit to individual subjects or to the aggregated group data. Both approaches are not ideal. When models are fit to individual subjects we neglect any similarity the parameters are likely to have. While we do not necessarily have to make use of this property to make useful inferences if we have lots of data, the ability to infer subject parameters based on the estimation of other subjects generally leads to more accurate parameter recovery (Wiecki et al., 2013a) in cases where little data is available as is often the case in clinical and neurocognitive experiments. One alternative is to aggregate all subject data into one meta-subject and estimate one set of parameters for the whole group. While useful in some settings, this approach is unsuited for the setting of computational psychiatry as individual differences play a huge role.

Hierarchical Bayesian models

Statistics and machine learning have developed efficient and versatile Bayesian methods to solve various inference problems (Poirier, 2006b). More recently, they have seen wider adoption in applied fields such as genetics (Stephens and Balding, 2009a) and psychology (e.g. Clemens et al., 2011a). One reason for this Bayesian revolution is the ability to quantify the certainty one has in a particular estimation. Moreover, hierarchical Bayesian models provide an elegant solution to the problem of estimating parameters of individual subjects outlined above (viz., the problem of neglecting similarities of parameters across subjects). Under the assumption that participants within each group are similar to each other, but not identical, a hierarchical model can be constructed where individual parameter estimates are constrained by group-level distributions (Nilsson et al., 2011a; Shiffrin et al., 2008a), and more so when group members are similar to each other.

Thus, hierarchical Bayesian estimation leverages similarity between individual subjects to share statistical power and increase sensitivity in parameter estimation. However, note that in our computational psychiatry application the homogeneity assumption that all subjects come from the same normal distribution is almost certainly violated (see above). For example, differences between subgroups of ADHD subjects would be decreased as the normality assumption pulls them closer together. To deal with the heterogeneous data often encountered in psychiatry we will discuss mixture models further down below. A detailed description of the mathematical details and inference methods of Bayesian statistics relevant for this endeavor can be found in the appendix.

1.4.4 Level 4: Supervised and unsupervised learning

Given that parameters have been estimated, or even given behavioral statistics alone, how can we group individuals into clusters that might be relevant for diagnostic categories or treatments? Bayesian clustering algorithms are particularly relevant to our objective as they (i) deal with the heterogeneity encountered in computational psychiatry and (ii) have the potential to bootstrap new classifications based on measurable, quantitative, computational endophenotypes. Because we are describing a toolbox using hierarchical Bayesian estimation techniques we focus this section on mixture models as they are easily integrated into this framework. Where possible, we highlight connections to more traditional clustering methods (e.g., “k-means”).

Gaussian Mixture Models

GMMs assume parameters to be distributed according to one of several Gaussian distributions (i.e. clusters). Specifically, given the number of clusters k, each cluster mean and variance gets estimated from the data. This type of model is capable of solving our above identified problem of assuming heterogeneous subjects to be normally distributed: by allowing individual subject parameters to be assigned to different clusters we allow estimation of different sub-groups in our patient and healthy population. Note, however, that the number k of how many clusters should be estimated must be specified a-priori in a GMM and remain fixed for the course of the estimation. This is problematic as we do not necessarily know how many sub-groups to expect in advance. Bayesian non-parametrics solve this issue by inferring the number of clusters from data. Dirichlet processes Gaussian mixture models (DPGMMs) belong to the class of Bayesian non-parametrics (Antoniak, 1974). They can be viewed as a variant of GMMs with the critical difference that they infer the number of clusters from the data (see Gershman and Blei (2012) for a review). An arguably simpler alternative, however, is to run multiple clusterings tested with different numbers of clusters and perform model comparison, as we discuss next.

Model Comparison

Model comparison provides measures to evaluate how well a model can explain the data while at the same time penalizing model complexity. Measures like the Bayesian Information Criterion (mathematical details can be found in the appendix) can be used to choose the GMM with the least number of clusters that still provide a good fit to the data. Moreover, model comparison is also used to select between computational cognitive models which often allow formulation of several plausible accounts of cognitive behavior. Of particular note are Bayes Factors that measure the evidence of a particular model in comparison to other, competing models (Kass and Raftery, 1993). More recently, and highly relevant to the field of computational psychiatry, these methods have been extended to provide proper random effects inference on model structure in heterogeneous populations (Stephan et al., 2009).

1.5 Example applications

In this last section we provide a proof of concept by demonstrating how the above described techniques (L1-L4) can be combined to (i) recover clusters associated with age, based on CMPs as extracted by the DDM, and (ii) predict brain state (DBS on/off).

1.5.1 Supervised and unsupervised learning of age

To demonstrate the concepts presented here-within we re-analyzed a data set collected and published by (Ratcliff et al., 2010). The data set consists of two groups, young (mean age 20.8) and old (mean age 68.6) human subjects tested on three different tasks: (i) a numerosity discrimination task that involved estimation of whether the number of asterisks presented on the screen was more or less than 50 (such that trials with close to 50 asterisks were harder than those with far fewer or far greater); (ii) a lexical decision task that required subjects to decide whether a presented string of letters is an existing word of the English language or not; and (iii) a memory recognition task that presented words to be remembered in a training phase that were subsequently tested for recall together with distractor words. Details of the tasks (including the conditions tested), subject characteristics, and DDM model analyses can be found in the original publication (Ratcliff et al., 2010).

We used the HDDM toolbox (Wiecki et al., 2013b) to perform hierarchical Bayesian estimation of DDM parameters from subjects RT and choice data without taking the different groups into account. We concatenated the DDM parameters of each subject in three tasks into one 22-dimensional CMP.

We next performed Factor Analysis (FA) on the CMP-vectors. FA is a statistical technique that uses correlations between parameters to find latent variables (called

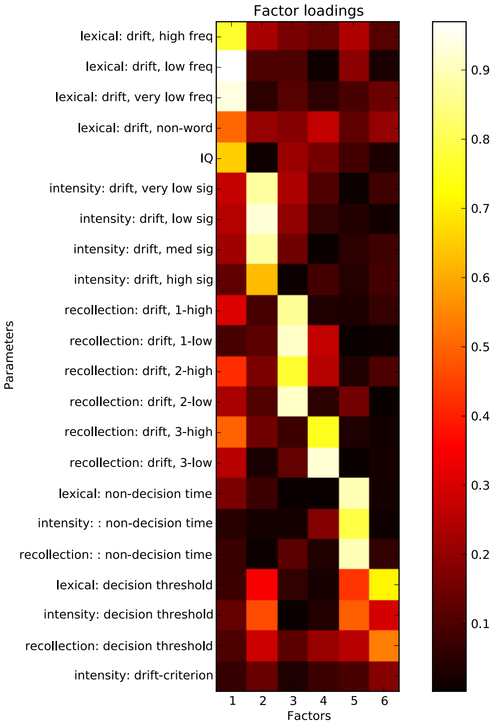

Figure 1.3: Factor loading matrix. Drift-diffusion model parameters of three tasks are presented along the y-axis while the extracted factors are distributed along the x-axis. Color-coded are the loading strengths. See the text for more details.

factors). Intuitively, highly correlated parameters will be loaded onto the same factor. As can be seen in figure 1.3, DDM parameters related to processing capability (i.e. drift-rate) in the three tasks are loaded onto the first four factors, while non-decision times and thresholds in the three tasks are loaded onto factor 5 and 6, respectively. Thus, instead of the 22 original dimensions we are able to describe the cognitive variables of individuals using 6 latent factors.

Classification of impairments and dysfunctions based on CMPs is a critical requirement for the clinical application of computational psychiatry. Although classification of age might not have clinical relevancy it provides an ideal testing environment as age is objectively measurable (as opposed to e.g. SZ, as described above). To classify young vs. old we employed logistic regression (using L2-regularization) on a

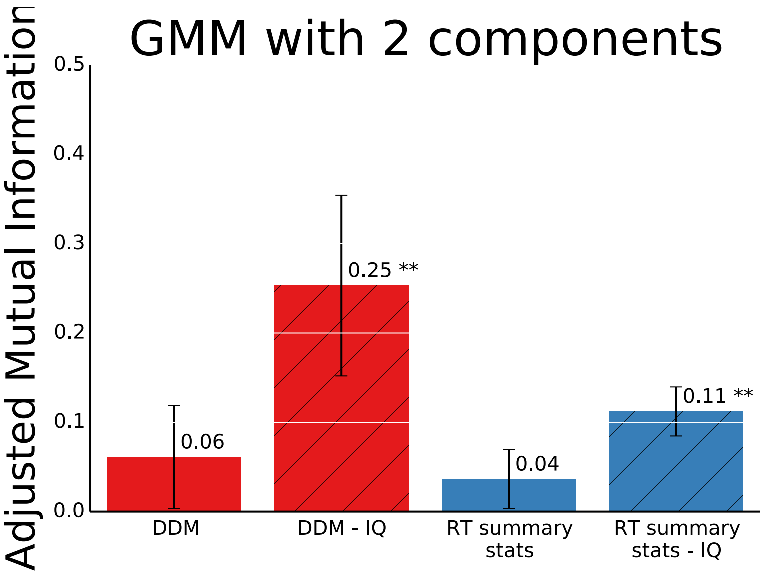

Figure 1.4: Adjusted mutual information scores (higher is better where 1 would mean perfect label recovery and 0 would mean chance level) for age after estimating a Gaussian Mixture Model with 2 components on DDM-factors (see text for more details on the factor analysis) and on DDM-factors after the contribution of IQ was regressed out. Error bars represent standard-deviation assessed via bootstrap. Asterisks ** denote significantly higher chance performance at p<0.01.

subset of the data and evaluated its prediction accuracy using held-out data (by using cross-validation). Classification performance was very high (up to 95% accuracy, not shown) demonstrating that cognitive tasks show great potential for classifying differences in brain functioning. In this case, there was no benefit to using DDM parameters compared to using summary statistics on RT and accuracy, as the differences in behavioral profiles between participants with large differences in age were quite stark. There are several examples where usage of a computational model does yield a significant increase in classification accuracy (see below and also Brodersen et al. (2013)) and may be more likely to do so when the patterns are more nuanced.

When applying these techniques to classify mental illness like SZ there is concern about the validity of our labels. If SZ does not represent a homogeneous, clearly defined group of individuals but rather patients with various cognitive and mental abnormalities, how could we expect a classifier to predict such an elusive, ill-defined

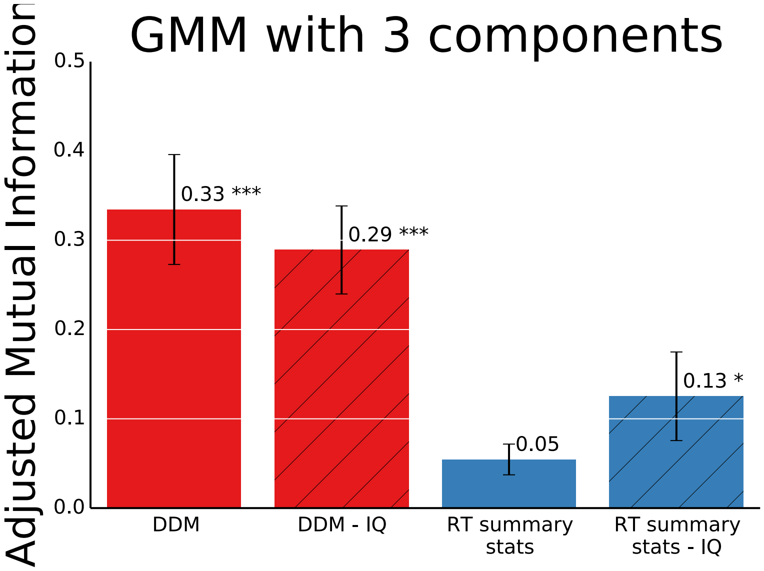

Figure 1.5: Adjusted mutual information scores (higher is better where 1 would mean perfect label recovery and 0 would mean chance level) for age after estimating a Gaussian Mixture Model with 3 components on DDM-factors (see text for more details on the factor analysis) and on DDM-factors after the contribution of IQ was regressed out. Error bars represent standard-deviation assessed via bootstrap. Asterisks * and *** denote significantly higher chance performance at p<0.05 and p<0.001, respectively.

label? One potential way to deal with this problem is to use an unsupervised clustering algorithm to find a new grouping which is hopefully more sensitive to the neurocognitive deficits Fair et al. (2012). As a proof of principle, we tested how well GMM clustering could recover age groupings in an unsupervised manner. Note that in a clinically more relevant setting we would not necessarily know the correct grouping ahead of time. Figure 1.4 shows the adjusted mutual information (which is 1 if we perfectly recover the original grouping and 0 if we group by chance) for age when estimating 2 clusters based on 6 latent factors extracted using FA (contrary to above we did not include IQ into the FA here). Notably, the age cluster is not recovered at all when using the DDM factors. Follow-up analysis suggests that the clustering selected by GMM picks up on some of the structure introduced by IQ (AMI = 0.1; not shown) . This indeed represents a potential problem for this unsupervised approach as there are many sources of individual variation like age, IQ, or education we might not be interested in when wanting clusters sensitive to pathological sources of variation. To address this problem we regressed the contribution of IQ out of every factor in order to remove this source of variation. Running GMM on these new regressed factors, we observe that the algorithm now clusters into different age groups (AMI=0.25 which corresponds to an accuracy of ˜75%). This might thus provide a viable technique in removing unwanted sources of inter-individual variation as variables like age, IQ, or education could just be regressed out before doing the clustering if these nuisance variables are known and measured.

The main issue here is that multiple factors can contribute to clusterings of neurocognitive parameters.

A different solution to this problem is presented in figure 1.5 where we estimated a GMM allowing for an additional cluster (3 clusters total). As can be seen, even when not regressing IQ out of the parameters, the clustering solution shows a clear sensitivity to age albeit none to IQ. Moreover, using summary statistics on RT and accuracy (mean and standard deviation) alone did not achieve a comparable level of recovery with the GMM (see figure 1.4 and 1.5). We also performed model comparison using BIC (not shown) to find the best number of clusters when successively testing different numbers of clusters. We found that adding more clusters monotonically decreased BIC thus favoring models with many clusters, despite the added complexity of these models. This might not be surprising given that there are many other individual differences beyond age and IQ that could affect group membership. It does represent a problem for this approach however as it is not immediately clear what level of representation should be chosen if a purely unsupervised measure like BIC does not provide guidance.

In conclusion, we demonstrated how computational modeling and latent variable models can be used to construct CMPs of individuals tested on multiple cognitive decision making tasks. Using supervised machine learning methods we were able to achieve up to 95% accuracy in classifying young vs. old age. Finally, after regressing IQ out as a nuisance variable, unsupervised clustering was able to group young and old individuals based on the structure of the CMP space.

1.5.2 Simulation experiment

Although the above example demonstrated a clear benefit in using the DDM for unsupervised clustering the model parameters were less beneficial compared to simple behavioral summary statistics (RT and accuracy) when performing supervised classification. This finding raises the question of whether DDM parameters derived based on behavioral measures alone can in principle provide a benefit in supervised learning over summary statistics. We thus performed a simple experiment where we simulated data from the DDM generating 2 groups with 40 subjects each. The mean parameters of the two groups differed in threshold, drift-rate and non-decision time (exact values can be found in the appendix). We then recovered DDM parameters by estimating the hierarchical HDDM (without allowing group to influence fit,

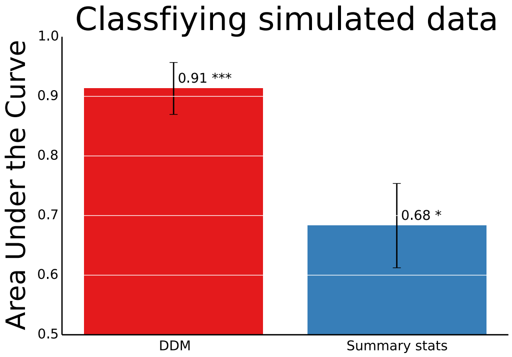

Figure 1.6: Area under the ROC curve which relates to classification accuracy of simulated RT data from the DDM. DDM represents parameters recovered in a hierarchical DDM fit ignoring the group labels. Summary statistics are mean and standard deviation of RT and accuracy. Error bars represent standard-deviation. Asterisks '***' and '*' indicate whether the accuracy is significantly higher than chance at p<0.001 and p<0.05, respectively.

which would be an unfair bias). Summary statistics consisted of mean and standard deviation of RT and accuracy. Figure 1.6 shows the area under the curve (AUC) using logistic regression with L2-regularization in a 10-fold cross-validation. As can be seen, for this parameter setting, the DDM-recovered parameters provide a large benefit over summary statistics. During the exploration of various generative parameter settings, however, we also found that other settings do not lead to an improvement, similar to the result obtained on the aging data set. Further research is necessary to establish conditions under which DDM modeling provides a clear benefit over using the simpler summary statistics.

1.5.3 Predicting brain state based on EEG

The above age example clearly demonstrated the potential of this approach in a data-driven, hypothesis-free manner. To complement this example we tested whether it was possible, using computational methods, to classify patients' brain state using computational parameters related to measures of impulsivity. We reanalyzed a data set from our lab in which Parkinsons Disease (PD) patients implanted with deep brain stimulators (DBS) in the subthalamic nucleus (STN) were tested on a reward-based decision making task (Cavanagh et al., 2011). STN-DBS is very effective in treating the motor symptoms of the disease but can sometimes cause cognitive deficits and impulsivity (Hälbig et al., 2009; Bronstein et al., 2011). Prior work has shown that when faced with conflict between different reward values during decision making, healthy participants and patients off DBS adaptively slow down to make a more considered choice, whereas STN-DBS induces fast impulsive actions. In this study, we showed that the degree of response time slowing for high conflict trials was related to the degree to which frontal theta power increased. DDM model fits revealed that theta-power increases were specifically related to an increase in decision threshold, leading to more cautious but accurate responding, whereas DBS prevented patients from increasing their threshold despite increases in cortical theta, leading to impulsive choice.

The above findings lend support to a computational hypothesis based on a variety of data across species regarding the neural mechanisms for decision threshold regulation. However, these findings were significant at the group level. Here, we tested whether we could classify individual patients' DBS status knowing only their DDM parameters, estimated from RT and choice data. We also included as a predictor the degree to which frontal theta modulated decision threshold (effectively another DDM parameter). Specifically, we used logistic regression with L2 regularization and cross-validation. The features for the classifier were the difference in thresholds in the two brain states (on and off DBS) and the difference in the theta-threshold regression coefficients in high and low conflict trials (on and off DBS). The classifier tries to predict which brain state a new subject is in based on these difference parameters without informing it which one corresponds to on or off state: we randomly sampled binary labels for each subject. The label indicated whether the features were coded relative to the on or off state. Intuitively, if the label was 0 for a subject, the features

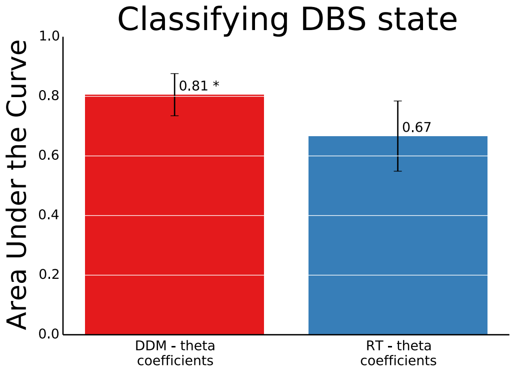

Figure 1.7: Out-of-sample classification accuracy using logistic regression to DBS state comparing DDM coefficients and using regression between RT and theta power. Error-bars indicate standard-deviation based on a bootstrap. The asterisk encodes significance at p<0.05.

would contain the change in regression coefficients (theta_diff_LC for low conflict and theta_diff_HC for high conflict) and threshold (a_dbs) when going from DBS on to off. Conversely, if the label was 1, the features would contain the change in regression coefficients and threshold when going from DBS off to on. The job of the classifier then becomes the classification of whether an individual is in the DBS on or off state based on the change in coefficients. The features based on raw RT data were created in a similar manner: Instead of using the regression coefficients of the influence of theta on decision threshold we included the influence of theta directly on RT in low and high conflict (found to be significantly correlated in (Cavanagh et al., 2011)) as well as the difference in mean RT between DBS on and off.

As can be seen in figure 1.7, using the DDM analysis greatly improved classification accuracy. Interestingly, of all the parameters fed into the classifier, the degree to which theta related to threshold adjustments in high-conflict trials was most predictive of DBS state (figure 1.8). This result is consistent with that obtained in (Cavanagh et al., 2011), but extends it to show how individual patients brain state, as a biomarker of impulsivity, can be diagnosed.

We thus demonstrated that this DDM analysis can be combined with brain measures

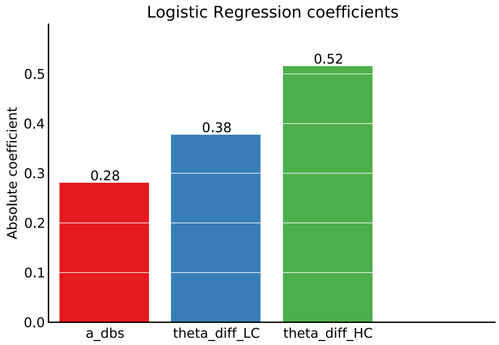

Figure 1.8: Absolute coefficients of logistic regression model using three predictors. Intuitively, the higher the coefficient, the more it contributes to separability of DBS state. a_dbs is the difference in threshold between DBS on and off, theta_diff_LC and theta_diff_HC are the differences in trial-by-trial regression coefficients between theta power (as measured via EEG) and decision threshold for low and high conflict trials, respectively.

(here EEG, but other measures such as fMRI are just as viable) to predict very specific changes in brain state. Critically, the influence of EEG on RT alone, although significant in Cavanagh et al. (2011), did not allow for the same accuracy as the DDM analysis. Moreover, this example shows the value of being hypothesis driven as this link between decision threshold and theta in high conflict trials (which was recovered as the most discriminative feature) was suggested by earlier, biologically plausible modeling efforts (Frank, 2005, 2006; Ratcliff and Frank, 2012; Wiecki & Frank, 2013).

While we show an increase in classification and clustering accuracy when using CMPs over aggregate performance scores (i.e. MPs, such as mean accuracy or mean RT) in certain cases, it could be argued that other feature extraction methods like Radial Basis Functions could achieve a similar goal. This is a compelling argument as these feature extraction methods are computationally much simpler and more flexible than task-specific computational models that parameterize the involved cognitive processes. Such an approach, if indeed comparable on the level of classification accuracy, might be favorable if classification was our sole goal. There are, however, unique benefits in describing behavior in terms of CMPs. Specifically:

- Computational models distill domain knowledge of the cognitive processes underlying task performance. As such, they can be seen as feature extraction methods that reduce nuisance variables, find a process-based representation of cognitive ability and thus make it easier for the classifier to separate different groups.
- Computational modeling can help with the task impurity problem. Aggregate performance scores summarize the contribution of a mixture of cognitive processes involved in a task. Computational models try to deconstruct behavior into its individual components and identify separable cognitive processes.
- Neurocognitive models often assume cognitive processes to be implemented by certain networks of the brain. As such, a computational parameter identified to have predictive power can be linked much easier to neural processes than aggregate performance scores.

1.6 Applications and challenges

How could this research program improve mental health research, diagnosis and treatment?

- Diagnosis: Ultimately the hope is that psychiatric diagnosis could move away from a symptom based classification of mental illness and instead use quantifiable biomarkers. CMPs could contribute to this by quantifying a subjects cognitive abilities in terms of psychological process variables that describe the efficacy of their neural circuitry.
- Treatment: Psychiatric drugs as well as other forms of treatment including deep-brain-stimulation have a high degree of variability in their efficacy across

individuals. By identifying pathological cascades and how they interact with treatment, we might be able to (I) predict which form of treatment will be effective for an individual and (ii) optimize treatment variables.

- Clinical research: Computational psychiatry can provide tools to link clinical symptoms to neurocognitive dysfunction that can open the door to a deeper level of understanding as well as provide novel targets for future studies into the causes of mental illness.
- Pharmacological research: Assessment of a drug mechanism and its efficacy by clinical ratings alone is often noisy, hard to interpret and biased due the placebo effect. More objective and quantitative measures of neurocognitive function are likely to improve on these current issues. Moreover, many psychiatric drugs fail in clinical phase 3 although they show promising results for a small subset of enrolled patients. If that subset could be identified by cognitive testing, the output of the drug discovery pipeline could be enhanced.

While the potential fruits of this research program are thus promising, the expected challenges to be overcome are nevertheless substantial. We cannot rely on DSM categories or a foundational understanding of the brain to bootstrap a new system in which to redefine mental illness. Among the main challenges are finding a good description of normal and abnormal cognitive function. Are there distinct clusters of cognitive dysfunction (and if so, how many) or is there a continuum with an arbitrary threshold on where mental illness begins? We provided an example here for how regressing out IQ can allow for better classification of age. Clearly in more complex psychiatric conditions, we may not always have access to variables that affect clustering of behavioral phenotypes in ways over which we would like to abstract.

While the new trans-dimensional approach of RDoC by the NIMH is very promising it must be open to additional levels of descriptions such as the neurocognitive computations of the brain. Computational psychiatry could then be embedded in this framework and translate neurocognitive research findings to other domains including genetics, neuroscience and clinical psychology.

1.7 Conclusions

In the light of the crisis in mental health research and practice and the widely recognized problems with conventional psychiatric classification based on the DSM, computational psychiatry is an emerging field that shows great promise for pursuing research aimed at understanding mental illness. Computational psychiatry provides powerful conceptual and methodological resources that enable management of the various features of mental illness and the various challenges with which researchers must cope. More specifically, by fitting computational models to behavioral data we can estimate computational parameters and construct computational multi-dimensional profiles (CMPs) which provide measures of functioning in one or another cognitive domain. Such measures are potentially of value in research contexts previously organized around symptom based classification as implemented by the DSM. CMPs may function as both more precise targets of research and more powerful explanatory resources for understanding individual differences, significant groupings, dynamic interactivity, and hierarchical organization of the brain.

Decision making appears to provide a good framework for studying mental illness as many disorders show abnormalities in core decision making processes. Strategically designed task batteries can provide the behavioral basis for studying such abnormalities. Sequential sampling models have a good track record in describing individual differences in decision making and can be linked to neuronal processes. Hierarchical Bayesian estimation provides a compelling toolbox to fit these models directly to data as it (i) provides an uncertainty measure; (ii) allows estimation of individual and group-level parameters simultaneously; (iii) allows for direct model comparison; and

(iv) enables deconstruction of symptoms by identifying latent clusters which correspond to different causal mechanisms. For example, impulsivity is a core symptom of many mental disorders like ADHD, OCD, Tourette syndrome, substance abuse and eating disorders (Robbins et al., 2012). Computational cognitive models have already started to deconstruct this broadly defined behavioral symptom and identified separate pathways that can all lead to alterations in impulse control (Dalley et al., 2011) including reduced motor inhibition (Chamberlain et al., 2006, 2008), early temporal discounting of future rewards, insensitivity towards negative relative to positive outcomes (Frank et al., 2007b; Cockburn and Holroyd, 2010), or an inability to adjust the decision threshold appropriately (Mulder et al., 2010a; Cavanagh et al., 2011; Frank et al., 2007a). Ultimately, the hope is to find novel ways to describe and assess mental illness based on objective computational neurocognitive parameters rather than the current subjective symptom-based approach. The bottom line is that computational psychiatry provides a combination of computational tools and strategies that are potentially powerful enough to underwrite a research program that will lead to a new level of understanding of mental illness and to new ways to describe, investigate, and assess mental illness, based on identifiable and reproducible neurocognitive computational multi-dimensional profiles.

1.8 Acknowledgements

The authors are grateful to Roger Ratcliff for generously providing the aging data set (Ratcliff et al., 2010) and useful discussions.

Supplementary online material can be found at http://ski.clps.brown.edu/papers/wieckietal_comp_p

Chapter 2

A computational model of inhibitory control in frontal cortex and basal ganglia

This chapter has been published and reflects contributions of other authors:

Wiecki T. V., & Frank M. J. (2013). A computational model of inhibitory control in frontal cortex and basal ganglia. *Psychological review*, 120(2), 32955.

2.1 Abstract

Planning and executing volitional actions in the face of conflicting habitual responses is a critical aspect of human behavior. At the core of the interplay between these two control systems lies an override mechanism that can suppress the habitual action selection process and allow executive control to take over. Here, we construct a neural circuit model informed by behavioral and electrophysiological data collected on various response inhibition paradigms. This model extends a well established model of action selection in the basal ganglia by including a frontal executive control network which integrates information about sensory input and task rules to facilitate well-informed decision making via the oculomotor system. Our simulations of the antisaccade, Simon and saccade-override task ensue in conflict between a prepotent and controlled response which causes the network to pause action selection via projections to the subthalamic nucleus. Our model reproduces key behavioral and electrophysiological patterns and their sensitivity to lesions and pharmacological manipulations. Finally, we show how this network can be extended to include the inferior frontal cortex to simulate key qualitative patterns of global response inhibition demands as required in the stop-signal task.

Download the model at: http://ski.clps.brown.edu/BG_Projects/

2.2 Introduction

"Before you act, listen. Before you react, think. Before you spend, earn. Before your criticize, wait." This quote by Ernest Hemingway highlights our basic tendency to act impulsively while reminding us that sometimes it is advisable to inhibit these prepotent response biases and act more thoughtful. Recent scientific advancements have shed light on the neural and cognitive mechanisms that implement inhibitory control of prepotent response biases (Andrs, 2003; Aron, 2007; Logan, 1985; Miyake et al., 2000; Stuphorn and Schall, 2006; Munoz and Everling, 2004). As part of this effort, a multitude of tasks exist to study response inhibition empirically. Among the tasks thought to require selective response inhibition are the antisaccade task, the Simon task, and the saccade-override task. Each of these tasks induces a prepotent response bias that sometimes needs to be overridden with a controlled response based on executive control. For example, the antisaccade task requires subjects to saccade in the opposite direction of an appearing stimulus. The Simon task requires subjects to respond according to an arbitrary stimulus-response rule (e.g., respond left or right depending on stimulus color), but where the stimulus is presented on one side of the screen, inducing a prepotent response bias to that side. In congruent trials the stimulus is presented on the same side as the correct response indicated by the rule, whereas on incongruent trials it is on the opposite side. Finally, the saccade-override task (Isoda and Hikosaka, 2007) requires subjects to saccade in the direction of a stimulus of a particular color for several repetitions in a row. On so-called switch-trials the instruction cue indicates that the other colored stimulus is now the target, so that the participant has to override the initial planned response and switch to the other one. While critical differences exist, all of these tasks require subjects to inhibit a prepotent response and replace it with a different response. In contrast, while also requiring response inhibition, the well-studied stop-signal task does not require subsequent initiation of an active response but only outright inhibition of the planned response (Verbruggen and Logan, 2008).

Electrophysiological and functional imaging data implicate key nodes in frontostriatal circuitry as being active during response inhibition and executive control. At the cortical level, these include the right inferior frontal gyrus (rIFG) (Aron et al., 2003; D. et al., 2007; Sakagami et al., 2001; Xue et al., 2008) the dorsolateral prefrontal cortex (DLPFC) (Wegener et al., 2008; Funahashi et al., 1993; Johnston and Everling, 2006), the supplementary eye fields (SEF) (Schlag-Rey et al., 1997), the presupplementory motor area (pre-SMA) (Congdon et al., 2009; Aron et al., 2007a; Isoda and Hikosaka, 2007), and the frontal eye fields (FEF) (Munoz and Everling, 2004). At the subcortical level, the striatum (Zandbelt and Vink, 2010; Watanabe and Munoz, 2011; Ford and Everling, 2009), the subthalamic nucleus (STN) (Eagle et al., 2008; Isoda and Hikosaka, 2008; Hikosaka and Isoda, 2008; Aron and Poldrack, 2006; Aron et al., 2007a) and the superior colliculus are involved. Manipulations that disrupt processing in either frontal or subcortical areas cause deficits in response inhibition (D. et al., 2007; Ray et al., 2009; Verbruggen et al., 2010). Moreover, response inhibition deficits are commonly observed in a wide range of psychiatric patients with frontostriatal dysregulation, including attentiondeficit/hyperactivity disorder (ADHD) (Nigg, 2001; Oosterlaan et al., 1998; Schachar and Logan, 1990), obsessive compulsive disorder (OCD) (Chamberlain et al., 2006; Menzies et al., 2007; Penadés et al., 2007; Morein-Zamir et al., 2009), schizophrenia (SZ) (Huddy et al., 2009; Bellgrove et al., 2006; Badcock et al., 2002), Parkinson's disease (PD) (van Koningsbruggen et al., 2009) and substance abuse disorders (Monterosso et al., 2005; Nigg et al., 2006).

Together, the above data suggest that intact functioning of the entire fronto-basal ganglia network is required to support response inhibition. However, it is far from clear that the underlying source of these deficits is the same. Inhibitory control is a very dynamic process, influenced by different interacting cognitive variables and neuromodulatory systems. Thus, response inhibition can be impacted by not only dysfunctional stopping *per se*, but can also be influenced by changes in motivational state (Leotti and Wager, 2010), attentional saliency (Morein-Zamir and Kingstone, 2006), maintenance and retrieval of task rules (Hutton and Ettinger, 2006; Nieuwenhuis et al., 2004; Reuter and Kathmann, 2004; Roberts et al., 1994), and separable modulations of selective vs global inhibition mechanisms (Aron, 2011), to name a few. Although electrophysiological recording studies demonstrate neuronal populations that differentiate between successful and unsuccessful stopping (Isoda and Hikosaka, 2008, 2007), or inhibition of prepotent responses in favor of controlled responses (Watanabe and Munoz, 2009; Ford and Everling, 2009), there is at present no coherent framework integrating all of these findings into a single model attempting to account for patterns of electrophysiological data, or selective disruptions of component parts and their effects on behavior.

The point of departure for our neural model builds on existing theorizing and data regarding the differential roles of the three main pathways linking frontal cortex with the basal ganglia (BG), often referred to as the direct, indirect and hyperdirect pathways. According to this framework, the corticostriatal direct “Go” and indirect “NoGo” pathways together implement a selective gating mechanism by computing the evidence for facilitating or suppressing each of the candidate motor actions identified by frontal cortex. Dopamine plays a critical role in this model by differentially modulating the activity levels in the two striatal populations, affecting both learning and choice. During rewards and punishments, phasic bursts and dips in dopamine neurons convey reward prediction errors (Montague et al., 1996) that transiently amplify Go or NoGo activity states, and therefore activity-dependent plasticity. In this manner, these striatal populations learn the positive and negative evidence for each cortical action (Frank, 2005). More chronic increases in tonic dopamine levels also directly affect choice by shifting the overall balance of activity toward the Go pathway over the NoGo pathway, thereby emphasizing learned positive relative to negative associations and speeding responding (and vice-versa for tonic decreases in dopamine). Many of this model’s predictions have been validated with behavioral studies involving dopaminergic manipulations and functional imaging in humans and monkeys (e.g., Frank et al., 2004; Nakamura and Hikosaka, 2006; Palminteri et al., 2009; Voon et al., 2010; Jocham et al., 2011), and synaptic plasticity and opto-genetic and genetic engineering studies in rodents (Kravitz et al., 2010; Hikida et al., 2010; Shen et al., 2008; Kravitz et al., 2012).

Note that in the above model, responses are *selectively* facilitated or suppressed via separate striatal Go and NoGo populations modulating the selection of particular cortical actions. However, more recent models have also incorporated the third hyperdirect pathway from frontal cortex to the STN to BG output. Communication along this pathway provides a *global* and dynamic regulation of the gating threshold, by transiently suppressing the gating of all responses when there is conflict between alternative actions (Frank, 2006; Ratcliff and Frank, 2012). Empirical studies using STN manipulations (Frank et al., 2007a; Wylie et al., 2010; Cavanagh et al., 2011) direct recordings (Cavanagh et al., 2011; Isoda and Hikosaka, 2008; Zaghloul et al., 2012), and fMRI/DTI (Aron et al., 2007a) have similarly supported this notion.

Nevertheless, the existing BG model cannot handle situations in which an initial prepotent response is activated but then needs to be suppressed – either altogether, or in favor of a more controlled response – situations typically studied under the rubric of "response inhibition". Here, we extend the model by incorporating additional cortical regions that facilitate executive control and can inhibit and override the more habitual response selection mechanism. We consider dynamics of the prepotent response process, the subsequent detection that this response needs to be inhibited, and the inhibition process itself – and how all of these factors are modulated by biological and cognitive variables. We consider electrophysiological data in various frontal (DLPFC, FEF, preSMA, ACC) and basal ganglia (striatum, STN) regions that are well captured by the model, and how these are linked to functional parameters of a high level decision making process embodied by a variant of the drift diffusion model.

Neural models are complex, in that they involve a number of parameters interacting to produce nonlinear effects on dynamics and behavior. There is also a risk of overfitting that could result from adjusting parameters to precisely match electrophysiological data from one experiment, which may make it difficult to precisely capture electrophysiological (or behavioral) data from a different experiment. Thus our aim was instead to capture qualitative patterns of data in both electrophysiology at multiple levels of cortical and subcortical network, and of the effects of their manipulation on behavior, with a single set of parameters.[1] In other work (Wiecki & Frank, in preparation) we show that systematic variations in neural model parameters are related in a lawful, monotonic fashion to more computational level parameters in a modified drift diffusion framework, providing a principled understanding and falsifiable experimental predictions. Moreover, despite the qualitative nature of model fits here, we nevertheless aim to distinguish our model from others in the literature based on general principles independent of particular parameterizations. Towards this goal we extracted a set of qualitative behavioral and neurocognitive benchmark results (listed in the results section) which we use to assess the validity of our model and compare to other models.

As noted above, despite surface features suggesting a single integrated response inhibition network, there are actually multiple dynamic components that can affect inhibition. Our contribution in this paper is to formalize these separable neural processes, to explore their interactive dynamics. To summarize and preview the core aspects of our work:

- We present a neural network model of the three main frontal-BG pathways supporting prepotent action selection, inhibitory control, conflict-induced slowing, and volitional action generation.
- We show that behavioral changes in a range of tasks dependent on these basic processes can result from alterations in brain connectivity and state and provide testable predictions for effects of distinct brain disorders.
- Selective response inhibition involves global conflict-induced slowing via the hy-

[1] By qualitative we mean that we do not attempt to quantitatively fit the precise shape of firing of any given cell type, but we do aim to show that a given population of cells increases or decreases firing rate at a particular point in time relative to some task event or to some estimated cognitive process. For example, for an area to be involved in inhibition it must show increased activity prior to the time it takes to inhibit a response. Or in striatum, particular cell populations are active related to biasing the prepotent response, suppressing that response, and then activating the controlled response - our model recapitulates this qualitative pattern.

perdirect pathway, raising the effective decision threshold to prevent prepotent responding, followed by DLPFC induction of striatal NoGo activity to inhibit the planned prepotent response. Subequently, the DLPFC provides top-down facilitation onto striatal Go populations encoding the controlled response.

- Response selection and inhibition are further regulated by neuromodulatory influences including dopamine linked to changes in motivational and attentional state. Dopamine reflects potential reward values and facilitates Go actions. In addition, our model suggests that while selective response inhibition is influenced by tonic levels of DA, global response inhibition is not.
- Our model is challenged in its ability to overcome prepotent responses and evaluated by its ability to reproduce key qualitative patterns reported in the literature, including:
  - Behavioral RT distribution patterns in selective response inhibition tasks.
  - Electrophysiological activity patterns of the FEF (Everling and Munoz, 2000), pre-SMA (Hikosaka and Isoda, 2008), the STN (Isoda and Hikosaka, 2008), striatum (Watanabe and Munoz, 2009), SC (Pouget et al., 2011; Pare and Hanes, 2003) and scalp recordings (Yeung et al., 2004a).
  - Psychiatric, developmental, lesion and pharmacological manipulations of frontal function and DA modulations.
- We show that when our model is extended to include the rIFG it can recover key electrophysiological and behavioral data from the stop-signal task literature.

In sum, this approach provides a mechanistic account of a major facet of cognitive control and executive functioning, which we hope will allow for a richer understanding of the relationship between behavioral, imaging, and patient findings.

2.3 Neural Network Model

We first introduce the neural circuit model of interacting dynamics among multiple frontal and basal ganglia nodes and their modulations by dopamine. We then describe how we vary model parameters to capture biological and cognitive manipulations.

Overview

The model is implemented in the Emergent software (Aisa et al., 2008) with the neuronal parameters adjusted to approximate known physiological properties of the different areas (Frank, 2005, 2006). The simulated neurons use a rate-code approximation of a leaky integrate-and-fire neuron (henceforth referred to as units) with specific channel conductances (excitatory, inhibitory and leak). Multiple units (simulated neurons) are grouped together into layers which correspond to distinct anatomical regions of the brain. Units within each layer project to those in downstream areas, and in some cases, when supported by the anatomy, there are bidirectional projections (e.g., bottom-up superior colliculus projection to cortex as well as top-down projections from cortex to colliculus). We summarize the general functionality of the model here to foster an intuitive understanding; implementational and mathematical details can be found in the appendix. While a single set of core parameters (i.e. integration dynamics and overall connection strength between layers) is used to simulate various electrophysiological and behavioral data in the intact state, each reported simulation is tested on 8 networks with randomly initialized weights between individual neurons. The model can be downloaded from our online-repository `http://ski.clps.brown.edu/BG_Projects` .

The model represents an extension of our established model of the BG (Frank, 2005, 2006; Wiecki and Frank, 2010). Because the extended model involves multiple components, we will progressively introduce each part, beginning with its core and then describing how each new component contributes additional functionality.

Basic basal ganglia model

The architecture of the core model is similar to Frank (2006). While the original model simulated manual motor responses, our model features a slightly adapted architecture in accordance to the neuroanatomy and physiology underlying rapid eye-movements (i.e. saccades) as reviewed in Hikosaka (2007) and Munoz and Everling (2004). Stimuli are presented to the network in the input layer, corresponding to high level sensory cortical representations. An arbitrary number of motor responses can be simulated, but here we include a model with just two candidate responses. The input layer projects directly to the cortical response units in the frontal eye fields (FEF) which implements action planning and monitoring and projects to the superior colliculus (SC), which acts as an output for saccade generation (Sparks, 2002). The SC consists of two units coding for a leftward and a rightward directed saccade. If the firing rate of one unit crosses a threshold, the corresponding saccade is initiated (Everling et al., 1999). The time it takes an SC unit to cross its threshold from trial onset is taken as the network's response time (RT). Stimulus-response mappings can be prepotently biased by changing projection strengths (i.e. weights) so that certain input patterns preferentially activate a set of FEF response units more than the alternative response units. (These sensory-motor cortical weights can also be learned from experience, such that they come to reflect the prior probability of selecting a particular response given the sensory stimulus; (Frank, 2006)). In fact, with only these three structures our model would only be capable of prepotent, inflexible responding.

By itself, FEF activation is not sufficiently strong to initiate saccade generation

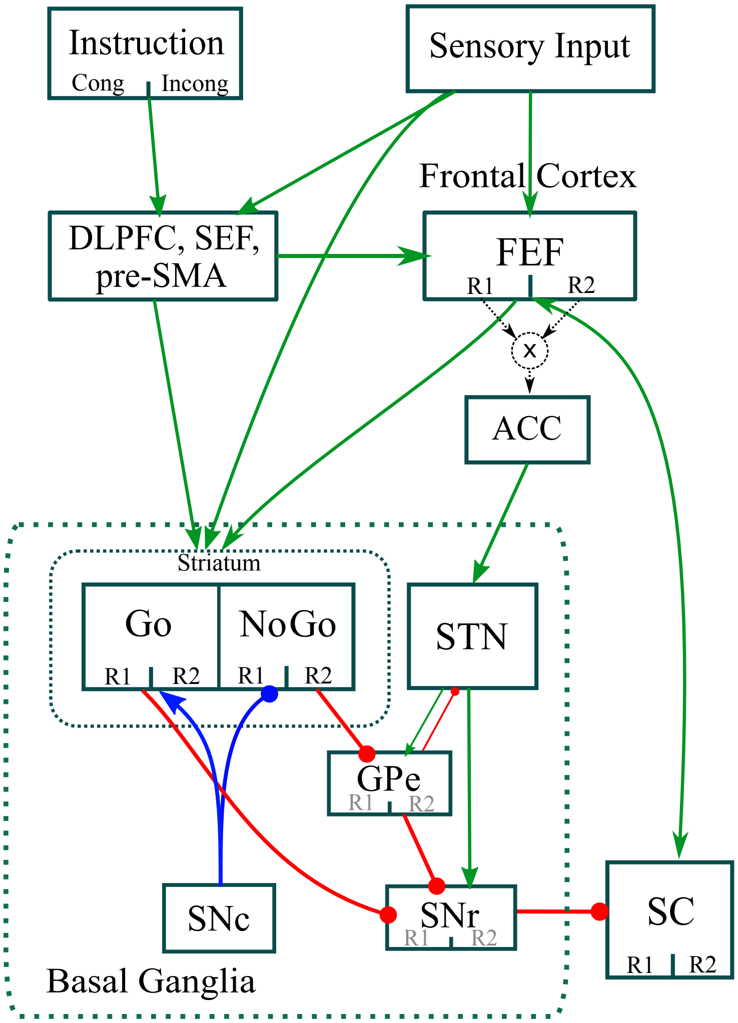

Figure 2.1: Box-and-arrow view of the neural network model. The sensory input layer projects to the FEF, striatum and executive control (i.e. DLPFC, SEF and pre-SMA). Via direct projections to FEF (i.e. cortico-cortical pathway), stimulus-response-mappings can become ingrained (habitualized). FEF has excitatory projections to the SC output layer that executes saccades once a threshold is crossed. However, under baseline conditions, SC is inhibited by tonically active SNr units. Thus, for SC units to become excited, they have to be disinhibited via striatal direct pathway Go unit activation and subsequent inhibition of corresponding SNr units. Conversely, responses can be selectively suppressed by striatal NoGo activity, via indirect inhibitory projections from striatum to GP and then to SNr. Coactivation of mutually incompatible FEF response units leads to dACC activity (conflict or entropy in choices), which activates STN. This STN surge makes it more difficult to gate a response until the conflict is resolved, via excitatory projections to SNr, effectively raising the gating threshold. Striatum is innervated by DA from SNc which amplifies Go relative to NoGo activity in proportion to reward value and allows the system to learn which actions to gate and which to suppress. The instruction layer represents abstract task rule cues (e.g. antisaccade trial). The DLPFC integrates the task cue together with the sensory input (i.e. stimulus location) to initiate a controlled re~~sponse corresponding to task rules, by activating the appropriate column of units in~~ FEF and striatum.

because the SC is under tonic inhibition from the BG output nucleus: the substantia nigra pars reticulata (SNr), whose neurons fire at high tonic rates. However, the tonic SNr-SC inhibition is removed following activation of corresponding direct (Go) pathway striatal units, which inhibit the SNr, and therefore disinhibit the SC (Hikosaka, 1989; Hikosaka et al., 2000; Goldberg et al., 2012). The indirect pathway acts in opposition to the direct pathway by further exciting the SNr (indirectly, via inhibitory projections to the globus pallidus (GP) which inhibits the SNr). Thus, direct pathway activity results in gating of a saccade (i.e. Go) while indirect pathway activity prevents gating (i.e. NoGo). Striking evidence for this classical model was recently presented by optogenetic stimulation selectively of direct or indirect pathways cells, showing inhibition or excitation of SNr respectively, and resulting in increased or decreased movement (Kravitz et al., 2010).

The Go and NoGo striatal populations include multiple units that code for the positive and negative evidence in favor of the FEF candidate actions given the sensory input context. Relative activity of the striatal pathways is modulated by dopaminergic innervation from the Substantia Nigra pars compacta (SNc) due to differential simulated D1 and D2 receptors present in the two pathways. In particular, dopamine further excites active Go units while inhibiting NoGo units. These effects on activity also produce changes in activity-dependent plasticity, allowing corticostriatal synaptic strength in the Go population to increase following phasic dopamine bursts during rewarding events, and those in the NoGo population to decrease (and vice-versa for negative events; (Frank, 2005)). For simplicity, in the present model we omit learning because the paradigms we simulate do not involve learning, and focus on associations that have already been learned. However, it is now well known that striatal unit activity is modulated by the reward value of the candidate action, such that rewarding saccades are more likely to be disinhibited (Hikosaka et al., 2006).

Bottom-up projections from SC to FEF allow action-planning to be modulated according to direct and indirect pathway activity (Sommer and Wurtz, 2006, 2004a,b, 2002). This effectively forms a closed loop in which FEF modulates the striatum which, via gating through SNr and SC, in turn modulates the FEF. Loosely, FEF considers the candidate responses and ”asks” the BG if the corresponding action should be gated or not. Thus, with these structures the model can selectively gate responses modulated by DA.

In addition to the above gating dynamics, the overall threshold for gating is controlled by the ease with which the SNr units are inhibited by the striatal Go units. The STN sends diffuse excitatory projections to the SNr (Parent and Hazrati, 1995), and therefore when STN units are active they increase the gating threshold for all responses, effectively contributing a ’global NoGo’ signal (Frank, 2006; Ratcliff and Frank, 2012). The STN does not however, act as a static increase in threshold. Rather, the STN receives input directly from frontal cortex, and becomes more active when there is response conflict (or choice entropy) during the early response selection process. In the current model, conflict is computed explicitly by the dorsal anterior cingulate cortex (dACC), which detects when multiple competing FEF response units are activated concurrently, and in turn activates the STN to make it more difficult to gate any response until this conflict is resolved. The full computational role of dACC is far from resolved and likely to be more complex than conflict detection and control (see, e.g. Holroyd and Coles, 2002; Botvinick et al., 2004; Alexander and Brown, 2011; Kolling et al., 2012). Nevertheless, alternative accounts of dACC function (Kolling et al., 2012) are entirely compatible with our model (an issue we return to in the discussion), but for convenience we label the computation as “conflict”.

Frontal Pathway model

Volitional response selection

So far our model is able to select/gate responses and slow down gating when an alternative response appears to have some value relative to the initial planned action.

However, SRITs require executive control: integration of the sensory state together with the task rule to not only inhibit the prepotent response but replace it with a volitional one. Such rule-based processing is effortful and time-consuming, and hence the controlled response process lags that of the initial fast response capture. Based on a variety of evidence, we ascribe the rule-based representations to the dorso-lateral prefrontal cortex (DLPFC) (e.g. Miller and Cohen, 2001; Chambers et al., 2009). This structure is involvedin the active maintenance of stimulus-response rule representations (Derrfuss et al., 2004, 2005; Brass et al., 2005), is necessary for correct antisaccade trials (Wegener et al., 2008; Funahashi et al., 1993; Johnston and Everling, 2006), and is involved in selective response inhibition (Garavan et al., 2006; Simmonds et al., 2008) and response selection (Braver et al., 2001; Rowe et al., 2002). Moreover, SEF (Schlag-Rey et al., 1997) and pre-SMA (Isoda and Hikosaka, 2007; Ridderinkhof et al., 2011) are also critically involved in correct SRIT performance.

We consequently added an abstract executive control layer to summarize the DLPFC, SEF and pre-SMA complex (in the future referred to as DLPFC). This layer selects FEF responses and biases BG gating according to task rules (see figure 2.1). Although not explicitly represented separately in the model architecture, we conceptualize the individual contribution of DLPFC as rule encoding and abstract action selection whereas SEF and pre-SMA are transforming this abstract action representation into concrete motor-actions (Schlag-Rey and Schlag, 1984; Schlag

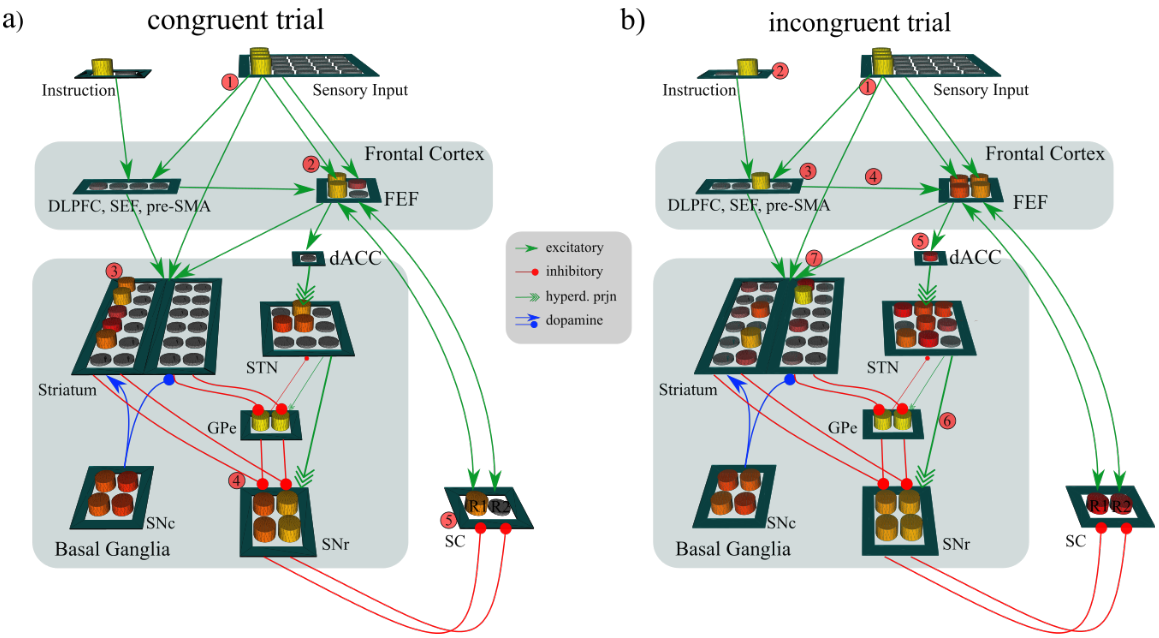

Figure 2.2: Neural network model in different task conditions. **a)** Prosaccade condition. (1) Left stimulus is presented in input layer; (2) Prepotent weights bias left response coding units in FEF; (3) Left response Go gating neurons in striatum are activated; (4) Left response coding units in SNr are inhibited; (5) The left response unit in SC is disinhibited, and due to recurrent excitatory projections with FEF, is excited and the action is executed. **b)** Antisaccade condition. The activity pattern early in the trial (i.e. before DLPFC comes online) is similar to that in the prosaccade condition. (1) Left stimulus is presented in input layer activating prepotent left response in FEF; (2) The unit coding for the antisaccade condition is externally activated in instruction layer; (3) DLPFC integrates sensory and instruction input according to task rules and activates *right* coding units in FEF together with *right* Go gating units *left* NoGo units in striatum; (4) in FEF, right coding units are activated due to DLPFC input in addition to the prepotent left coding units already active; (5) dACC detects co-activation of multiple FEF action plans and activates (6) hyperdirect pathway to excite STN and SNr, globally preventing gating until conflict is resolved. Eventually, stronger controlled DLPFC activation of the right coding FEF response results in gating of the correct antisaccade (7). In some trials, DLPFC activation is too late and the prepotent left saccade will have already crossed threshold, resulting in an error.

and Schlag-Rey, 1987; Curtis and DEsposito, 2003). In turn, these planned motor actions can influence the selected response in FEF and bias gating via projections to striatal Go and NoGo neurons (Munoz and Everling, 2004).

Anatomical and functional studies demonstrate projections from both DLPFC to SEF and pre-SMA (Lu et al., 1994; Wang et al., 2005) and to striatum to affect response gating (Haber, 2003; Doll et al., 2009; Frank and Badre, 2011); and from SEF to FEF (Huerta et al., 1987). We explore how these projections impact dynamics of response selection. But how does the executive controller in our model 'know' which rule to activate? We do not address here how these rule representations arise via learning, which is the focus of other PFC-BG modeling studies (see Rougier et al., 2005; Frank and Badre, 2011; Collins and Frank, 2012). Instead, we simulate the state of the network after learning by simply including an Instruction layer as a second input layer to the model encoding task condition (e.g. antisaccade trial). In case of the antisaccade task, the sensory input layer encodes the direction of the visual stimulus and the instruction layer encodes whether the network should perform a pro or antisaccade. The DLPFC complex then integrates these two inputs and activates a (pre-specified) rule unit that (i) projects to the correct FEF response units supporting the antisaccade; (ii) activates striatal NoGo units to prevent gating of the active prepotent pro-saccade response, and (iii) activates striatal Go units encoding the controlled antisaccade.

Critically, DLPFC units are relatively slow to activate the appropriate rule unit. This is due to the need to formulate a conjunctive rule representation between the visual location of the stimulus and the task instruction (either one of these is not sufficient to determine the correct response, and indeed, each individual input provides evidence for multiple potential rules). Time constants of membrane potential updating is reduced to support this integration, which also is intended to approximate slower time course of rule retrieval and subsequent computation to determine the correct action (via interactions with preSMA and SEF). Moreover, we include considerable inter-trial noise in DLPFC activation dynamics so that executive control is available earlier on some trials while later on others. The slowed integration and the increase of inter-trial noise in executive control are necessary for the model to capture the quantitative benchmark results (demonstrated below). Moreover, the slower controlled processing is also a core feature of classical dual process models of cognition (e.g. Sloman, 1996) and the increased noise accords with the general statistical observation that longer latencies are typically associated with greater variability.

Competition between the two response selection mechanisms

As outlined above, our model features two response selection mechanisms: (i) a fast, prepotent mechanism driven by a biased projection from sensory input to FEF; and (ii) a slow, volitional mechanism that originates in the DLPFC which integrates instruction and sensory input to select and gate the correct response. Importantly, the volitional mechanism is slower but stronger than the prepotent one. If, due to noise in the speed of integration, executive control is slower on some trials, it might be too late to activate the correct rule representation before the prepotent response is gated. In contrast, when the executive controller is faster, it activates the alternative FEF response, leading to conflict-induced slowing, and then actively suppresses the prepotent response via projections to striatal NoGo units encoding the prosaccade. This conceptualization can be regarded as a biologically plausible implementation of the cognitive activation-suppression model (Ridderinkhof, 2002; Ridderinkhof et al., 2004). Note however that our implementation involves two suppression mechanisms, one in which conflict results in global threshold adjustment, and another in which the prepotent response is selectively inhibited.

Modulations

To test the influence of different biological manipulations on executive control paradigms, we modify various parameters in the network model. Here, we list the different modulations and their implementation.

- *Prepotency*: To simulate differences in the strength of the prepotent response capture of an appearing stimulus (e.g., the prosaccade stimulus) we modulate the projection strength between sensory input to the dominant response units in FEF and striatum.

- *Speed of DLPFC*: To simulate efficacy of prefrontal function we modulate the speed of DLPFC integration, by adjusting the time constant of membrane potential updating in these units. Faster updating implies proactive control.

- *Connectivity of DLPFC*: To simulate differences in intra-cortical connectivity we modulate the DLPFC→FEF projection strength.

- *Speed-accuracy trade-off*: To simulate strategic adjustments in the speed-accuracy trade-off, we modulate the connection strength between frontal cortex and striatum (Forstmann et al., 2010a). In particular, when speed is emphasized, the FEF more effectively activates striatal Go units so that it is easier to reach gating threshold. In contrast, accuracy adjustments are reflected in increased STN baseline ultimately increasing the response gating threshold.

- *STN impact*: STN contributions are simulated by manipulating the relative synaptic strengths from STN to SNr, effectively changing the amount of STN activity required to prevent BG gating (Ratcliff and Frank, 2012; Cavanagh et al., 2011).

- *tonic DA*: Pharmacological and disease modulations of DA levels are simulated by either decreasing (e.g., PD) or increasing (e.g., SZ) tonic DA activity, which in turn modulates relative activity of Go vs NoGo units.

2.3.1 Selective Response Inhibition

Methods

As summarized earlier, all SRITs have a common task structure. (i) A prepotent response bias is induced by priming an action. In the antisaccade task this is a result of the appearance of a stimulus that initiates a ’visual grasping reflex’ (Hess et al., 1946); in the Simon task this is the result of placing the target stimuli on either side of the screen, initiating a response capture (Ridderinkhof, 2002); in the saccade-overriding task this is the result of repeated responding to the same colored stimulus which renders this response habitual. (ii) In congruent trials, the correct response is the same as the prepotently biased one. (iii) In incongruent trials, the correct response is incompatible with the prepotently biased response, and subjects can use executive control to suppress the initially predominant action in favor of the task-appropriate one.

We implemented this common task structure as follows in our neural network model (alternative task implementations that accommodate the differences between the tasks lead to similar patterns so we simplified in order to use a single task representation of this basic process, but nevertheless simulate patterns of data evident in specific tasks below). Two stimulus positions, left and right, were encoded in the input layer as two distinct columns of activated units. The prepotent bias toward an appearing target was hard-coded by strong weights from each input stimulus to corresponding response units in FEF. This prepotent weight facilitates fast responding for congruent trials, but biases responding in the erroneous direction for incongruent trials. The DLPFC layer integrates sensory input and instruction input to activate a conjunctive rule unit encoding the unique combination of sensory and instruction input, which then projects to the associated correct response unit in FEF. Each of the four DLPFC units project to the appropriate FEF response unit. Note that weights from the DLPFC to the FEF are stronger than the prepotent bias connection from the input layer to the FEF so that the DLPFC would eventually override an erroneous prepotent response. (The same functionality could be achieved by simply allowing DLPFC units to reach a higher firing rate or to engage a larger population of units, instead of adjusting the weights). In addition, DLPFC units activate corresponding Go and NoGo units in the striatum (e.g. in an antisaccade trial, Go units coding for the correct response and NoGo units coding for the incorrect response get activated by top-down PFC input).

Results

We identified a set of key behavioral and neurophysiological qualitative patterns across SRITs that form desiderata for our model to capture:

#1 Incongruent trials are associated with higher error rates than congruent trials (e.g Reilly et al., 2006; McDowell et al., 2002; Isoda and Hikosaka, 2008).

#2 Reaction times (RTs) are faster for errors than correct trials (e.g Reilly et al., 2006; McDowell et al., 2002; Isoda and Hikosaka, 2008).

#3 Strategic adjustments in the speed-accuracy trade-off (via changes in decision threshold) modulates functional connection strength between frontal cortex and striatum (Forstmann et al., 2010a). Similarly, STN activity is associated with modulations of the decision threshold (Ratcliff and Frank, 2012; Cavanagh et al., 2011).

#4 Various psychiatric diseases associated with frontostriatal cathecholamine dysregulation lead to increased error rates and speeded responses (e.g. Reilly et al., 2006; Harris et al., 2006; Reilly et al., 2007; McDowell et al., 2002).

#5 Early activation of prepotent motor response, e.g. in EMG measurements (Burle et al., 2002).

#6 At least four different types of activation dynamics in FEF neurons during correct and error incongruent trials (Everling and Munoz, 2000). Specifically, neurons coding for the erroneous (i.e. prepotent) response are fast to activate and their activity is greater on error trials than correct trials. In contrast, neurons coding for the correct (i.e. controlled) response are slower to activate and their activity is reduced and delayed on error trials. See figure 2.6c for the quantitative data that forms the basis of this qualitative pattern.

#7 At least four different types of striatal neurons with dissociable dynamics and direction selectivity in congruent and incongruent trials (Watanabe and Munoz, 2009; Ford and Everling, 2009). Specifically, (i) during prosaccades, distinct neural populations code for facilitation of the correct response and suppression of the alternative; (ii) during antisaccade trials, (iia) neurons coding for facilitation of the *incorrect* prepotent response initially become active but return to baseline when (iib) neurons coding for the suppression of that response become active together with (iic) neurons coding for facilitation of the correct controlled response (see figure 2.9b).

#8 Neurons forming part of the hyperdirect pathway from frontal cortex (pre-SMA, dACC) to the STN show increased activity (i) *before* correct incongruent responses and (ii) *after* incorrect incongruent responses, but (iii) baseline activity during congruent response (Isoda and Hikosaka, 2007, 2008; Yeung et al., 2004a; Zaghloul et al., 2012). This pattern of activity co-occurs with delayed but more accurate incongruent responding.

In the following, we demonstrate how our model reproduces these qualitative patterns, before linking its dynamics to a higher level computational description.

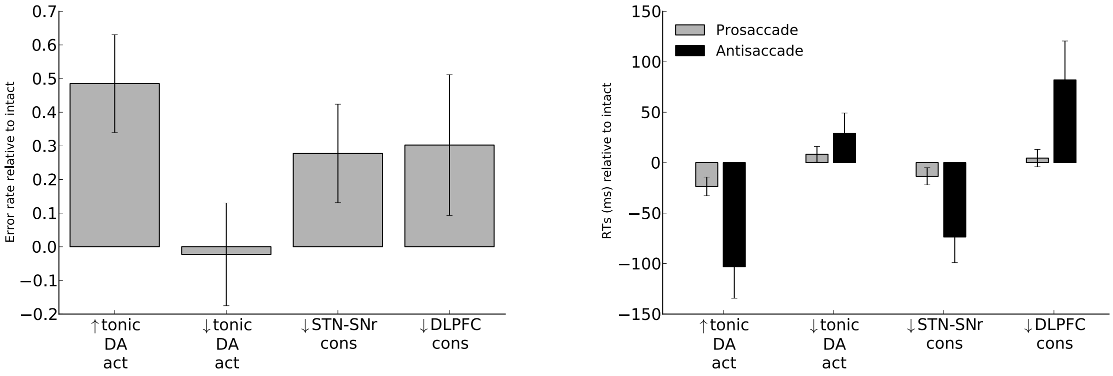

Figure 2.3: **a)** Error rates in incongruent trials ± SEM relative to intact networks for different neural manipulations. Networks make more errors with increased tonic DA levels, or STN dysfunction, compared to intact networks. **b)** Response Times (RTs) ± SEM relative to intact networks, for pro and antisaccade trials as a function of neural manipulations. For more analysis see the main text.

Behavior

As expected, intact networks make considerably more errors on incongruent trials (error rate of 15%) as compared to perfect performance in congruent trials (error rate close to 0%, not shown), thereby capturing qualitative pattern #1.

Further, networks in general have longer response times (RTs) in incongruent trials (see figure 2.3(b)) thus capturing qualitative pattern #2. Incongruent trials are slower for two reasons: (i) it takes time for executive control (DLPFC) computations due to the requirement to integrate two sources of input to activate the associated rule; and (ii) once activated, the controlled response conflicts with the prepotent response, leading to STN activation and associated increases in BG gating threshold.

Additional analysis revealed that incongruent error trials are associated with faster RTs compared to correctly performed incongruent trials (figure 2.4). In our model, errors are made when the faster prepotent action reaches threshold before the inhibitory process can cancel it. This mechanism allows the model to capture qualitative pattern #2 and #3.

We next investigated how these behavioral patterns were affected by manipulations (see figure 2.3(a)). Incongruent error rates were most exaggerated with increased tonic DA levels, and by disrupted STN function to simulate deep brain stimulation. The effect of increased striatal DA on incongruent error rates captures corresponding patterns (see #4) observed in non-medicated schizophrenia patients, who have elevated striatal DA (e.g Reilly et al., 2006; Harris et al., 2006; Reilly et al., 2007; McDowell et al., 2002). Tonic DA elevations are associated with speeded responding in both congruent and incongruent trials, due to shifted balance toward the Go pathway facilitating response gating. This same mechanism explains the increased antisaccade error rate. Conversely, decreased tonic DA leads to slowed responding due to increased excitability of the indirect NoGo pathway. The model also predicts that STN dysfunction produces increased error rates, due to an inability to raise the threshold required for striatal facilitation of prepotent responses. Indeed, STN-DBS induces impulsive (fast but inaccurate) responding in SRITs (Wylie et al., 2010).

Finally, we tested in more detail how systematic parametric changes in a biological variable affect RT and accuracy. Figure 2.5(a) shows how RT distributions change under different settings of FEF→striatum connection strength. Figure 2.5(b) shows quantitatively how increases in FEF→striatum connectivity leads to faster RT and decreased accuracy (qualitative pattern #3). Loosely, increasing FEF connection strength onto Go-units in the direct pathway leads to faster gating of responses. Conversely, increases in STN→SNr connectivity lead to slower RT and improved accuracy (figure 2.5(c)). The reason for both of these effects is that they differentially modulate SNr activity. Recall that the SNr tonically inhibits the thalamus, unless it is itself inhibited by the striatal direct pathway. Hence any modulation of the ease with which SNr units are inhibited – either via stronger connections from cortex onto Go units, or by increasing the SNr via the STN – will change the threshold required for

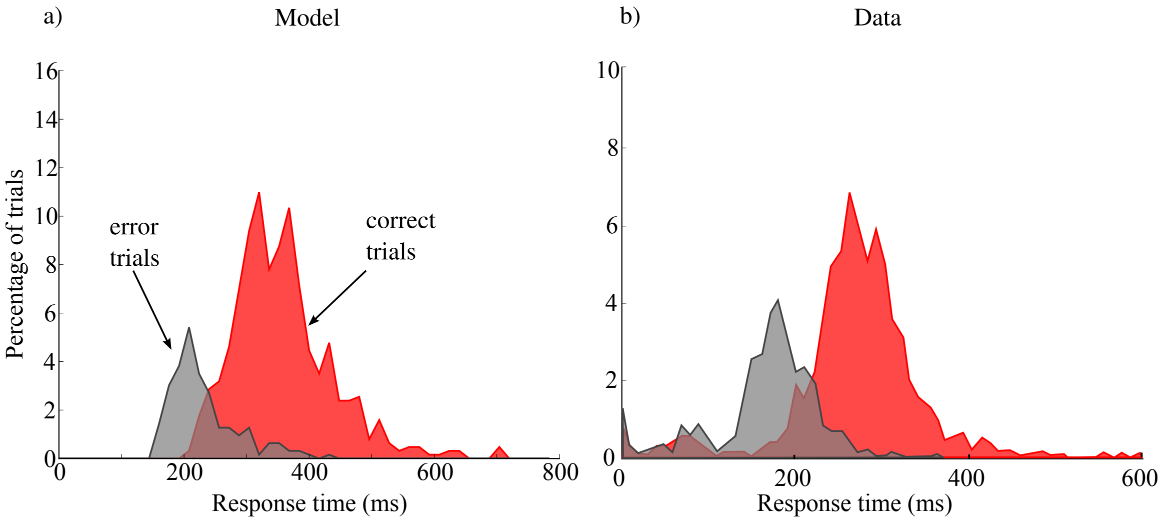

Figure 2.4: **a)** RT histogram for correct and erroneous incongruent trials in the model. Error RT distributions were shifted to the left due to prepotent response capture. This pattern is exaggerated with increased tonic DA due to lowered effective gating threshold. **b)** RT histograms of a monkey during the switch-task (reproduced from Isoda and Hikosaka (2008)). In blocks of trials, monkeys are continuously rewarded following saccades to one of two targets. On so-called 'switch-trials' a cue indicates that the monkey should perform a saccade to the opposite target, requiring the monkey to inhibit his planned saccade and perform a saccade to the opposite direction. As in the model, errors are associated with shorter reaction time. **c)** Reaction time distribution of an alternative model with fast DLPFC integration speeds. Correct trials are in red and errors in gray (not present). This model cannot account for the behavioral pattern of errors and RTs as a function of congruency, in contrast to models with slowed DLPFC integration (panel a).

the BG to gate an action. Indeed, Ratcliff and Frank (2012) and Lo and Wang (2006) have shown that these two mechanisms are related to changes in the decision threshold in sequential sampling models. Our model subsumes both of these mechanisms, and suggests that these different routes are themselves modulated by distinct cognitive variables, such as volitional speed-accuracy modulation and conflict/choice entropy (cortico-striatal and STN). We return to this issue in the Discussion.

In sum, our model captures key qualitative behavioral patterns described in the literature (see above). Moreover, these patterns hold over varying biologically plausible parameter ranges leading to predictable changes in the behavioral patterns. However, given the complexity of the underlying model, it is also important to establish whether the internal dynamics of the different nodes of the network are consistent with available electrophysiological data in this class of tasks.

Neurophysiology

DLPFC, SEF and pre-SMA activity Our model summarizes the computations of the executive control complex as a single layer corresponding to DLPFC, SEF and pre-SMA. One of our central predictions is that DLPFC activation must be delayed relative to the habitual response mechanism in order to produce the desired qualitative patterns. To demonstrate the plausibility of this account we simulated networks with increased DLPFC speed (time constant of membrane potential updating). Consequently, networks ceased to make fast errors while correct RTs became much faster and more peaked (figure 2.4c). The reason for this pattern is that active executive control now dominates and overrides the prepotent mechanism during early processing. This result implies that some delay in executive control is needed to account for empirical findings in which incongruent RTs are delayed.

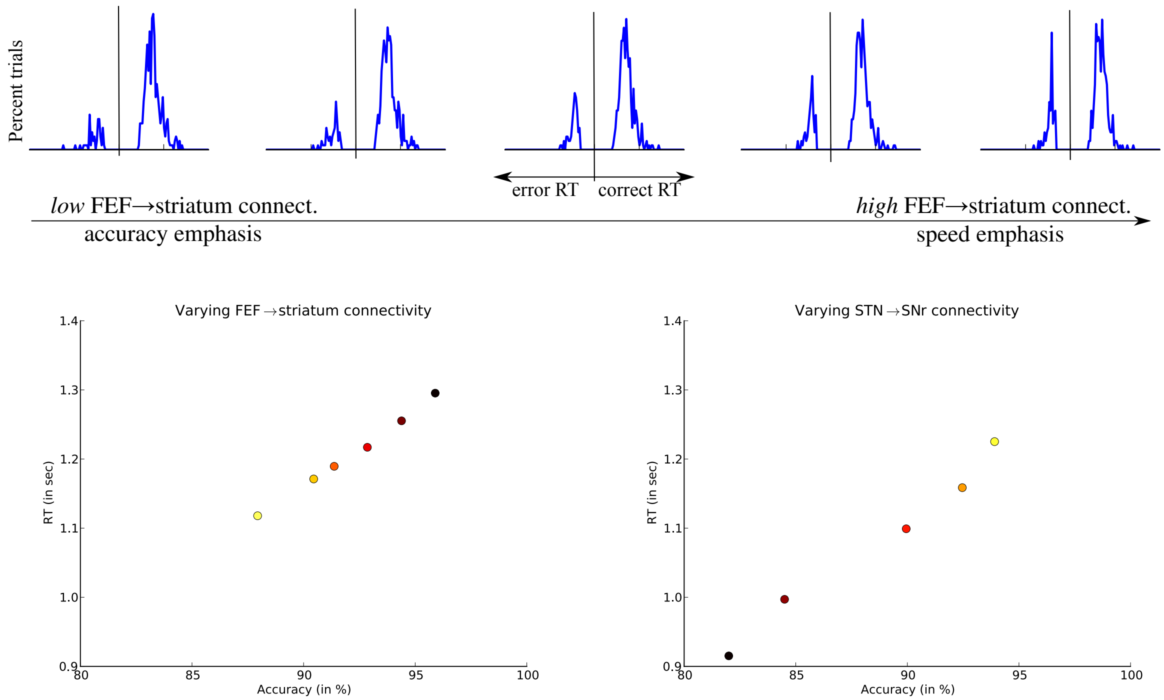

Figure 2.5: **a)** RT distributions for incongruent trials by network models. FEF→striatum projection strengths were varied along the x-axis. Correct RT distributions are on the right side of each panel and incorrect RT distributions are on the left side, mirrored on the y-axis. This manipulation is equivalent to a speed-accuracy adjustment, as shown empirically to vary with pre-SMA→striatal communication (Forstmann et al, 2008; 2010), where here FEF plays the role of pre-SMA for eye movements as compared to manual movements studied in Forstmann et al. **b+c)** Speed-accuracy tradeoff under parametric modulation of **(b)** FEF→striatum connection strength and **(c)** STN→SNr connection strength (color coded). Black represents low and yellow high connection strength. This pattern is consistent with decision threshold modulation. The absolute values of connection strengths in these different routes are chosen to lie on a sensitive range producing observable effects for demonstration purposes.

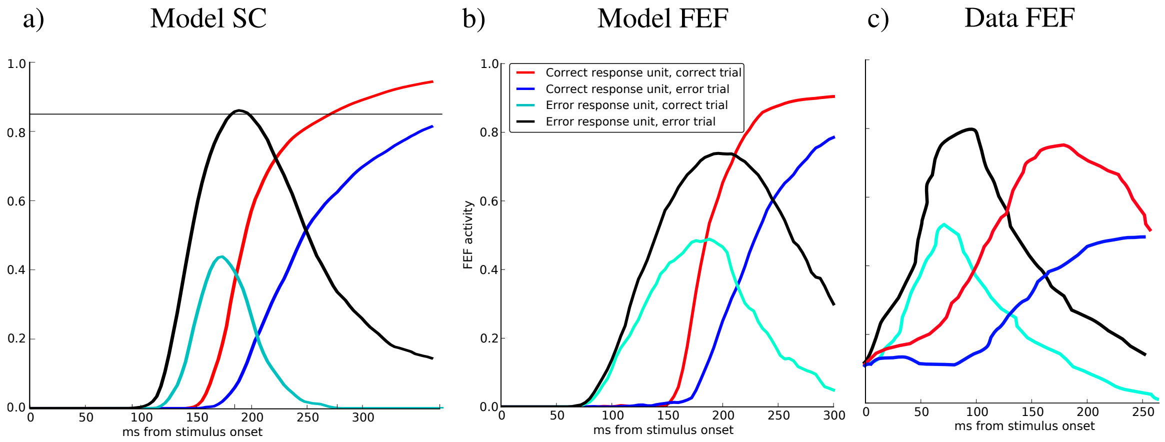

Figure 2.6: **a)** Average activity of individual superior colliculus (SC) units coding for the correct and error response in correct and incorrect trials during incongruent trials aligned to stimulus onset. The prepotent (i.e. erroneous) response comes on before the volitional, correct response. In incorrect trials the error-unit threshold is crossed before the volitional response unit gets active. In correct trials the error-unit is inhibited in time. **b)** Average activity of individual FEF units coding for prepotent error responses and volitional correct responses during incongruent trials aligned to stimulus onset (benchmark pattern #6) **c)** Electrophsyiological recordings in FEF of monkeys (reproduced from Everling and Munoz (2000)).

SC and FEF activity Comparing single unit activation patterns of SC (see figure 2.6a) to those of FEF (see figure 2.6b) reveals that the activation dynamics are very similar between those two regions. Our model thus predicts that FEF can be interpreted as a cortical saccade planning/monitoring area that directly influences saccade generation via its projections to SC (Munoz and Everling, 2004). Moreover, SC activity reveals that in both, correct and incorrect incongruent trials, the incorrect prepotent response unit becomes active before the controlled one, thus matching qualitative pattern #5.

dACC activity As described earlier, the dACC computes co-activation of both response units in FEF (i.e. when average activity is > 0.5) – a direct measure of conflict (or value of the alternative action to that initially considered; see above). Consequently, its activity (see figure 2.7a) follows a similar pattern as average FEF layer activity: conflict is present but resolved prior to responding in correct trials

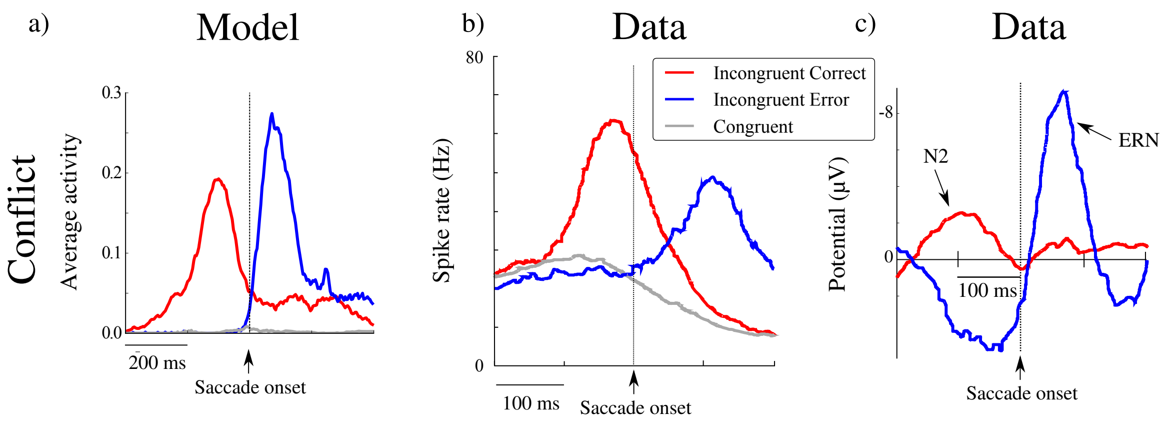

Figure 2.7: **a)** Averaged dACC activity (corresponding to conflict in FEF) in prosaccade and correct and incorrect incongruent trials. No conflict is present in congruent trials. During correct incongruent trials, conflict is detected and resolved before the response is gated. During incorrect incongruent trials, an incorrect response is made before conflict is detected. **b)** Activity recorded in monkey pre-SMA during the switch-task (reproduced from Isoda and Hikosaka (2007)). **c)** EEG recordings from the central scalp of humans during the Flanker task (reproduced from Yeung et al. (2004a)), thought to originate from dACC. The N2 and ERN component closely match our modeling results, replicating this aspect of the Yeung model.

while conflict is present *after* responding in error trials. However, dACC does not get active in congruent trials, because it never shifts from one action to the other.

This qualitative pattern of peak conflict activation *before* correct incongruent trials but *after* incorrect incongruent trials matches event-related potentials (ERPs) commonly observed in human EEG studies (see figure 2.7c). The so-called *error related negativity* (ERN) which is measured *after* response errors whereas the so-called *N2* potential is measured *before* correct high conflict responses (Falkenstein et al., 1991; Gehring et al., 1993). The idea that these two signals could merely represent 'two sides of the same conflict coin' and reflective of underlying dACC activity was first presented in the modeling work by Yeung and colleagues (Yeung and Cohen, 2006; Yeung et al., 2004b).

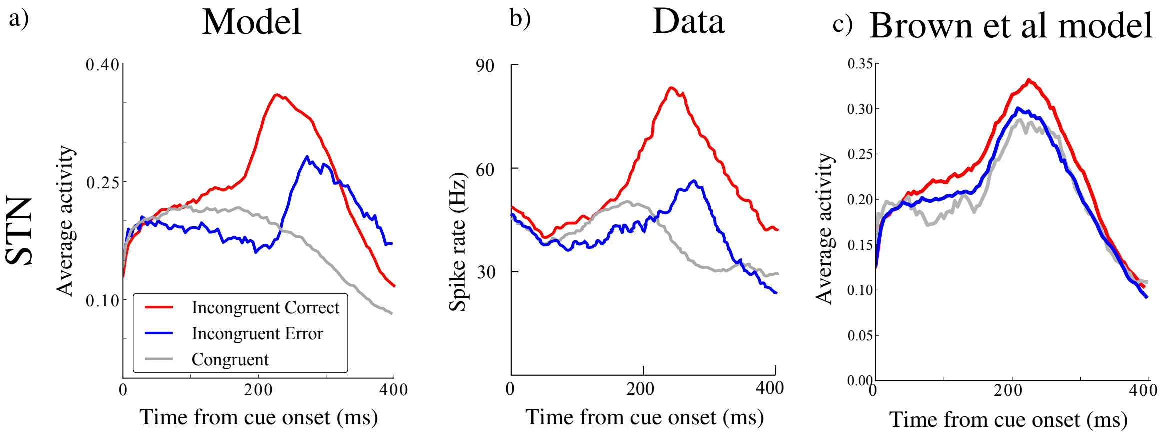

Figure 2.8: **a)** Averaged activity of the model STN layer during prosaccade and correct and incorrect incongruent trials relative to response execution. During congruent trials STN units exhibit a small early increase in activity that subsides. Correct incongruent trials show increased activity early on in the trial which causes the conflict-induced slowing and prevents prepotent response gating. In error trials, this mechanism is triggered too late and the incorrect response gets executed. **b)** Electrophysiological recordings of the monkey STN (reproduced from Isoda and Hikosaka (2008)) on correct and incorrect switch trials and non-switch trials. **c)** Average activity of the STN layer of an alternative model in which STN is not excited by dACC but instead by saccadic output (SC in our model) as proposed by Brown et al. (2004). This model does not predict differences between trial types.

STN activity As noted in the model description, conflict detection in the dACC results in delayed (and more accurate) responding by recruiting the STN to prevent gating until conflict is resolved. Indeed, this mechanism is in part responsible for the rightward-shifted RT distributions in correct incongruent trials. Accordingly, this same pattern of increased activity before correct responses and increased activity after error responses can be observed in STN (see figure 2.8a). Again, this qualitative pattern was also found in STN recordings in monkeys by Isoda and Hikosaka (2008) (see figure 2.8b), who showed that timing of STN firing relative to pre-SMA was consistent with communication along this hyperdirect pathway.

The neurocomputational model of (Brown et al., 2004) interprets the role of STN differently. In their model, STN is activated by the output structure (FEF in their case) to lock out the influence of competing responses *after* a response has been selected. This is a critical difference to the account presented herein where STN plays a role in the selection of a response by raising the threshold prior to response selection, thereby delaying execution but increasing accuracy. To show explicitly how our model predictions can be qualitatively differentiated from this alternative model of STN function, we disconnected dACC inputs into the STN and instead allowed only the output structure (SC in our model) to project to it, so that STN function operates as it does in Brown et al (2004). As can be seen in figure 2.8c, the activity pattern changes dramatically. Specifically, there is no more differentiation of activation patterns between the different trial types as is observed in our model and the empirical data (Isoda and Hikosaka, 2008). Because STN only influences processing after response selection, it also does not lead to delayed responding or decision threshold adjustment. This qualitative difference in model predictions is fundamental and not subject to parameter tuning, as it reflects a distinct computational role for the STN. Although we focused on the Brown model for demonstration purposes here, other models of STN function with different connectivity would similarly not account for these data. For example, the biophysical model of Rubchinsky et al. (2003) assumes that STN neurons provide focused selection of a particular action (by disinhibiting SNr, taking the role of the direct Go pathway) while simultaneously inhibiting competing actions (by exciting SNr in other columns). This model cannot explain this activity pattern because co-activation of multiple cortical inputs does *not* result in increased STN activity (see figure 6b in Rubchinsky et al. (2003)).

Striatal activity Figure 2.9a shows striatal activity in congruent and incongruent trials (column I and column II, respectively) for direct-path Go and indirect-path NoGo units (upper and lower rows, respectively). In each case, activity selective to the correct and error responses are color coded. The model closely captures the qualitative pattern across four cell populations (#7) identified in monkey dorsal striatum recordings during the antisaccade task (see figure 2.9b and Watanabe and Munoz

(2009); Ford and Everling (2009)). In particular, for congruent trials, correct-coding Go neurons gate the response while error-coding NoGo units suppress the alternative. In incongruent trials, Go neurons for the error-coding prepotent response are initially activated, but are then followed by increased activity of the corresponding NoGo population which then suppresses the initiated Go activity via NoGo→Go inhibitory projections (Taverna et al., 2008). Finally, the controlled Go-correct units are activated and an incongruent response is executed. Thus our model predicts that the pattern of electrophysiology observed in empirical recordings arises due to top-down cognitive control modulation of direct and indirect pathway neurons.

Note again that we can distinguish our model's predictions from those of other models that omit the indirect pathway as a distinct source of computation (there are several) or from models that do include it but assign a different function. The neural network model of Brown et al. (2004) assumes the indirect pathway activation defers execution of the correct action plan until the time is appropriate. This would suggest that the executive control complex would activate NoGo units coding for the *correct* response, not the incorrect response as in our model. To demonstrate how this leads to qualitatively different patterns than is observed in our model and the data (see pattern #7 and figure 2.9c) in which this alternative account is simulated in our model. (However, we note that the Brown et al model could potentially accord with our model in the sense that they also advocate a mechanism by which negative prediction errors drive learning in the NoGo cells, which after training on the AST may also produce the patterns we observe here given that the prepotent response would be punished). Similarly, the prominent model of Gurney et al. (2001) suggests that this pathway serves as a control pathway rather than providing negative evidence against particular actions as in our model, and it is unclear how this control function (while not disputed per se) would reproduce the patterns observed here.

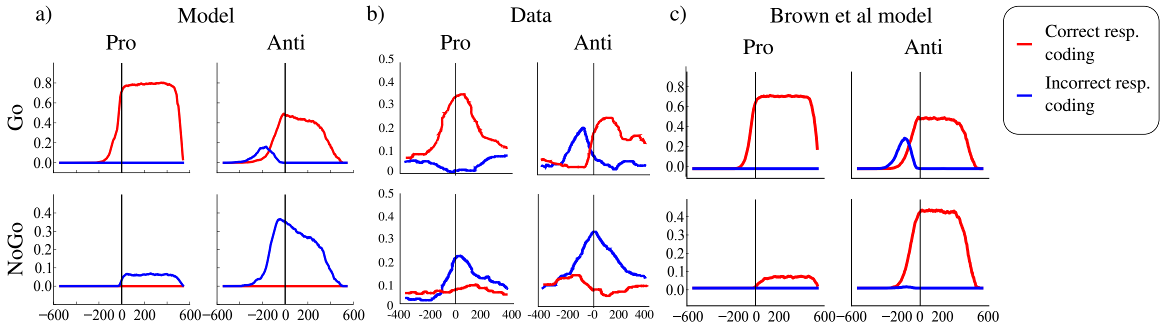

Figure 2.9: **a)** Averaged striatal activity during correct pro (first column) and incongruent trials (second column) in Go (first row) and NoGo (second row) neuronal populations. In each case, activity for correct (red) and error (blue) response coding units are shown separately. As described in the text, the Go units for the prepotent response become active early in the trial for both trial types, but in antisaccade trials these are followed by NoGo units which veto the Go activity and finally Go activity for the controlled response due to top-down DLPFC activity. **b)** Electrophysiological recordings of the monkey striatum (reproduced from Watanabe and Munoz (2009)). The first row represents neurons coding corresponding to the executed response (i.e. Go neurons) and the second row represents neurons coding that suppress execution of the corresponding action (i.e. NoGo neurons). **c)** Alternative model simulating Brown et al. (2004) assumption that the indirect pathway acts to defer the execution of the correct response, rather than suppress the alternative response. Note predictions for Go pathway accord with those of our model and the data, but prediction of NoGo neurons differ.

2.3.2 Global Response Inhibition

Methods

In SRITs the selectively inhibited prepotent response must be replaced with another, controlled response. Conversely, the stop-signal task (SST) requires outright response inhibition (e.g. Logan and Cowan, 1984; Aron and Poldrack, 2006; Cohen and Poldrack, 2008) and is used to assess global inhibitory control (Aron, 2011). Specifically, subjects are required to make press left and right keys in response to Go-cues appearing on a screen. On a subset of trials *after* the Go-cue has been presented, a stop-signal is presented after variable delay (i.e. stop-signal delay; SSD) instructing the subject to withhold responding.

Here we show that our model can also simulate the SST after we included the right inferior frontal cortex (rIFG) with direct projections to STN (Aron et al., 2007a) see figure 2.10. Given the assumptions of the race model (i.e., a race between Go and Stop processes), one can estimate the stop-signal reaction time (SSRT) by measuring the probability of successful inhibition at different SSDs. This inhibition function is then compared to the distribution of Go reaction times in non-stop signal trials. There are several extensive reviews of the SST (Verbruggen and Logan, 2009b), so here we focus on how our model captures the available evidence. Note that the SST typically refers to the task involving manual movements (and inhibition thereof), but a well studied equivalent has been used in the oculomotor domain, where it is referred to as the *'countermanding task'*. While the neuronal circuitry involved in Go-responding depends on the response modality, the neuronal circuitry involved in the global mechanism may be independent of the response modality (Leung and Cai, 2007).

Networks are presented with one of two input stimuli (left or right), represented

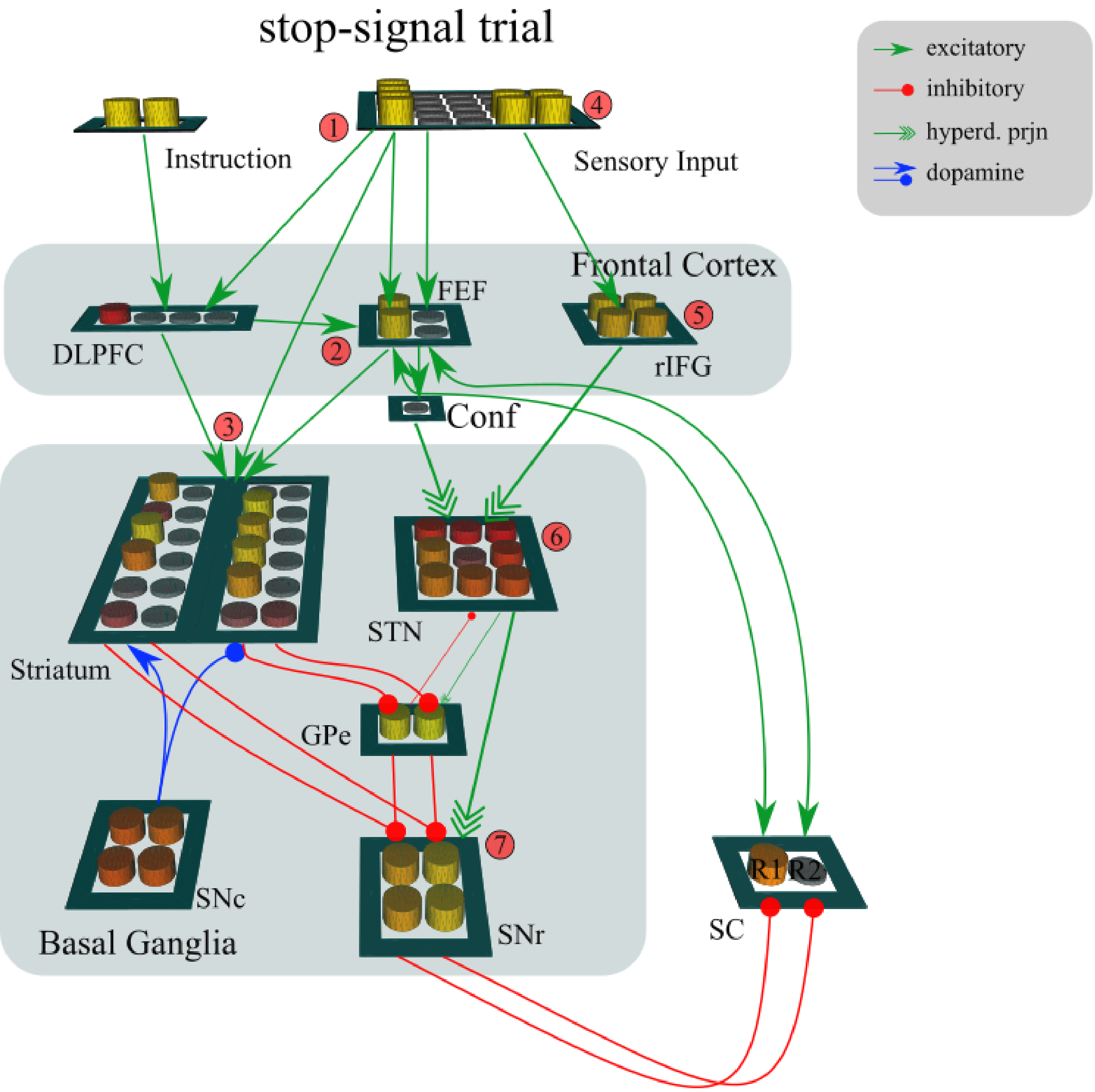

Figure 2.10: Extended neural network model including rFIG during stop-signal trials. (1) Left input stimulus activates (2) left-coding FEF response units and (3) initiates gating via striatum (similar to pro-saccade trial in a). After a delay, (4) the stop-signal is presented which activates (5) rIFG, which in turn (6) transiently activates the STN and finally (7) the whole SNr to globally prevent gating. Note, that DLPFC is beginning to get active to initiate selective response inhibition via striatal NoGo units.

by a column of four units each. As in prior simulations, prepotent responses are implemented by weights from the input units to the corresponding FEF response units, such that a left stimulus suggests a left response. On 25% of trials, a stop-signal is presented with variable delay (by activating devoted units in the sensory input layer). The stop signal units send excitatory projections directly to the rIFG layer. rIFG units in the hyperdirect pathway excite the STN (Aron et al., 2007a; Neubert et al., 2010) and prevent striatal response gating, and therefore inhibit responding if the SC has not already surpassed threshold. In addition to this global rIFG-STN response suppression mechanism, the DLPFC combines the stop-signal input and the stimulus location to selectively inhibit the associated response via activation of the corresponding population of striatal NoGo units. Critically, this selective mechanism is slower but remains active after the STN returned to baseline and prevents subsequent responding. Thus, the model uses a fast, global but transient response inhibition mechanism and a slower, selective but lasting mechanism (Aron, 2011). To estimate the SSRT, we use the dynamic one-up / one-down *staircase procedure* for adjusting the SSD (e.g. Logan et al., 1997; Osman et al., 1986).

We tested the influence of rIFG lesions on the SSRT (Aron et al., 2004) by parametrically reducing the projection strength of rIFG to the STN.

The selective norepinephrine (NE) reuptake inhibitor Atomoxetine increases NE release and improves stop-signal performance in animals, healthy adults and adult ADHD patients (Chamberlain et al., 2007, 2009). NE is hypothesized to adaptively change the activation gain of neurons in frontal cortex (Aston-Jones and Cohen, 2005). We consequently tested the influence of decreasing the gain parameter in units of the frontal cortex[2].

[2] Gain modulates how step-like the activation-dynamics of units are in relation to their input activity. Low gain leads to linear activation dynamics while high gain levels make a unit respond in a binary-like fashion.

Finally, we simulated different motivational influences on stop-signal accuracy. Evidence for the neural underpinnings of motivational biases comes from an fMRI study by Leotti and Wager (2010), who reported that subjects instructed to focus on speed instead of accuracy exhibited a greater increase in activations in brain regions associated with response facilitation, including the FEF and the striatum. Conversely, when instructed to focus on accuracy, subjects exhibited greater activity in IFG regions associated with response inhibition. We thus simulated these activation patterns to account for speed-accuracy tradeoff in a similar manner as in the antisaccade simulations. In the speed-condition, we manipulated the strength of FEF to striatum connections due to evidence that frontostriataal connectivity is enhanced under speed emphasis (Forstmann et al., 2008, 2010a; Mansfield et al., 2011). Conversely, in the accuracy condition we increased baseline excitatory input to rIFG, allowing it to be more excitable and hence facilitating STN recruitment. This simulation approximates the effect of a putative PFC rule based representation to focus on accuracy. Recent data supports the notion that the (right) STN, which receives input from rIFG, shows increased excitability associated with an increased response caution during accuracy focus (Mansfield et al., 2011).

Results

As with the SRITs above we extracted a list of key qualitative results from the literature we use to evaluate the fit of our model.

#1 The probability of inhibiting a response decreases monotonically as SSD increases (Verbruggen and Logan, 2008).

#2 Error responses that escape inhibition are, on average, faster than Go responses on no-stop-signal trials. However, while the distributions begin at the same minimum value, the responses that escape inhibition have a shorter maximum value (Verbruggen and Logan, 2008).

#3 STN neurons are excited to stop signals but show little differentiation between stop-signal inhibition and stop-respond error trials (Aron et al., 2007a). Contrary, downstream SNr neurons are excited in correct trials but are disinhibited during error trials (Schmidt et al., 2012).

#4 SEF neurons are activated in stop-signal and stop-response trials *after* SSRT and can thus not contribute to successful stopping (Stuphorn et al., 2000).

Behavior

To illustrate the staircase procedure, figure 2.11(a) shows an example trace of how SSDs are adjusted to assess 50% stop-signal accuracy. As can be seen, the network with rIFG lesion is impaired at stopping and requires shorter SSD on average to inhibit successfully.

As can be seen in figure 2.11(b) the inhibition function resulting from testing the neural network systematically with different SSDs reveals a monotonically decreasing probability of correctly stopping (qualitative pattern #1).

Cumulative RT distributions of Go and non-canceled Stop trials are presented in figure 2.12. Both distributions match closely up until SSD+SSRT (qualitative pattern #2) suggesting that both are generated by the same process.

Different modulations affect GoRT and SSRT in different ways (figures 2.13(a) and 2.13(b)). While DA manipulations certainly speed GoRT, SSRT remains largely unaffected. On the other hand, when the network is tested with reduced gain (simulating low NE levels), or has lesions to either STN or rIFG, it exhibits SSRT deficits (increases). Finally, simulated accuracy emphasis results in slowed Go RT

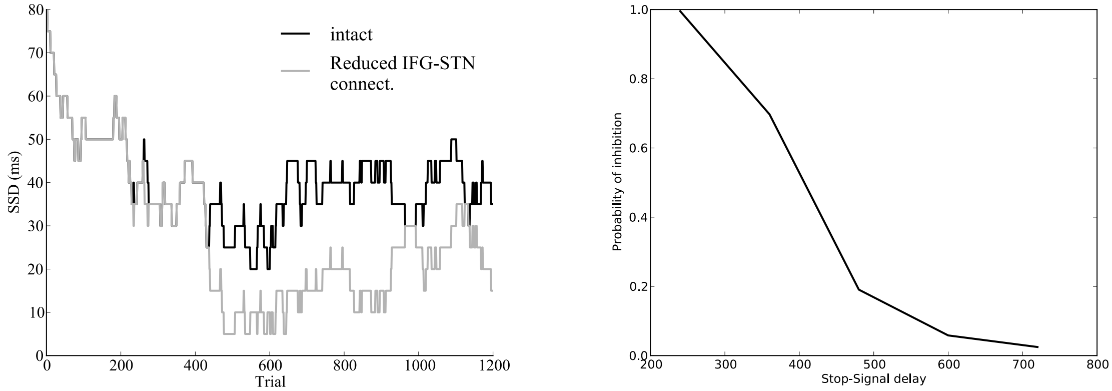

Figure 2.11: **a)** Progression of the staircase procedure for manipulating SSD in networks with reduced rIFG-STN connectivity. Trial number is plotted on the x-axis and the stop-signal delay (SSD) in ms (converted from simulator time) is plotted on the y-axis. If a response is successfully inhibited on stop-signal trial, the SSD is increased by 20 ms to make it harder. If a response is erroneously made on a stop-signal trial, the SSD is decreased by 20 ms. Networks without lesion are highest in general representing the most effective Stop-process that is able to withhold responses even when the SSD is quite long. **b)** Inhibition function of the neural network model in the stop-signal task. The model is tested on systematically varying levels of stop-signal delay (SSD) in ms and the proportion of correctly inhibited trials is plotted along the y-axis.

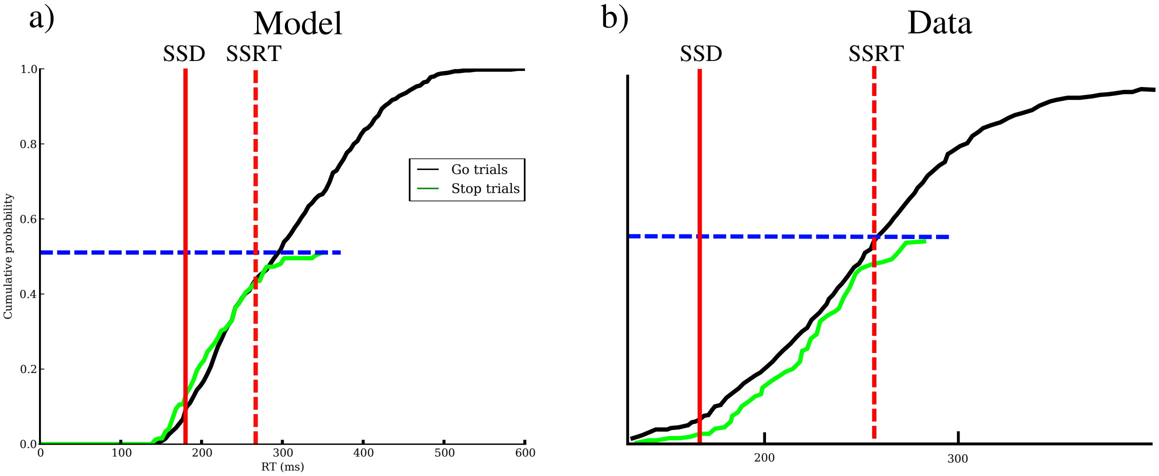

Figure 2.12: **a)** Cumulative reaction time distributions of the neural network model and from a monkey experiment. **b)** Cumulative reaction time distribution from a monkey experiment for comparison. Reproduced from (Lo et al., 2009). The solid red line denotes mean stop-signal delay (SSD); the broken red line denotes stop-signal reaction time (SSRT) offset at SSD. The broken blue horizontal line represents 50% stopping accuracy. Note that the response distribution sums to the response probability – not necessarily to 1.

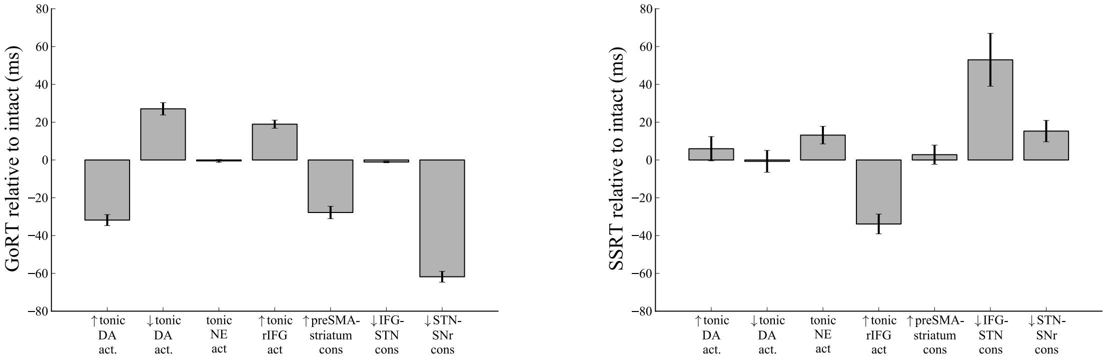

Figure 2.13: **a)** Mean RTs in ms ± SEM (converted from simulator time) for Go trials under different modulations (see text). **b)** Mean SSRTs in ms ± SEM (converted from simulator time) under different modulations (see text).

but faster SSRT (more effective inhibition). The pattern that emerges from these results is that SSRT is changed by modulations of parameters that are part of the global inhibitory pathway: rIFG and STN.

Neurophsyiology

To assess the neural correlates of stopping behavior in our model we analyzed STN and SNr activity aligned to stop-signal onset. As can be seen in figure 2.14, there is little differentiation between stop-signal inhibition and error trials while SNr units show a marked dip in error trials that is less pronounced in inhibition trials (qualitative pattern #3).

We moreover analyzed the activity pattern of our executive control complex which consists of DLPFC, SEF and pre-SMA. As can be seen in figure 2.14, activation is observed in stop-signal trials (both stop-respond and successful inhibitions) only *after* SSRT and could thus had no influence on the stopping (qualitative pattern #3). This result implies that global stopping to salient stop signals is most likely driven by the fast stop process along the rIFG-STN hyperdirect pathway. We ascertain that

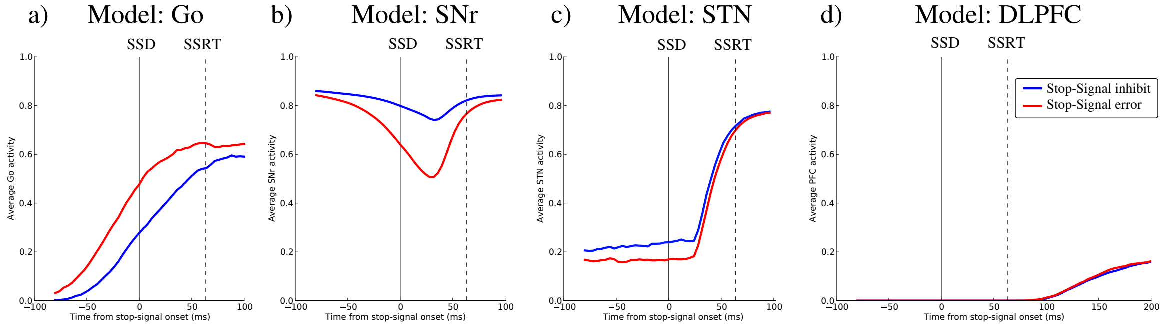

Figure 2.14: Average activity aligned to stop-signal onset for inhibited and error stop-signal trials. **a)** Striatal Go-neuronal activity. **b)** Substantia nigra pars reticulata activity. **c)** Subthalamic nucleus activity. **d)** Activity of the executive control complex consisting of DLPFC, SEF and pre-SMA.

executive control processes are delayed relative to this global stopping mechanism, and may participate in selective response inhibition (and in the stop-change task, activation of the correct response) after the global response pause has passed.

2.4 Discussion

The interaction between executive control and habitual behavior is a central feature of higher-level brain function, and plays a role in various domains from cognitive psychology (under the rubric of "system 1" vs. "system 2"; (Evans, 2005)) to machine learning (model-free vs. model based control (Daw et al., 2005)). At the core of this interaction is a mechanism that allows executive control to override the habitual system and guide action selection. A multitude of psychological cognitive tasks have been used to probe the nature of this interaction. The stop-signal task requires outright stopping of a response already in the planning stage. The antisaccade (Hallett, 1979), Simon (Simon, 1969), and saccade override (Isoda and Hikosaka, 2007, 2008) tasks all involve inhibition of a prepotent action together with initiation of an action incompatible with the prepotent one. Despite the apparent behavioral simplicity of these tasks, various lines of research have revealed a highly complex and tightly interconnected brain network underlying response inhibition consisting of frontal areas including DLPFC, SEF, pre-SMA, FEF, rIFG, and dACC and basal ganglia structures including the striatum and STN.

We presented a dynamic neural network model of selective and global response inhibition which provides a description of the distributed computations carried out by individual brain regions and neurotransmitters. The complexity of this model is grounded by well established neuroanatomical and physiological considerations, and accounts for a wealth of key data including electrophysiology, psychiatric and pharmacological modulations, behavioral, lesion and imaging studies. Moreover, this model is constrained (i) by using a single parameterization across all simulations of intact function and (ii) by the multitude of qualitative results from different levels of analysis it is required to reproduce. Although we used one parameterization across the intact model simulations, we also generalized the functionality via systematic manipulations across a range of parameters. In other work (Wiecki & Frank, in prep), we have shown that the emerging fundamental computational properties of this complex system as a whole are captured by analysis using a modified drift diffusion model, in which distinct mechanisms within the neural model (e.g., STN projection strength, DLPFC speed) are monotonically related to high level decision parameters (e.g., decision threshold, and drift rate of the executive process).

2.4.1 Selective Response Inhibition

In our SRIT simulations, the model assumes that prepotent, reflexive actions such as a saccade to an appearing stimulus (e.g. a prosaccade) are selected via the cortico-cortical route and swiftly gated by the BG. An abundance of data supports the general involvement of the BG in saccade generation and inhibition (e.g. Hikosaka et al., 2000; Hikosaka, 1989; Hikosaka and Wurtz, 1986). Conversely, the cognitive control system not only represents the task rules needed to respond appropriately (e.g. in DLPFC), but also incorporates a downstream mechanism in dACC-STN to detect when these rules indicate an alternative action than was originally initiated. Thus our model synthesizes the popular account of dACC in terms of response conflict (Botvinick et al., 2001) with recent studies suggesting that dACC rather reflects the value of the alternative action (Kolling et al., 2012). Moreover, via the hyperdirect pathway to the STN, this mechanism serves to transiently increase the BG gating threshold to prevent prepotent actions from being facilitated and allows more time for the controlled PFC-striatal mechanisms to selectively suppress this response and to facilitate appropriate alternative courses of action. It has also been shown that the SEF, FEF (Munoz and Everling, 2004), dACC (Botvinick et al., 2004), pre-SMA (Isoda and Hikosaka, 2007) and STN (Isoda and Hikosaka, 2008) are involved in detecting conflict between a planned response and the current rule, and for switching from an automatic to a volitional response (e.g., antisaccades).

To detect conflict between reflexive and controlled responses, the system needs to be able to compute the correct identity of the controlled response itself. In the model, the DLPFC integrates task instructions and current stimulus location and forms a conjunctive rule representation (Wallis and Miller, 2003; Bunge and Wallis, 2008) that then provides evidence for the associated controlled response via its projection to the FEF, and further biases the gating of this response (and the selective suppression of the reflexive response) via striatum. We demonstrated that this is a necessary condition for our model by showing that a model with faster integration speeds fails to account for key behavioral patterns.

Thus it should be clear that compared to a congruent response, an incongruent response should (i) be more prone to error because it depends on successful inhibition of prepotent actions which may be close to threshold by the time conflict is detected and (ii) take longer due to (iia) additional computation needed for the DLPFC to perform the requisite vector inversion (activation of correct rule representation among multiple competitors based on an integration of input and instruction), and (iib) the delay in commitment to a response resulting from the increase in decision threshold along the hyperdirect pathway.

Early cognitive models of interference control assumed a dual-route mechanism for action selection, including an automatic response route and a volitional one (Kornblum et al., 1990; Eimer, 1995; DeLiang et al., 2005; Ridderinkhof, 2002). This model was extended to include selective suppression of the automatic response by the volitional response mechanism (i.e. the activation-suppression model (Ridderinkhof, 2002; Ridderinkhof et al., 2011)). Our model shares these attributes but makes two crucial contributions to this discussion: (i) strong predictions on the neural correlates of these abstract cognitive processes, and (ii) a raise in decision threshold requiring more evidence to gate *any* response. This latter mechanism may not only be adaptive as a fast route to prevent gating of prepotent actions, but could also serve to increase the likelihood that the alternative action selected is the most accurate (particularly when there may be more than one, as is often the case in more realistic executive control scenarios than those typically studied in simple response inhibition tasks).

Response time distributions and errors: Neural underpinnings

At the behavioral level, our intact model reproduces the same patterns found empirically – networks made more errors (see figure 2.3(a)) and were in general slower (see figure 2.3(b)) on incongruent trials compared to congruent trials (e.g. Reilly et al., 2006; Harris et al., 2006; Reilly et al., 2007; McDowell et al., 2002). Incongruent errors were more likely to occur when networks responded fast (see figure 2.4, 2.5(a) and Ridderinkhof et al. (2011)). These errors result primarily from variance in the speed of cognitive control (DLPFC), but also in the prepotent response (in some trials gating is faster than others) and in the inhibition process (in some trials the hyperdirect pathway and/or striatal NoGo process is slower). Moreover, reduced DLPFC connectivity also degrades accuracy on incongruent trials, mirroring the empirical performance degradation in antisaccade tasks during development associated with reduced DLPFC connectivity (Hwang et al., 2010). A more explicit investigation into the dynamics of these processes comes from the simulated electrophysiology across brain regions and trial types.

Conflict- and error-related activity: relation to existing models

The Error Related Negativity (ERN) is an event-related potential associated with errors made in forced-choice reaction time tasks (Falkenstein et al., 1991; Gehring et al., 1993). The ERN reaches its peak within 100 ms *after* the erroneous response. Using a connectionist model, Yeung and collegues hypothesize the ERN to reflect conflict between the executed, erroneous response and the still-evolving activation of the correct response (Yeung and Cohen, 2006; Yeung et al., 2004b). Thus, the error detection mechanism reflects an internal correction of the executed response, leading to a transient period of response conflict. According to this same framework, a similar potential should be observed in high conflict trials *before* correct responses, when conflict is resolved prior to responding. These authors indeed reported that the N2 potential exhibited just this profile and argued that it reflected the same underlying conflict mechanism in the dACC.

Our dACC node exhibits the same qualitative pattern of increased activity (i) before correct incongruent responses, (ii) after incorrect incongruent responses and (iii) baseline activity during congruent responses. However, this pattern is not unique to ERPs thought to originate from dACC, but is also found in electrophysiological recordings in pre-SMA, SEF (Emeric et al., 2010) and STN (Isoda and Hikosaka, 2008). Our model provides an explicit framework that recapitulates these effects and explores their influences on behavior. Together, these dynamics accord with our earlier assertion that our model synthesizes the conflict model with the notion that the dACC reflects the value of the alternative action: this network only becomes activated when the alternative action is deemed to be more correct than the prepotent one. This process occurs either prior or following response execution (as in the conflict monitoring account), but must always occur after the initial activation of an incorrect (often prepotent) response (not specified by the conflict account but consistent with the alternative action value account). To more formally describe the computational dynamic implicated, we devised a modified drift diffusion model which explicitly incorporates this reversal in evidence.

2.4.2 Global Response Inhibition

By adding a single rIFG layer to our model we generalized our model to capture data from global response inhibition tasks such as the SST. As we demonstrated above, this model recovers key qualitative behavioral patterns reported in the literature. Moreover, model neurophysiology revealed interesting similarities to recent rat electrophysiological recordings in the SST (Schmidt et al., 2012). Specifically, while STN activity surges in response to the stop signal to the same extent regardless of whether the response is successfully inhibited or not, activity in the SNr strongly differentiates between these trial types. During errors, the striatal Go signals were potent and early enough to inhibit SNr activity in spite of the STN surge. These results suggest that the source of response inhibition errors is variance in the Go process, but that the duration of the stop-process is rather fixed. This conceptualization matches closely with the interactive horse-race model (Verbruggen and Logan, 2009a). Here, we hypothesize that the critical point of interaction between the two processes is the SNr.

Why did we add an rIFG layer given that our initial model already contained an executive control complex including DLPFC? As described above, rIFG and STN involvement in the SST is well established, and moreover, simulations showed that the activations in our executive control complex needed to account for SRITs was too slow to account for global response inhibition needed in SST. Nevertheless, the nature of the (potentially separable) mechanisms engaged for detecting when inhibitory control is necessary, and how it should be implemented, remains largely elusive. In particular, the role of rIFG is actively debated. Some studies specifically implicate the rIFG in response inhibition (Verbruggen et al., 2010; Aron et al., 2003; D. et al., 2007; Sakagami et al., 2001; Xue et al., 2008), whereas others report rIFG activity in tasks lacking pure response inhibition demands, suggesting that it is more involved in monitoring or salience detection (Sharp et al., 2010; Verbruggen et al., 2010; Hampshire et al., 2010; Fleming et al., 2010; Chatham et al., 2012; Munakata et al., 2011). Our model unifies these two seemingly opposing views by arguing that the rIFG in fact detects salient events and, via downstream processing, engages a stopping mechanism whether or not it is required by the task rules. In both the stop-signal and stop-change task, subjects have to detect an infrequent signal which tells them to update their current action plan. We argue that these signals are salient events and, via noradrenergic modulation, enhance processing in the rIFG which, in turn, causes an *orienting* or *circuit breaker* response by activating the STN (Swann et al., 2011a) to pause response selection. This pause enables the volitional DLPFC based response selection mechanism to take control and either inhibit a specific response (as in the stop-signal task) or initiate a new response (as in the stop-change task). This theory of a rIFG triggering a global response-pause is supported by rIFG involvement in the oddball task (Stevens, 2000; Huettel and McCarthy, 2004) which requires no behavior adaptation whatsoever, yet still causes response slowing (Barcelo et al., 2006; Parmentier et al., 2008). Indeed, in many of the above-reported studies in which rIFG is activated under conditions of monitoring or saliency, when they have been reported, subject RTs were nevertheless delayed despite no overt inhibitory demands (Sharp et al., 2010; Fleming et al., 2010; Chatham et al., 2012).

2.4.3 Different forms of response inhibition

Inhibitory control can be issued globally or selectively (Aron and Verbruggen, 2008; Aron, 2011). The brain seems to revert to a global inhibitory mechanism when unexpected events occur that require quick response adaptation (e.g., stop-signals), and to a selective inhibitory control mechanism when response inhibition can be prepared (Greenhouse et al., 2011; Hu and Li, 2011). We propose that selective inhibition of the prepotent response is initiated by the DLPFC and implemented via the indirect corticostriatal NoGo pathway (Zandbelt and Vink, 2010; Watanabe and Munoz, 2009, 2010; Hu and Li, 2011; Jahfari et al., 2011). Global response inhibition on the other hand is driven by a salience detection mechanism implemented in the rIFG which directly projects to the STN to inhibit responding (Mink, 1996; Nambu et al., 2000, 2002; Kuhn et al., 2004; Aron et al., 2007b; Eagle et al., 2008; Isoda and Hikosaka, 2008; Jahfari et al., 2011, 2012).

In addition to the selectivity of inhibitory control, differences exist between proactive and reactive initiation of response inhibition (Aron, 2011; Greenhouse et al., 2011; Swann et al., 2011b; Cai et al., 2011). Our modeling work suggests multiple possible sources for proactive control. Speed-accuracy adjustments are implemented by increasing functional connectivity between frontal motor regions and striatum to decrease the decision threshold under speed emphasis (see figure 2.5(b), 2.5(c) and Lo and Wang (2006); Forstmann et al. (2010a)). The second proactive mechanism increases response caution by increasing baseline rIFG activity to prime saliency detection and slow responding via the rIFG-STN hyperdirect pathway (see figure 2.13(b)).

Interestingly, while FEF→striatum functional connectivity influence speed and accuracy in our SRIT simulations, SSRT in the stop-signal task is unaffected by this modulations and is only improved by an increase in tonic rIFG activity. This suggests that proactive control in form of mere response slowing is uneffective in reducing SSRT – the staircase procedure adapts to slower overall responding – but that enhanced attentional monitoring has preferential influence on global inhibitory control. In other words, although all these mechanisms can lead to adjustments in decision threshold, only those associated with active engagement of the stop process will facilitate inhibitory control per se. If confirmed, this result may have implications for refining therapy of inhibitory control disorders like addiction, obesity and OCD. Nevertheless, it remains important to emphasize that the striatal NoGo pathway is also thought to help to prevent the proactive selection of maladaptive responses.

2.4.4 Multiple mechanisms of response threshold regulation in fronto-basal-ganglia circuitry at different time scales

Different mechanisms in our neural network influence the gating threshold for initiating motor responses at distinct time scales, and modulated by distinct cognitive variables. First, the strength of cortico-striatal projections regulate the ease with which cortical motor plans can be gated by the BG, allowing for speed emphasis in the speed-accuracy tradeoff (see figure 2.5(c)). This aspect of our model is quite similar to the model of Lo and Wang (2006) and was subsequently corroborated by Forstmann et al. (2010a). Our model converges on the same conclusion but extends this view by showing that gating threshold is also more dynamically regulated on a shorter time-scale by (i) motivational state (changes in DA levels, which are modulated by reinforcement and also facilitate striatal Go signals); and (ii) response conflict and saliency (via the hyperdirect pathway, making it more difficult or Go signals to drive BG gating (Jahfari et al., 2011)). Moreover, STN efficacy in the neural model is positively correlated with increases in estimated decision threshold (Ratcliff and Frank, 2012). Evidence for conflict-induced decision threshold adjustment via the hyperdirect pathway has been recently described in a reinforcement-based decision making task (Cavanagh et al., 2011). Increases in frontal EEG activity during high conflict decisions were related to increases in decision threshold estimated by the drift diffusion model. Intracranial recordings directly within the STN also revealed decision conflict-related activity during the same time period and frequency range as observed over frontal electrodes (see also Zaghloul et al. (2012)). Moreover, disruption of STN function with deep brain stimulation led to a reversal of the relationship between frontal EEG and decision threshold, without altering frontal activity itself. These data thus support the notion that frontal-STN communication is involved in decision threshold adjustment as a function of conflict. Similarly, proactive preparation to increase decision threshold in the stop signal task when stop signals are likely is associated with hyperdirect pathway activity (Jahfari et al., 2012).

In our neural models, conflict-related STN activity subsides with time (see figure 2.8), due to resolution of conflict in FEF/ACC, feedback inhibition from GPe, and neural accommodation. Thus a more refined description of this transient STN surge is that it initially increases the decision threshold (more so with conflict), followed by a dynamic collapse of the decision threshold over time. Indeed, a recent multilevel computational modeling and behavioral study by Ratcliff and Frank (2012) supported this idea by showing that a collapsing threshold diffusion model provided good fits to both the BG model and to human participant choices in a reinforcement conflict task. Moreover, the temporal profile of the best-fitting collapsing threshold corresponded well to the time course of the collapse in STN activity across time.

2.4.5 Psychiatric disorders and differential effects of dopamine and norepinephrine

Abnormal striatal DA signaling is hypothesized to be at the core of many disorders, including PD (Bernheimer et al., 1973), SZ (Breier et al., 1998) and ADHD (Casey et al., 2007; Frank et al., 2007b). Intriguingly, all of these disorders are linked to response inhibition deficits in the stop-signal task. Our earlier BG models have successfully accounted for a wide variety of findings associated with striatal DA manipulations across reinforcement learning and working memory tasks (for review, Wiecki and Frank, 2010). Yet, we found here that striatal DA manipulations, while affecting overall RT, had negligible effects on response inhibition deficits as assessed by SSRT (see figure 2.13(b)). This prediction converges with recent evidence (reviewed in, Munakata et al., 2011) showing that levodopa, a drug that increases DA levels in striatum (Harden and Grace, 1995), had no influence on SSRT in PD patients (Obeso et al., 2011a,b).

This lack of DA effect raises the question of the source of the response inhibition deficits in the aforementioned disorders. One conspicuous candidate is abnormal NE functioning as suggested by evidence in both ADHD (Faraone et al., 2005; Ramos and Arnsten, 2007; Frank et al., 2007c) and PD (Farley et al., 1978). In our simulations, NE modulation influences SSRT via its gain-modulatory effects in rIFG (Aston-Jones and Cohen, 2005). Additional support for this account comes from pharmacological experiments using the selective norepinephrine reuptake inhibitor atomoxetine, which improves response inhibition performance in animals, healthy adults and ADHD patients (Chamberlain et al., 2007, 2009). Moreover, fMRI analysis revealed that atomoxetine exerted its beneficial effects via modulation of rIFG (Chamberlain et al., 2009), providing additional support for the model mechanisms. Finally, this highlights an alternative source for response inhibition deficits observed in PD patients previously linked to DA dysfunction (see Vazey and Aston-Jones (2012) for a review highlighting the importance of aberrant NE signaling in cognitive deficits of PD patients).

2.5 Limitations

Despite our model's success in reproducing and explaining a wide array of data and offering potential solutions for long standing issues in the field, we certainly acknowledge that there are many errors of omission and – although we did not include any biological features that are unsupported by data – perhaps some errors of commission. We note however note that most of our assumptions and simulations are largely orthogonal to each other. Thus, each aspect of the model is falsifiable on its own, without necessarily falsifying other aspects. We discuss a few salient limitations below; it is by no means exhaustive.

2.5.1 Specificity of PFC regions and function

While the BG of our neural network model is fairly concrete and solidly grounded on ample anatomical electrophysiological, and functional evidence, the individual contributions of frontal regions including DLPFC, SEF, pre-SMA, FEF and dACC are not as well established currently. For example, we identified an executive control network in our model consisting of DLPFC, SEF and pre-SMA. The task rules and necessary motor commands to follow them are implemented by hard-coded input and output weight patterns of its extended network (i.e. sensory input, instruction, FEF and striatum). This implementation short-circuits a lot of the computational complexities the biological system has to solve; (i) the executive controller has to selectively retrieve the appropriate rule for the current trial from short or long-term working memory; (ii) integrate the sensory evidence to compute the correct response (e.g. via vector inversion); (iii) compute the necessary motor sequences to perform the correct action; and (iv) identify incorrectly activated prepotent responses and selectively suppress them. While neural network models with a more detailed representation of PFC exist (e.g. O'Reilly and Frank, 2006) in which rule-like representations can develop through experience, how exactly the necessary computations can be implemented dynamically is as-of-today a still unresolved question.

Critically, our focus in this work was on how PFC and BG interact when inhibitory control is required by extending the detailed BG model by Frank (2006). We also account for some electrophysiologcal data in frontal cortex, while acknowledging that there is still some uncertainty in the respective roles of these areas and their interactions which will be open for revision as more data become available.

2.5.2 Learning

Previous BG models explored the role of DA in feedback driven learning (Wiecki and Frank, 2010). As humans (but not monkeys) are able to perform this task without learning, we chose to remain agnostic about the type of learning that takes place prior to performing the task. We thus hard-coded task rules into the model. An additional driving factor is the lack of published reports on specific learning phenomena in the SST and AST.

2.6 Conclusions

We presented a comprehensive, biologically plausible model of global and selective response inhibition which takes known properties of the neuronal underpinnings into account and tries to link them with results from cognitive science, electrophysiology, imaging studies and pharmacological experiments. Here, we showed that augmenting our previously described BG model with the addition of the FEF, DLPFC, and rIFG allows us to simulate control over prepotent responses and to capture a wealth of data in this domain across multiple levels of analysis. We furthermore provide multiple mechanisms that can lead to disruptions in inhibitory control processes and which have implications for interpretation of data from patients with psychiatric disorders such as SZ and ADHD. Our model shows that the observed deficits in inhibitory control paradigms do not necessarily have to reflect dysfunctional response inhibition per se but could be due to other factors such as salience, conflict detection and/or motivation, and related to distinct neural mechanisms.

2.7 Acknowledgments

All modeling was performed by TVW under supervision of MJF. We thank Jeffrey Schall, Gordon Logan, Christopher H. Chatham, James F. Cavanagh, and three anonymous reviewers for helpful comments on an earlier version of the manuscript.

This project was supported by NIMH grant R01MH080066-01 and NSF grant1125788 to MJF, and partially supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract number D10PC20023. The US Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation therein. The views and conclusion contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the US Government.

Chapter 3

HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python

This chapter has been published and reflects contributions of other authors:

Wiecki T. V., Sofer I., & Frank M.J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics* 7:14

3.1 Abstract

The diffusion model is a commonly used tool to infer latent psychological processes underlying decision making, and to link them to neural mechanisms based on response times. Although efficient open source software has been made available to quantitatively fit the model to data, current estimation methods require an abundance of response time measurements to recover meaningful parameters, and only provide point estimates of each parameter. In contrast, hierarchical Bayesian parameter estimation methods are useful for enhancing statistical power, allowing for simultaneous estimation of individual subject parameters and the group distribution that they are drawn from, while also providing measures of uncertainty in these parameters in the posterior distribution. Here, we present a novel Python-based toolbox called HDDM (hierarchical drift diffusion model), which allows fast and flexible estimation of the the drift-diffusion model and the related linear ballistic accumulator model. HDDM requires fewer data per subject / condition than non-hierarchical method, allows for full Bayesian data analysis, and can handle outliers in the data. Finally, HDDM supports the estimation of how trial-by-trial measurements (e.g. fMRI) influence decision making parameters. This paper will first describe the theoretical background of drift-diffusion model and Bayesian inference. We then illustrate usage of the toolbox on a real-world data set from our lab. Finally, parameter recovery studies show that HDDM beats alternative fitting methods like the χ^2-quantile method as well as maximum likelihood estimation. The software and documentation can be downloaded at: http://ski.clps.brown.edu/hddm_docs/

3.2 Introduction

Sequential sampling models (SSMs) (Townsend and Ashby, 1983b) have established themselves as the de-facto standard for modeling response-time data from simple two-alternative forced choice decision making tasks (Smith and Ratcliff, 2004). Each decision is modeled as an accumulation of noisy information indicative of one choice or the other, with sequential evaluation of the accumulated evidence at each time step. Once this evidence crosses a threshold, the corresponding response is executed. This simple assumption about the underlying psychological process has the appealing property of reproducing not only choice probabilities, but the full distribution of response times for each of the two choices. Models of this class have been used successfully in mathematical psychology since the 60s and more recently adopted in cognitive neuroscience investigations. These studies are typically interested in neural mechanisms associated with the accumulation process or for regulating the decision threshold (e.g. Forstmann et al., 2008; Cavanagh et al., 2011; Ratcliff et al., 2009). One issue in such model-based cognitive neuroscience approaches is that the trial numbers in each condition are often low, making it difficult to estimate model parameters. For example, studies with patient populations, especially if combined with intra-operative recordings, typically have substantial constraints on the duration of the task. Similarly, model-based fMRI or EEG studies are often interested not in static model parameters, but how these dynamically vary with trial-by-trial variations in recorded brain activity. Efficient and reliable estimation methods that take advantage of the full statistical structure available in the data across subjects and conditions are critical to the success of these endeavors.

Bayesian data analytic methods are quickly gaining popularity in the cognitive sciences because of their many desirable properties (Lee and Wagenmakers, 2013; Kruschke, 2010). First, Bayesian methods allow inference of the full posterior distribution of each parameter, thus quantifying uncertainty in their estimation, rather than simply provide their most likely value. Second, hierarchical modeling is naturally formulated in a Bayesian framework. Traditionally, psychological models either assume subjects are completely independent of each other, fitting models separately to each individual, or that all subjects are the same, fitting models to the group as if they are all copies of some "average subject". Both approaches are sub-optimal in that the former fails to capitalize on statistical strength offered by the degree to which subjects are similar with respect to one or more model parameters, whereas the latter approach fails to account for the differences among subjects, and hence could lead to a situation where the estimated model cannot fit any individual subject. The same limitations apply to current DDM software packages such as DMAT (Vandekerckhove and Tuerlinckx, 2008) and fast-dm (Voss and Voss, 2007). Hierarchical Bayesian methods provide a remedy for this problem by allowing group and subject parameters to be estimated simultaneously at different hierarchical levels (Lee and Wagenmakers, 2013; Kruschke, 2010; Vandekerckhove et al., 2011). Subject parameters are assumed to be drawn from a group distribution, and to the degree that subjects are similar to each other, the variance in the group distribution will be estimated to be small, which reciprocally has a greater influence on constraining parameter estimates of any individual. Even in this scenario, the method still allows the posterior for any given individual subject to differ substantially from that of the rest of the group given sufficient data to overwhelm the group prior. Thus the method capitalizes on statistical strength shared across the individuals, and can do so to different degrees even within the same sample and model, depending on the extent to which subjects are similar to each other in one parameter vs. another. In the DDM for example, it may be the case that there is relatively little variability across subjects in the perceptual time for stimulus encoding, quantified by the "non-decision time" but more variability in their degree of response caution, quantified by the "decision threshold". The estimation should be able to capitalize on this structure so that the non-decision time in any given subject is anchored by that of the group, potentially allowing for more efficient estimation of that subject's decision threshold. This approach may be particularly helpful when relatively few trials per condition are available for each subject, and when incorporating noisy trial-by-trial neural data into the estimation of DDM parameters.

HDDM is an open-source software package written in Python which allows (i) the flexible construction of hierarchical Bayesian drift diffusion models and (ii) the estimation of its posterior parameter distributions via PyMC (Patil et al., 2010). User-defined models can be created via a simple Python script or be used interactively via, for example, the IPython interpreter shell (Pérez and Granger, 2007). All run-time critical functions are coded in Cython (Behnel et al., 2011) and compiled natively for speed which allows estimation of complex models in minutes. HDDM includes many commonly used statistics and plotting functionality generally used to assess model fit. The code is released under the permissive BSD 3-clause license, test-covered to assure correct behavior and well documented. An active mailing list exists to facilitate community interaction and help users. Finally, HDDM allows flexible estimation of trial-by-trial regressions where an external measurement (e.g. brain activity as measured by fMRI) is correlated with one or more decision making parameters.

This report is intended to familiarize experimentalists with the usage and benefits of HDDM. The purpose of this report is thus two-fold; (i) we briefly introduce the toolbox and provide a tutorial on a real-world data set (a more comprehensive description of all the features can be found online); and (ii) characterize its success in recovering model parameters by performing a parameter recovery study using simulated data to compare the hierarchical model used in HDDM to non-hierarchical or non-Bayesian methods as a function of the number of subjects and trials. We show that it outperforms these other methods and has greater power to detect dependencies of model parameters on other measures such as brain activity, when such relationships are present in the data. These simulation results can also inform experimental design by showing minimum number of trials and subjects to achieve a desired level of precision.

3.3 Methods

3.3.1 Drift Diffusion Model

SSMs generally fall into one of two classes: (i) diffusion models which assume that *relative* evidence is accumulated over time and (ii) race models which assume independent evidence accumulation and response commitment once the first accumulator crossed a boundary (LaBerge, 1962; Vickers, 1970). Currently, HDDM includes two of the most commonly used SSMs: the drift diffusion model (DDM) (Ratcliff and Rouder, 1998; Ratcliff and McKoon, 2008) belonging to the class of diffusion models and the linear ballistic accumulator (LBA) (Brown and Heathcote, 2008) belonging to the class of race models. In the remainder of this paper we focus on the more commonly used DDM.

As input these methods require trial-by-trial RT and choice data (HDDM currently only supports binary decisions) as illustrated in the below example table:

RT	response	condition	brain measure
0.8	1	hard	0.01
1.2	0	easy	0.23
0.25	1	hard	-0.3

The DDM models decision making in two-choice tasks. Each choice is represented as an upper and lower boundary. A drift-process accumulates evidence over time until it crosses one of the two boundaries and initiates the corresponding response (Ratcliff and Rouder, 1998; Smith and Ratcliff, 2004) (see figure 3.1 for an illustration). The speed with which the accumulation process approaches one of the two boundaries is called drift-rate v. Because there is noise in the drift process, the time of the boundary crossing and the selected response will vary between trials. The distance between the two boundaries (i.e. threshold a) influences how much evidence must be accumulated until a response is executed. A lower threshold makes responding faster in general but increases the influence of noise on decision making and can hence lead to errors or impulsive choice, whereas a higher threshold leads to more cautious responding (slower, more skewed RT distributions, but more accurate). Response time, however, is not solely comprised of the decision making process – perception, movement initiation and execution all take time and are lumped in the DDM by a single non-decision time parameter t. The model also allows for a prepotent bias z affecting the starting point of the drift process relative to the two boundaries. The termination times of this generative process gives rise to the response time distributions of both choices.

An analytic solution to the resulting probability distribution of the termination times

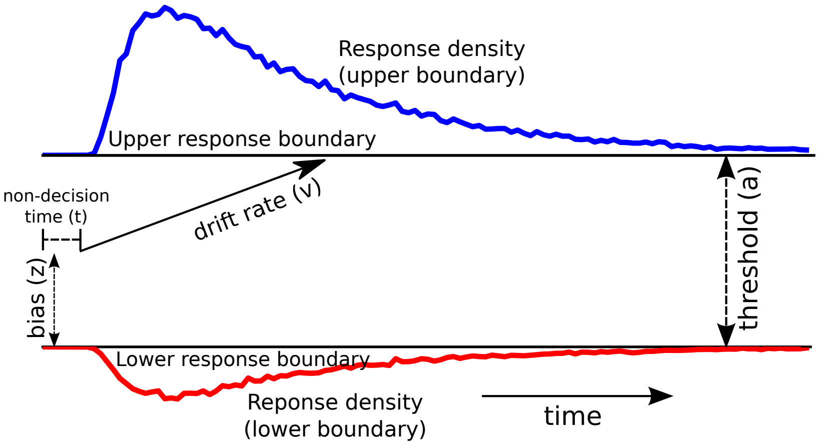

Figure 3.1: Trajectories of multiple drift-processes (blue and red lines, middle panel). Evidence is noisily accumulated over time (x-axis) with average drift-rate v until one of two boundaries (separated by threshold a) is crossed and a response is initiated. Upper (blue) and lower (red) panels contain density plots over boundary-crossing-times for two possible responses. The flat line in the beginning of the drift-processes denotes the non-decision time t where no accumulation happens. The histogram shapes match closely to those observed in response time measurements of research participants. Note that HDDM uses a closed-form likelihood function and not actual simulation as depicted here.

was provided by Wald (1947); Feller (1968):

$$f(x|v,a,z) = \frac{\pi}{a^2} \exp\left(-vaz - \frac{v^2\,x}{2}\right) \times \sum_{k=1}^{\infty} k \exp\left(-\frac{k^2\pi^2 x}{2a^2}\right) \sin\left(k\pi z\right)$$

Since the formula contains an infinite sum, HDDM uses an approximation provided by (Navarro and Fuss, 2009).

Subsequently, the DDM was extended to include additional noise parameters capturing inter-trial variability in the drift-rate, the non-decision time and the starting point in order to account for two phenomena observed in decision making tasks, most notably cases where errors are faster or slower than correct responses. Models that take this into account are referred to as the full DDM (Ratcliff and Rouder, 1998). HDDM uses analytic integration of the likelihood function for variability in drift-rate and numerical integration for variability in non-decision time and bias (Ratcliff and Tuerlinckx, 2002).

3.3.2 Hierarchical Bayesian Estimation of the Drift-Diffusion Model

Statistics and machine learning have developed efficient and versatile Bayesian methods to solve various inference problems (Poirier, 2006a). More recently, they have seen wider adoption in applied fields such as genetics (Stephens and Balding, 2009b) and psychology (Clemens et al., 2011b). One reason for this Bayesian revolution is the ability to quantify the certainty one has in a particular estimation of a model parameter. Moreover, hierarchical Bayesian models provide an elegant solution to the problem of estimating parameters of individual subjects and groups of subjects, as outlined above. Under the assumption that participants within each group are similar to each other, but not identical, a hierarchical model can be constructed where individual parameter estimates are constrained by group-level distributions (Nilsson et al., 2011b; Shiffrin et al., 2008b).

HDDM includes several hierarchical Bayesian model formulations for the DDM and LBA. For illustrative purposes we present the graphical model depiction of a hierarchical DDM with informative priors and group-only inter-trial variability parameters in figure 3.2. Note, however, that there is also a model with non-informative priors which the user can opt to use. Nevertheless, we recommend using informative priors as they constrain parameter estimates to be in the range of plausible values based on past literature (Matzke and Wagenmakers, 2009) (see the supplement), which can aid in reducing issues with parameter collinearity, and leads to better recovery of true parameters in simulation studies – especially with few trials as shown below.

Graphical nodes are distributed as follows:

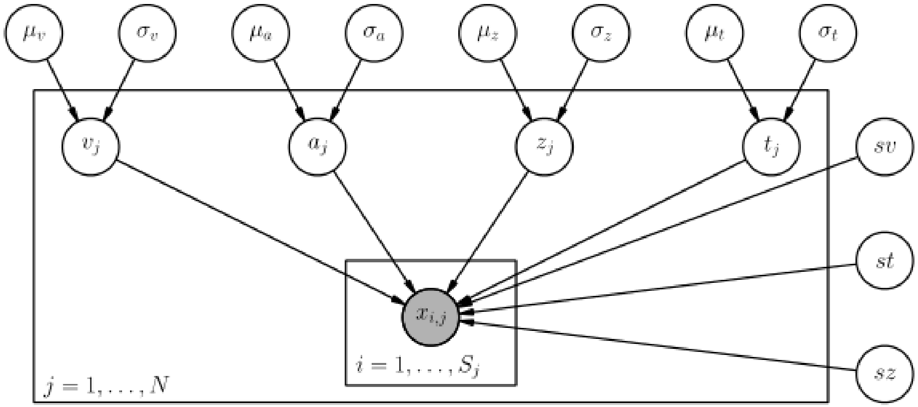

Figure 3.2: Basic graphical hierarchical model implemented by HDDM for estimation of the drift-diffusion model. Round nodes represent random variables. Shaded nodes represent observed data. Directed arrows from parents to children visualize that parameters of the child random variable are distributed according to its parents. Plates denote that multiple random variables with the same parents and children exist. The outer plate is over subjects while the inner plate is over trials.

$$\begin{array}{l|l|l}
\mu_a \sim \mathcal{G}(1.5, 0.75) & \sigma_a \sim \mathcal{HN}(0.1) & a_j \sim \mathcal{G}(\mu_a, \sigma_a^2) \\
\mu_v \sim \mathcal{N}(2, 3) & \sigma_v \sim \mathcal{HN}(2) & v_j \sim \mathcal{N}(\mu_v, \sigma_v^2) \\
\mu_z \sim \mathcal{N}(0.5, 0.5) & \sigma_z \sim \mathcal{HN}(0.05) & z_j \sim \text{invlogit}(\mathcal{N}(\mu_z, \sigma_z^2)) \\
\mu_t \sim \mathcal{G}(0.4, 0.2) & \sigma_t \sim \mathcal{HN}(1) & t_j \sim \mathcal{N}(\mu_t, \sigma_t^2) \\
sv \sim \mathcal{HN}(2) & st \sim \mathcal{HN}(0.3) & sz \sim \mathcal{B}(1, 3)
\end{array}$$

and $x_{i,j} \sim F(a_i, z_i, v_i, t_i, sv, st, sz)$ where $x_{i,j}$ represents the observed data consisting of response time and choice of subject i on trial j and F represents the DDM likelihood function as formulated by Navarro and Fuss (2009). $\mathcal{N}$ represents a normal distribution parameterized by mean and standard deviation, $\mathcal{HN}$ represents a positive-only, half-normal parameterized by standard-deviation, $\mathcal{G}$ represents a Gamma distribution parameterized by mean and rate, $\mathcal{B}$ represents a Beta distribution parameterized by α and β. Note that in this model we do not attempt to estimate individual parameters for inter-trial variabilities. The reason is that the influence of these parameters onto the likelihood is often so small that very large amounts of data would be required to make meaningful inference at the individual level.

HDDM then uses Markov chain Monte Carlo (MCMC) (Gamerman and Lopes, 2006) to estimate the joint posterior distribution of all model parameters (for more information on hierarchical Bayesian estimation we refer to the supplement).

Note that the exact form of the model will be user-dependent; consider as an example a model where separate drift-rates v are estimated for two conditions in an experiment: easy and hard. In this case, HDDM will create a hierarchical model with group parameters $\mu_{v_{\text{easy}}}$, $\sigma_{v_{\text{easy}}}$, $\mu_{v_{\text{hard}}}$, $\sigma_{v_{\text{hard}}}$,and individual subject parameters $v_{j_{\text{easy}}}$, and $v_{j_{\text{hard}}}$.

3.4 Results

In the following we will demonstrate how HDDM can be used to infer different components of the decision making process in a reward-based learning task. While demonstrating core features this is by no means a complete overview of all the functionality in HDDM. For more information, an online tutorial and a reference manual see http://ski.clps.brown.edu/hddm_docs.

Python requires modules to be imported before they can be used. The following code imports the `hddm` module into the Python name-space:

```
import hddm
```

3.4.1 Loading data

It is recommended to store your trial-by-trial response time and choice data in a csv (comma-separated-value, see below for exact specifications) file. In this example we will be using data collected in a reward-based decision making experiment in our lab (Cavanagh et al., 2011). In brief, at each trial subjects choose between two symbols. The trials were divided into win-win trials (WW), in which the two symbols were associated with high winning chances; lose-lose trials (LL), in which the symbols were associated with low winning chances, and win-lose trials (WL), which are the easiest because only one symbol was associated with high winning chances. Thus WW and LL decisions together comprise high conflict (HC) trials (although there are other differences between them, we do not focus on those here), whereas WL decisions are low conflict (LC). The main hypothesis of the study was that high conflict trials induce an increase in the decision threshold, and that the mechanism for this threshold modulation depends on communication between mediofrontal cortex (which exhibits increased activity under conditions of choice uncertainty or conflict) and the subthalamic nucleus (STN) of the basal ganglia (which provides a temporary brake on response selection by increasing the decision threshold). The details of this mechanism are described in other modeling papers (e.g. Ratcliff and Frank, 2012). Cavanagh et al. (2011) tested this theory by measuring EEG activity over mid-frontal cortex, focusing on the theta band, given prior associations with conflict, and testing whether trial-to-trial variations in frontal theta were related to adjustments in decision threshold during high conflict trials. They tested the STN component of the theory by administering the same experiment to patients who had deep brain stimulation (DBS) of the STN, which interferes with normal processing and was tested in the on and off condition.

The first ten lines of the data file look as follows.

```
subj_idx,stim,rt,response,theta,dbs,conf
0,LL,1.21,1.0,0.65,1,HC
0,WL,1.62,1.0,-0.327,1,LC
0,WW,1.03,1.0,-0.480,1,HC
0,WL,2.77,1.0,1.927,1,LC
0,WW,1.13,0.0,-0.2132,1,HC
0,WL,1.14,1.0,-0.4362,1,LC
0,LL,2.0,1.0,-0.27447,1,HC
0,WL,1.04,0.0,0.666,1,LC
0,WW,0.856,1.0,0.1186,1,HC
```

The first row represents the column names; each following row corresponds to values associated with a column on an individual trial. While `subj_idx` (unique subject identifier), `rt` (response time) and `response` (binary choice) are required, additional columns can represent experiment specific data. Here, `theta` represents theta power as measured by EEG, `dbs` whether DBS was turned on or off, `stim` which stimulus type was presented and `conf` the conflict level of the stimulus (see above).

The hddm.load_csv() function can then be used to load this file.

```
data = hddm.load_csv('hddm_demo.csv')
```

3.4.2 Fitting a hierarchical model

The HDDM class constructs a hierarchical DDM that can later be fit to subjects' RT and choice data, as loaded above. By supplying no extra arguments other than data, HDDM constructs a simple model that does not take our different conditions into account. To speed up convergence, the starting point is set to the maximum a-posterior value (MAP) by calling the HDDM.find_starting_values method which uses gradient ascent optimization. The HDDM.sample() method then performs Bayesian inference by drawing posterior samples using the MCMC algorithm.

```
# Instantiate model object passing it our data.
# This will tailor an individual hierarchical DDM around the dataset.
m = hddm.HDDM(data)
# find a good starting point which helps with the convergence.
m.find_starting_values()
# start drawing 2000 samples and discarding 20 asburn-in
m.sample(2000, burn=20)
```

We recommend drawing between 2000 and 10000 posterior samples, depending on the convergence. Discarding the first 20-1000 samples as burn-in is often enough in our experience. Auto-correlation of the samples can be reduced by adding the thin=n keyword to sample() which only keeps every n-th sample, but unless memory is an issue we recommend keeping all samples and instead drawing more samples if auto-correlation is high.

Note that it is also possible to fit a non-hierarchical model to an individual subject by setting is_group_model=False in the instantiation of HDDM or by passing in data

Figure 3.3: Posterior plots for the group mean (left half) and group standard-deviation (right half) of the threshold parameter a. Posterior trace (upper left inlay), autocorrelation (lower left inlay), and marginal posterior histogram (right inlay; solid black line denotes posterior mean and dotted black line denotes 2.5% and 97.5% percentiles).

which lacks a `subj_idx` column. In this case, HDDM will use the group-mean priors from above for the DDM parameters.

The inference algorithm, MCMC, requires the chains of the model to have properly converged. While there is no way to guarantee convergence for a finite set of samples in MCMC, there are many heuristics that allow identification of problems of convergence. One analysis to perform is to visually investigate the trace, the autocorrelation, and the marginal posterior. These can be plotted using the `HDDM.plot_posteriors()` method (see figure 3.3). For the sake of brevity we only plot two here (group mean and standard deviation of threshold). In practice, however, one should examine all of them.

```
m.plot_posteriors(['a','a_var'])
```

Problematic patterns in the trace would be drifts or large jumps which are absent here. The autocorrelation should also drop to zero rather quickly (i.e. well smaller than 50) when considering the influence of past samples , as is the case here.

The Gelman-Rubin $\hat{R}$ statistic (Gelman and Rubin, 1992) provides a more formal test for convergence that compares within-chain and between-chain variance of different runs of the same model. This statistic will be close to 1 if the samples of the different chains are indistinguishable. The following code demonstrates how 5 models can be run in a for-loop and stored in a list (here called `models`).

```
models = list()
for i in range(5):
    m = hddm.HDDM(data)
    m.find_starting_values()
    m.sample(5000, burn=20)
    models.append(m)

hddm.analyze.gelman_rubin(models)
```

Which produces the following output (abridged to preserve space):

```
{'a': 1.000,
 'a_std': 1.001,
 't': 1.000}
```

Values should be close to 1 and not larger than 1.02 which would indicate convergence problems.

Once convinced that the chains have properly converged we can analyze the posterior values. The HDDM.print_stats() method outputs a table of summary statistics for each parameters' posterior).

```
m.print_stats()
```

```
              mean       std      2.5q       25q       50q       75q 97.5q

a         2.058015  0.102570  1.862412  1.988854  2.055198  2.123046 2.261410
a_var     0.379303  0.089571  0.244837  0.316507  0.367191  0.426531 0.591643
a_subj.0  2.384066  0.059244  2.274352  2.340795  2.384700  2.423012 2.500647
```

The output contains various summary statistics describing the posterior of each parameter: group mean parameter for threshold a, group variability a_var and individual subject parameters a_subj.0. Other parameters are not shown here for brevity but would be outputted normally.

As noted above, this model did not take the different conditions into account. To test whether the different conflict conditions affect drift-rate we create a new model which estimates separate drift-rate `v` for the three conflict conditions. HDDM supports splitting by condition in a between-subject manner via the `depends_on` keyword argument supplied to the `HDDM` class. This argument expects a Python `dict` which maps the parameter to be split to the column name containing the conditions we want to split by. This way of defining parameters to be split by condition is directly inspired by the fast-dm toolbox (Voss and Voss, 2007).

```
m_stim = hddm.HDDM(data, depends_on={’v’: ’stim’})
m_stim.find_starting_values()
m_stim.sample(2000, burn=20)
```

Note that while every subject was tested on each condition in this case, this is not a requirement. The `depends_on` keyword can also be used to test between-group differences. For example, if we collected data where one group received a drug and the other one a placebo we would include a column in the data labeled ’drug’ that contained ’drug’ or ’placebo’ for each subject. In our model specification we could test the hypothesis that the drug affects threshold by specifying `depends_on={’a’: ’drug’}`. In this case `HDDM` would create and estimate separate group distributions for the two groups/conditions. After selecting an appropriate model (e.g. via model selection) we could compare the two group mean posteriors to test whether the drug is effective or not.

We next turn to comparing the posterior for the different drift-rate conditions. To plot the different traces we need to access the underlying node object. These are stored inside the `nodes_db` attribute which is a table (specifically, a `DataFrame` object as provided by the `Pandas` Python module) containing a row for each model parameter (e.g. `v(WW)`) and multiple columns containing various information about that parameter (e.g. the mean, or the node object). The `node` column used here represents the `PyMC` node object. Multiple assignment is then used to assign the 3

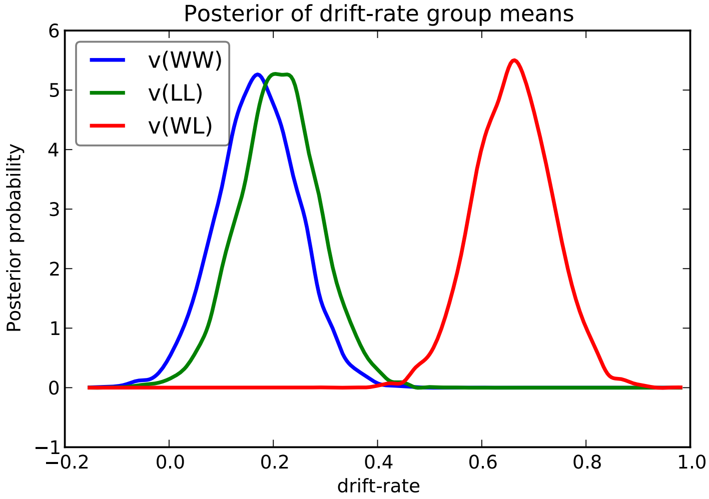

Figure 3.4: Posterior density plot of the group means of the 3 different drift-rates v as produced by the `hddm.analyze.plot_posterior_nodes()` function. Regions of high probability are more credible than those of low probability.

drift-rate nodes to separate variables. The `hddm.analyze.plot_posterior_nodes()` function takes a list of `PyMC` nodes and plots the density by interpolating the posterior histogram (see figure 3.4).

```
v_WW, v_LL, v_WL = m_stim.nodes_db.node[['v(WW)', 'v(LL)', 'v(WL)']]
hddm.analyze.plot_posterior_nodes([v_WW, v_LL, v_WL])
```

Based on figure 3.4 we might reason that the `WL` condition drift-rate is substantially greater than that for the other two conditions, which are fairly similar to each other.

One benefit of estimating the model in a Bayesian framework is that we can do significance testing directly on the posterior rather than relying on frequentist statistics (Lindley, 1965) (see also Kruschke (2010) for many examples of the advantages of this approach). For example, we might be interested in whether the drift-rate for `WW` is larger than that for `LL`, or whether drift-rate for `LL` is larger than `WL`. The below code computes the proportion of the posteriors in which the drift rate for one condition is greater than the other. It can be seen that the posteriors for `LL` do not overlap at all for `WL`, and thus the probability that `LL` is greater than `WL` should be near zero.

```
print "P(WW > LL) = ", (v_WW.trace() > v_LL.trace()).mean()
print "P(LL > WL) = ", (v_LL.trace() > v_WL.trace()).mean()
```

Which produces the following output.

```
P(WW > LL) =  0.34696969697
P(LL > WL) =  0.0
```

In addition to computing the overlap of the posterior distributions we can compare whether the added complexity of models with additional degrees of freedom is justified to account for the data using model selection. The deviance information criterion (Spiegelhalter et al., 2002a) (DIC; lower is better) is a common method for assessing model fit in hierarchical models. The DIC is known to be somewhat biased in selecting the model with greater complexity, although alternative forms exist which improve this issue (see Plummer, 2008). Nevertheless, DIC can be a useful metric with this caveat in mind. One suggested approach is to generate simulated data from alternative models and use DIC to determine whether it properly selects the correct model given the same task contingencies. This exercise can help determine whether to rely on DIC, and also to provide an expected quantitative difference in DIC scores between models if one of them was correct, as a benchmark to compare DIC differences for fits to real data. We recommend interpreting significant differences in parameter estimates only within the models that fit the data the best penalized for complexity. By accessing the `dic` attribute of the model objects we can print the model comparison measure:

```
print "Lumped model DIC: %f" % m.dic
print "Stimulus model DIC: %f" % m_stim.dic
```

Which produces the following output:

```
Lumped model DIC: 10960.570932

Stimulus model DIC: 10775.615192
```

Based on the lower DIC score for the model allowing drift-rate to vary by stimulus condition we might conclude that it provides better fit than the model which forces the drift-rates to be equal, despite the increased complexity.

Note that Bayesian hypothesis testing and model comparison are areas of active research. One alternative to analyzing the posterior directly and the DIC score is the Bayes Factor (e.g. Wagenmakers et al., 2010).

3.4.3 Fitting regression models

As mentioned above, cognitive neuroscience has embraced the DDM as it enables to link psychological processes to cognitive brain measures. The Cavanagh et al. (2011) study provides a useful illustration of the functionality. EEG recordings provided a trial-ty-trial measure of brain activity (frontal theta), and it was found that this activity correlated with increases in decision threshold in high conflict trials. Note that the data set and results exhibit more features than we consider here for the time being (specifically the manipulation of deep brain stimulation), but for illustrative purposes, we show only the code here to reproduce the main theta-threshold relationship in a model restricted to participants without brain stimulation. For more information, see Cavanagh et al. (2011).

The `HDDMRegressor` class allows individual parameters to be described by a linear model specification. In addition to the data argument, `HDDMRegressor` expects a linear model descriptor string to be provided. This descriptor contains the `outcome` variable that should be replaced with the output of the linear model – in this case `a`. The expression `theta:C(stim)` specifies an interaction between theta power and stimulus. The `C()` specifies that the `stim` column contains categorical data and will result in WL, LL, and WW being dummy coded. The Treatment argument encodes which condition should be used as the intercept. The two other conditions – LL and WW – will then be expressed *relative* to WL. For more information about the linear model specification syntax we refer to the Patsy documentation. In summary, by selecting data from the dbs off condition and specifying a linear model that uses categorical dummy-coding we can estimate a within-subject effect of theta power on threshold in different conditions.

```
m_reg = hddm.HDDMRegressor(data[data.dbs == 0], "a ~theta:C(conf, Treatment('LC'))",
 depends_on={'v': 'stim'})
```

Which produces the following output:

```
Adding these covariates:
['a_Intercept', "a_theta:C(conf, Treatment('LC'))[HC]",
"a_theta:C(conf, Treatment('LC'))[LC]"]
```

Instead of estimating one static threshold per subject across trials, this model assumes the threshold to vary on each trial according to the linear model specified above (as a function of their measured theta activity). Cavanagh et al. (2011) illustrates that this brain/behavior relationship differs as a function of whether patients are on or off STN deep brain stimulation, as hypothesized by the model that STN is responsible for increasing the decision threshold when cortical theta rises).

As noted above, this experiment also tested patients on deep brain stimulation (DBS). Figure 3.5 shows the regression coefficient of theta on threshold when the above model is estimated in the DBS off condition (in blue) and the DBS on condition (in green; code to estimate not shown). As can be seen, the influence of theta on threshold reverses. This exercise thus shows that HDDM can be used both to assess the influence of trial-by-trial brain measures on DDM parameters, but also how parameters vary when brain state is manipulated.

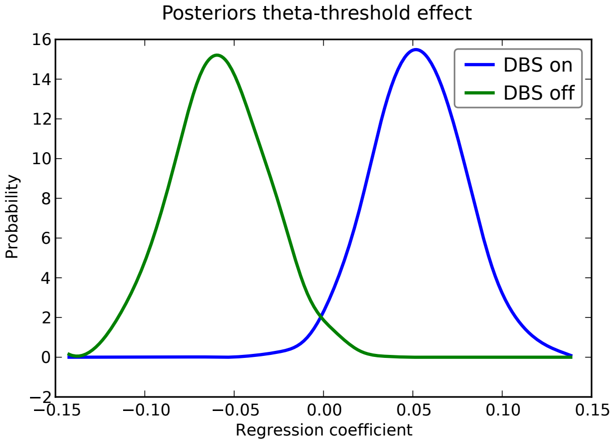

Figure 3.5: Posterior density of the group theta regression coefficients on threshold a when DBS is turned on (blue) and off (green).

Finally, HDDM also supports modeling of within-subject effects as well as robustness to outliers. Descriptions and usage instructions of which can be found in the supplement.

3.5 Simulations

To quantify the quality of the fit of our hierarchical Bayesian method we ran three simulation experiments. All code to replicate the simulation experiments can be found online at https://github.com/hddm-devs/HDDM-paper.

3.5.1 Experiment 1 and 2 setup

For the first and second experiments, we simulated an experiment with two drift-rates (v_1 and v_2), and asked what the likelihood of detecting a drift rate difference is using each method. For the first experiment, we fixed the number of subjects at 12 (arbitrarily chosen), while manipulating the number of trials (20, 30, 40, 50, 75, 100, 150). For the second experiment, we fixed the number of trials at 75 (arbitrary chosen), while manipulating the number of subjects (8, 12, 16, 20, 24, 28).

For each experiment and each manipulated factor (subjects, trials), we generated 30 multi-subject data-sets by randomly sampling group parameters. For the first and second experiment, the group parameters were sampled from a uniform distribution $[v_1 \sim \mathcal{U}(0.1, 0.5), a \sim \mathcal{U}(0.5, 0.2), t \sim \mathcal{U}(0.2, 0.5), sv \sim \mathcal{U}(0, 2.5)]$, sz and st were set to zero, and v_2 was set to $2 * v_1$. To generate individual subject parameters, zero centered normally distributed noise was added to v_1, a, t, and sv, with standard deviation of 0.2, 0.2, 0.1, and 0.1 respectively. The noise of v_2 was identical to that of v_1.

We compared four methods: (i) the hierarchical Bayesian model presented above with a within subject effect (HB); (ii) a non-hierarchical Bayesian model, which estimates each subject individually (nHB); (iii) the χ^2-Quantile method on individual subjects (Ratcliff and Tuerlinckx, 2002); and (iv) maximum likelihood (ML) estimation using the Navarro and Fuss (2009) likelihood on individual subjects.

To investigate the difference in parameter recovery between the methods, we computed the mean absolute error of the recovery for each parameter and method in the trials experiment (we also computed this for the subjects experiment but results are qualitatively similar and omitted for brevity). We excluded the largest errors (5%) from our calculation for each method to avoid cases where unrealistic parameters were recovered (this happened only for ML and the quantiles method).

For each dataset and estimation method in the subject experiment we computed whether the drift-rate difference was detected (we also computed this for the trials experiment but results are qualitatively similar and omitted for brevity). For the non-hierarchical methods (ML, quantiles, nHB), a difference is detected if a paired t-test found a significant difference between the two drift rate of the individuals (p ¡ .05). For HB, we used Bayesian parameter estimation (Lindley, 1965; Kruschke, 2010). Specifically, we computed the 2.5 and 97.5 quantiles of the posterior of the group variable that models the difference between the two drift rates. An effect is detected if zero fell outside the quantiles. The detection likelihood for a given factor manipulation and estimation method was defined as the number of times an effect was detected divided by the total number of experiments.

3.5.2 Experiment 3 setup

In the third experiment, we investigated the detection likelihood of trial-by-trial effects of a given covariate (e.g. a brain measure) on the drift-rate. We fixed the number of subjects at 12, and manipulated both the covariate effect-size (0.1, 0.3, 0.5) and the number of trials (20, 30, 40, 50, 75, 100, 150). To generate data, we first sample an auxiliary variable, α_i from $\mathcal{N}(1, 0.1)$ for each subject i. We then sampled a drift-rate for each subject and each trial from $\mathcal{N}(\alpha_i, 1)$. The drift rate of each subject was set to be correlated to a standard normally distributed covariate (i.e. we generated correlated covariate data) according to the tested effect size. The rest of the variables were sampled as in the first experiments.

We compared all previous methods except the quantiles method, which cannot be used to estimate trial-by-trial effects. For the non-hierarchical methods (ML, quantiles, nHB), an effect is detected if a one sample t-test finds the covariate to be significantly different than zero ($p < .05$). For the HB estimation, we computed the 2.5 and 97.5 quantiles of the posterior of the group covariate variable. If zero fell outside the quantiles, then an effect was detected.

3.5.3 Results

The detection likelihood results for the first experiment are very similar to the results of the second experiment, and were omitted for the sake of brevity. The HB method had the lowest recovery error and highest likelihood of detection in all experiments (figure 3.6, 3.7, 3.8). The results clearly demonstrates the increased power the hierarchical model has over non-hierarchical ones. To validate that the increase in detection

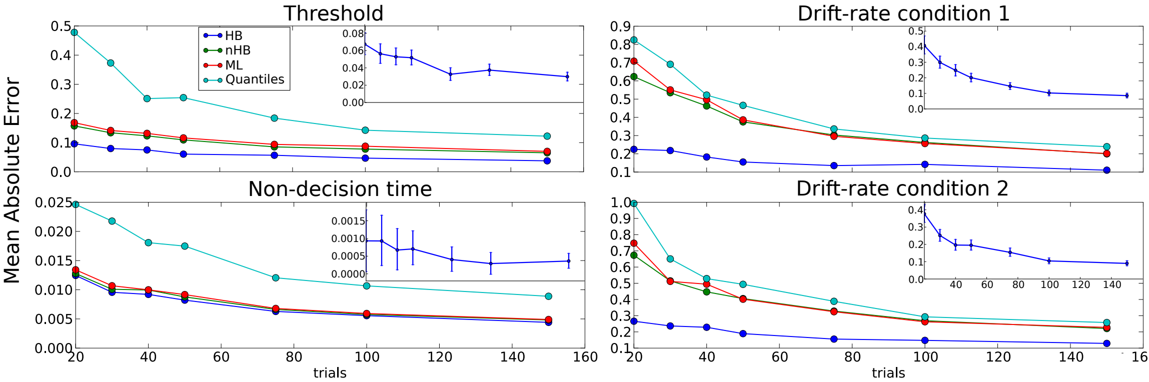

Figure 3.6: Trials experiment. Trimmed mean absolute error (MAE, after removing the 2.5 and 9.75 percentiles) as a function of trial number for each DDM parameter. Colors code for the different estimation methods (HB=Hierarchical Bayes, nHB=non-hierarchical Bayes, ML=maximum likelihood, and Quantiles=χ^2-Quantile method). The inlay in the upper right corner of each subplot plots the difference of the MAEs between HB and ML, and the error-bars represent 95% confidence interval. HB provides a statistically significantly better parameter recovery than ML when the lower end of the error bar is above zero (as it is in each case, with largest effects on drift rate with few trials).

rate is not due to the different statistical test (Bayesian hypothesis testing compared to t-testing), but rather due to the hierarchical model itself, we also applied a t-test to the HB method. The likelihood of detection increased dramatically, which shows that the Bayesian hypothesis testing is not the source of the increase. However, the t-test results were omitted since the independence assumption of the test does not hold for parameters that are estimated using a hierarchical model.

The differences between the hierarchical and non-hierarchical methods in parameters recovery are mainly noticeable for the decision threshold and the two drift rates for every number of trials we tested, and it is most profound when the number of trials is very small (figure 3.6). To verify that the HB method is significantly better than the other methods we chose to directly compare the recovery error achieved by the method in each single recovery to the recovery error achieved by the other methods for the same set dataset (inlay). For clarity purposes, we show only the comparison of HB with ML. The results clearly show that under all conditions HB outperforms the other methods.

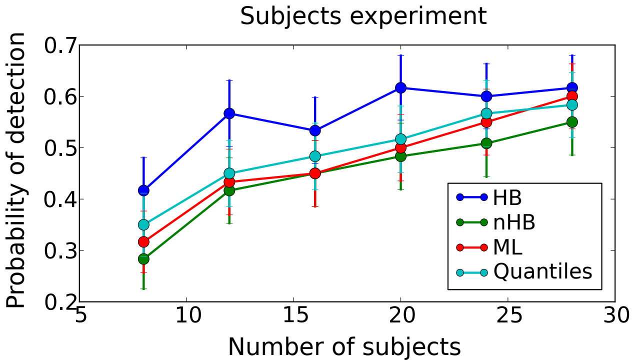

Figure 3.7: Subjects experiment: Probability of detecting a drift-rate difference (y-axis) for different numbers of subjects (x-axis) and different estimation methods (color coded; HB=Hierarchical Bayes, nHB=non-hierarchical Bayes, ML=maximum likelihood, and Quantiles=χ^2-Quantile method). HB together with Bayesian hypothesis testing on the group posterior results in a consistently higher probability of detecting an effect.

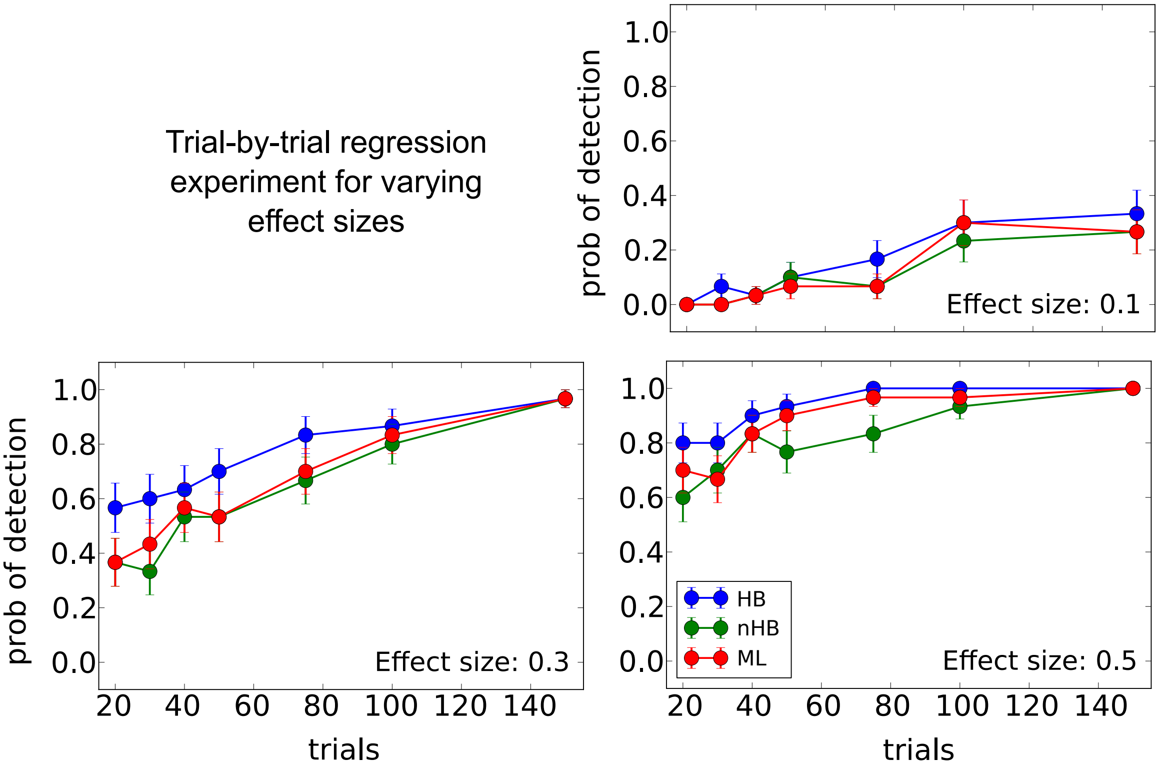

Figure 3.8: Trial-by-trial covariate experiment: Probability of detecting a trial-by-trial effect on drift-rate (y-axis) with effect-sizes 0.1 (top plot), 0.3 (middle plot) and 0.5 (bottom plot) for different estimation methods (color coded; HB=Hierarchical Bayes, nHB=non-hierarchical Bayes, ML=maximum likelihood). While there is only a modest increase in detection rate with the smallest effect size, HB provides an increase in detection rate of up to 20% with larger effect sizes and fewer trials.

3.6 Discussion

Using data from our lab on a reward-based learning and decision making task (Cavanagh et al., 2011) we demonstrate how HDDM can successfully be used to estimate differences in information processing based solely on RT and choice data. By using the HDDMRegression model we are able to not only quantify latent decision making processes in individuals but also how these latent processes relate to brain measures (here theta power as measured by EEG had a positive effect on threshold) on a trial-by-trial basis. Critically, changing brain state via DBS revealed that the effect of theta power on threshold was reversed. As these trial-by-trial effects are often quite noisy, our hierarchical Bayesian approach facilitated the detection of this effect as demonstrated by our simulation studies (figure 3.8), due to shared statistical structure among subjects in determining model parameters. This analysis is more informative than a straight behavioral relationship between brain activity and RT or accuracy alone. While we used EEG to measure brain activity this method should be easily extendable towards other techniques like fMRI (e.g. van Maanen et al., 2011). While trial-by-trial BOLD responses from an event-related study design are often very noisy, initial results in our lab were promising with this approach.

In a set of simulation studies we demonstrate that the hierarchical model estimation used in HDDM can recover parameters better than the commonly used alternatives (i.e. maximum likelihood and χ^2-Quantile estimation). This benefit is largest with few number of trials (figure 3.6) where the hierarchical model structure provides most constraint on individual subject parameter estimation. To provide a more applicable measure we also compared the probability of detecting a drift-rate and trial-by-trial effect and show favorable detection probability.

In conclusion, HDDM is a novel tool that allows researchers to study the neurocognitive implementations of psychological decision making processes. The hierarchical modeling provides power to detect even small correlations between brain activity and decision making processes. Bayesian estimation supports the recent paradigm shift away from frequentist statistics for hypothesis testing (Lindley, 1965; Kruschke, 2010; Lee and Wagenmakers, 2013).

3.7 Acknowledgements

The authors are thankful to Guido Biele, ystein Sandvik and Eric-Jan Wagenmakers for useful feedback and/or code contributions. This work was supported by NIMH Grant RO1 MH080066-01 and NSF Grant #1125788.

Chapter 4

Bridging the gap: Relating biological and psychological models of the response inhibition

4.1 Introduction

In Wiecki et al. (a) (see chapter 2 above) we introduced a neural circuit model informed by behavioral and electrophysiological data collected on various response inhibition paradigms. The neural dynamics explicitly simulated in this model allowed us to accurately map known aspects of the neuroanatomy and recover key electrophysiological patterns. While these neural networks allow us to model the underlying neurobiology more accurately, their complexity and overwhelming number of parameters prohibit the use of quantitative measures to fit them directly to behavioral data. Psychological process models are agnostic about the underlying neurobiology and instead model behavior at the cognitive level. The benefit of model of this type is that they have fewer parameters and can be directly fit to behavior.

Here, we further develop a higher level description of the associated processes based on a combination of the drift diffusion model (DDM) (Ratcliff and McKoon, 2008; Smith and Ratcliff, 2004) and the Noorani and Carpenter (2012) antisaccade LATER model. We call this model the selective inhibition DDM, or SIDDM. This model can be viewed as summarizing the computations performed by the neural network model from chapter 2. To establish this link, we fit the observed behavioral outputs (error rates, response time distributions) produced by the more complex neural model from chapter 2 with the SIDDM model. Our results show that while certain biological manipulations impact dissociable SIDDM parameters, other manipulations impact the same parameter albeit with different underlying objective functions for regulating this parameter. For example, there are multiple biological mechanisms that affect the decision threshold (e.g. frontostriatal connectivity, dopamine, fronto-subthalamic connectivity), but these mechanisms are themselves differentially impacted by motivation, decision conflict, and the speed-accuracy trade-off. This simulation exercise allows us to formulate predictions about the consequences of specific biological manipulations on estimated SIDDM parameters and associated error rates and RT distributions. An ultimate goal of this line of work would allow inverse inference of neurobiological factors underlying psychiatric disease based on patterns of behavior when appropriately probed with diagnostic task manipulations and assessed with quantitative modeling.

4.2 Methods

Given the complexity of the neural model, we tested how variations of certain biological processes can be explained in more functional terms using a minimalist model of cognitive control. The model consists of various interacting Wald accumulators (Schwarz, 2001). The accumulators belong to the class of sequential sampling, or rise-to-threshold models which simulate decision making as a noisy drift-process that accumulates evidence over time. In our case, this drift-process is contained by two boundaries – an upper absorbing boundary which registers a response once it is crossed, and a lower reflective boundary that can not be crossed. Drift-processes always start at the lower boundary. The rate of the accumulation over time, or (*"drift-rate v"*), influences the speed by which the threshold is reached and corresponding responses are generated. The distance between the lower and upper boundary (i.e. *"threshold a"*) influences how much evidence needs to accumulate before a decision is made. The duration of processes not belonging to decision making (e.g. perception, motor execution) are captured by a single *"non-decision time"* parameter.

Selective response inhibition tasks (SRITs) like the antisaccade task consist of two trial types – congruent and incongruent. In the antisaccade task, congruent trials (also called prosaccade trials) requires subjects to initiate a saccade towards an appearing target. As humans as well as primates have a reflexive, prepotent tendency to look towards changes in their environment, this type of trial illicits no response conflict and is almost automatic. Incongruent or antisaccade trials on the other hand require the subject to inhibit this prepotent response tendency to look at the appearing stimulus and instead look to the *opposite* side by initiating executive control.

A single, prepotent accumulator with drift-rate v_{pre} is responsible for congruent trials. Incongruent trials, however, involve interaction of 3 Wald accumulators (note that this is the architecture as Noorani and Carpenter (2012); see also figure 4.1). The first accumulator shares the same drift-rate v_{pre} with the accumulator from congruent trials but leads to errors if it reaches its threshold before the others do (i.e. the prepotent accumulator). A response inhibition process with drift-rate v_{inhib} stops the prepotent accumulator upon reaching its threshold. Correct incongruent

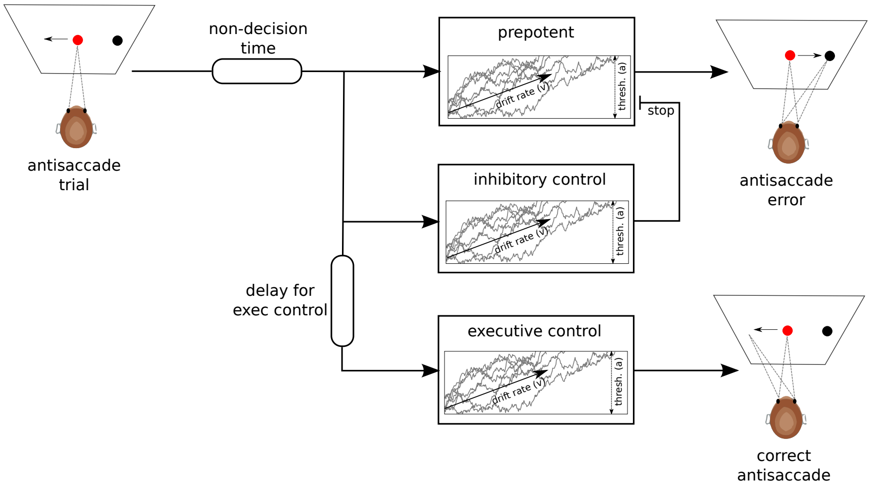

Figure 4.1: Computational process model of the antisaccade task. Depicted is the architecture of accumulators during an antisaccade trial. During prosaccade trials, only the prepotent process is used. See the main text for a description of the model.

responses are committed when the executive control accumulator with drift-rate v_{exec} reaches its threshold before the prepotent accumulator. Start of the executive control accumulator is further delayed by a constant time t_{exec} to capture cognitive processes like rule-retrieval and rule-application (i.e. vector inversion). This delay is also required to capture the commonly observed pattern of fast errors and slow correct responses.

The benefit of an abstract model like the DDM compared to our neural network model is its simplicity and small number of parameters which make it possible to fit it directly to behavior (error rates and reaction time data) and to determine whether changes in behavioral measures are more likely to be related to changes in one or another underlying decision parameter. However, these models make no concrete assumptions of the underlying neurobiological implementation. Here, we aim to combine the strengths of both approaches by fitting the SIDDM to simulated RTs from our neural network model. To identify neural correlates we compare systematic modulations of certain biological parameters of the neural network model and test which DDM parameter best explains the resulting change in RT distributions.

While Ratcliff and Frank (2012) followed a similar approach, the authors fit the standard DDM for 2 alternative forced choice cases to the original BG neural network model (Frank, 2006) which lacked an executive control mechanism for volitional response selection. We thus extend upon this work by including an executive control mechanism in both models to simulate cases in which a prepotent response tendency must be overcome.

The neural network model consists of a sensory input region where stimuli are presented. Different stimulus locations are hard-wired to bias certain responses prepotently via direct projections to the frontal eye fields (FEF). The FEF has directional connections to the output layer of the network – the superior colliculus (SC). In addition, the sensory input area has structured connections to the basal ganglia (BG) which disinhibits the SC and acts as a selective response gate. The striatum of the BG is innervated by dopamine (DA) which excites the response facilitating direct pathway and inhibits the response suppressing indirect pathway. So far, these parts of the network allow correct responding in congruent trials in which the prepotent response matches the target direction. In incongruent trials, however, this prepotent mechanism would only be able to generate errors. Executive control is implemented via the dorso-lateral prefrontal cortex (DLPFC) which integrates sensory input to determine the correct response. DLPFC then (i) activates the correct response unit in FEF, (ii) facilitates the correct response in the direct pathway, and (iii) suppresses the incorrect response via activating corresponding indirect pathway units. The dorsal anterior cingulate cortex (dACC) detects conflict in the FEF (as is the case in incongruent trials where the prepotent and executive control responses differ) and activates the subthalamic nucleus (STN). Once activated, the STN raises the gate of the BG output structures by exciting the substantia nigra pars reticulata (SNr) and suppresses gating of all considered actions. For more details on specific aspects of the network see Wiecki et al. (a).

We generated RTs from the neural network model under the following systematic biological manipulations:

- varying strengths of prepotency by modulating sensory input→FEF connectivity;
- varying degrees of DLPFC→FEF connectivity;
- varying levels of tonic DA in striatum;
- varying degrees of STN→SNr connectivity;
- varying degrees of FEF→striatum connectivity.

Parameter ranges for each modulation were chosen so as to be in a region that resulted in visible differences in simulated RT distributions.

Behavioral RTs from the neural network model are generated by measuring the time taken for a SC unit to cross a pre-specified threshold for response execution. We then fit the SIDDM to the simulated response time data of our neural network model. As a closed-form solution to the SIDDM likelihood is difficult to compute we used probability density approximation (PDA) introduced by Turner and Sederberg (2013). This likelihood-free method only requires simulation of data from a generative process and approximates a likelihood function using kernel density estimation. We can then evaluate the data on the approximated likelihood to compute the summed log probability and find the best fitting parameters using Powell optimization (Powell, 1964) with basin-hopping (Wales and Doye, 1997) to avoid getting stuck in local maxima. For each biological manipulation we allowed one SIDDM parameter to vary while fixing all others. We then repeated this process for each SIDDM parameter and for each simulated biological manipulation. We performed model selection by comparing the log probability (logp) to assess which SIDDM parameter alteration best accounts for the specific manipulation in our neural network model, and then

	a	t_{exec}	v_{exec}	v_{pre}	v_{inhib}
sensory→FEF connectivity	2437	2345	2325	**2558**	2370
sensory→striatum connectivity	2384	1969	1965	**2681**	2231
DLPFC connectivity	2102	**2399**	2355	1980	1944
tonic DA	**2104**	1926	1966	2046	1925
STN connectivity	**2262**	2213	2248	2181	2098
FEF→striatum connectivity	**1767**	1744	1733	1738	1716

Table 4.1: Goodness-of-fit (assessed by log probability) for different SIDDM parameters for different manipulations in the neural network model. All models have the same complexity. Best-fitting log probabilities are in bold.

plotted the estimated parameters to determine the nature of this relationship.

4.3 Results

We used the SIDDM to quantitatively fit error rates and RT distributions produced by the neural model, and performed model selection to determine which free parameter of the SIDDM best accounted for variations of each network model parameter. Higher logp values [1] represent better fit. As can be seen in table 4.1 the different manipulations in the neural network were accounted for by both distinct and overlapping SIDDM model parameters.

Based on our design of the neural network we had several hypotheses of how biological parameters relate to SIDDM parameters. As sensory→FEF and sensory→striatum connections are associated with the prepotent response mechanism of the model, we expected changes along these pathways to be captured by changes in prepotent drift-rate v_{pre}. Contrary, DLPFC→FEF connectivity (i.e., the degree to which rule representations influence motor units) should associate with executive control parameters like t_{exec} and v_{exec} (note that the neural network does not have a separate delay com-

[1] Because all models have the same number of parameters we can use the logp instead of other model comparison measures like the Bayesian information criterion which penalize model complexity.

ponent so we are unable to dissociate the two). Indeed, t_{exec} and, to a slightly lesser extent, v_{exec} captured DLPFC→FEF connectivity modulations. Finally, striatal DA, STN→SNr connectivity and FEF→striatum projection strength all modulate the ease by which the BG gates responses and should thus be associated with decision threshold, which is exactly the pattern we observed. However, inspection of figure 4.2 shows that whereas increases in DA and FEF→striatal projection strength were associated with lower threshold, increases in STN were associated with higher threshold.

Moreover, these different biological measures themselves are modulated by distinct cognitive variables, such as reward and conflict/choice entropy (dopamine and STN). We return to this issue of multiple routes to decision threshold regulation in the Discussion. Across all fits, the relationship between the biological manipulation and the best-fitting DDM parameters was monotonic and largely linear within the selected parameter ranges (see figure 4.2).

Figure 4.3 shows an example of how well the best fitting parameter values in each condition are recovered by the simulated RT distributions of the neural network model. A variation in a single network parameter – sensory→FEF projection strength – changes the amount of evidence needed before the BG gating threshold is reached. Larger projection strengths result in a greater proportion of fast errors. This pattern is very well captured in SIDDM fits by altering only v_{pre} across these runs.

4.4 Discussion

In Wiecki et al. (a) (also see chapter 2), we presented a dynamic neural network model of selective and global response inhibition which provides a description of the distributed computations carried out by individual brain regions and neuro-

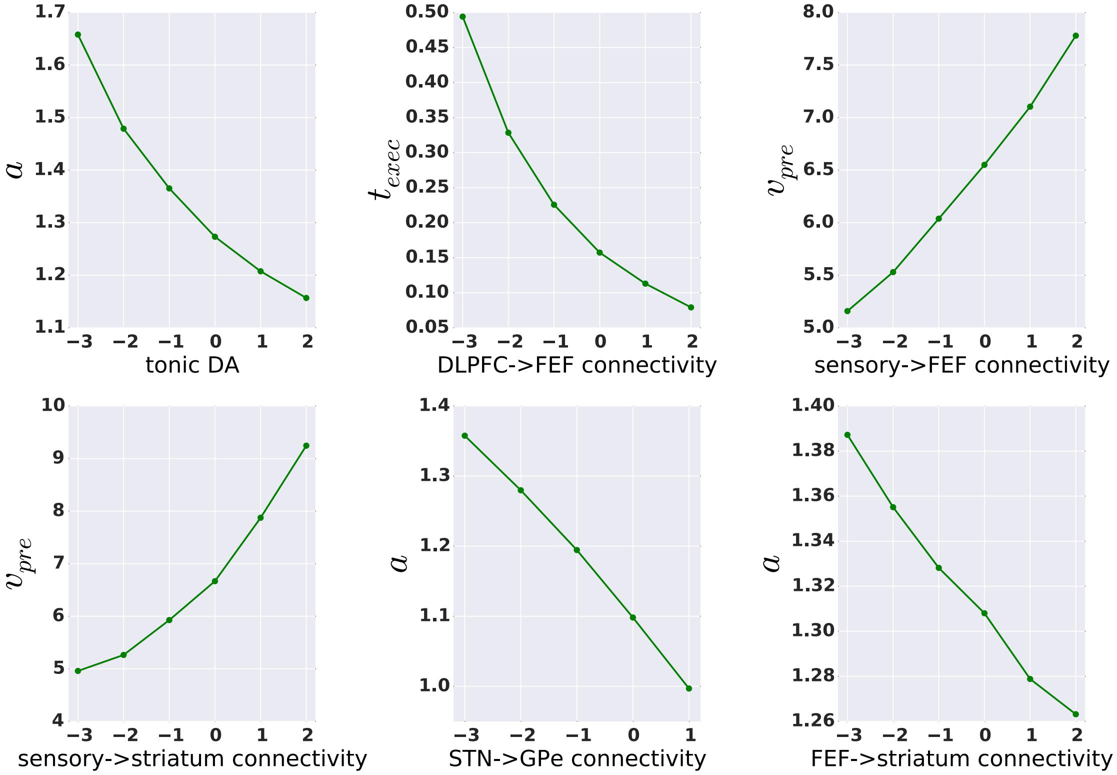

Figure 4.2: Best fitting values of SIDDM parameters for systematic modulations of different biological parameters in the neural network model. For each condition (e.g. DLPFC speed) the best fitting model was chosen (see table 4.1). As can be seen, systematic modulations of the biological parameters within the sensitive range results in a monotonic and linear relationship to the SIDDM parameters that explain the behavioral data. Note that the x-axis has been re-scaled as the absolute value of parameters is dependent on the neural network model and different parameters are influenced quite differently by quantitative changes.

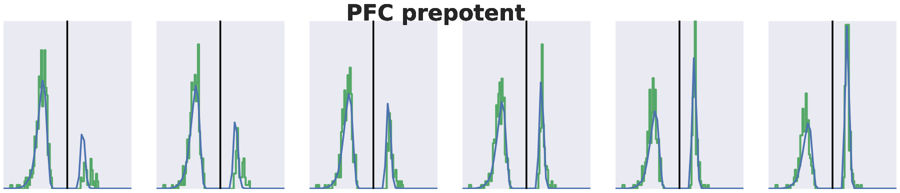

Figure 4.3: Each subplot contains RT distributions for incongruent trials by network models (green) and likelihood (blue) using the best-fitting parameters of a SIDDM model. Correct RT distributions are on the left side of each panel and incorrect RT distributions are on the right side, mirrored on the y-axis. Neural network manipulations of sensory→FEF projection strengths were increased from left to right, with only v_{pre} varying accordingly for the SIDDM fits.

transmitters. Briefly, a prepotent response mechanism connects the sensory cortex directly to FEF which starts to activate the prepotent response in the output area of the superior colliculus (SC) and gates it via the BG. The correct response is selected in the same manner by the slower volition response selection mechanism implemented by the DLPFC which integrates sensory information together with the trial instruction. The complexity of this model is grounded by well established neuroanatomical and physiological considerations and accounts for a wealth of key data from electrophysiological, behavioral, lesion, pharmacological and imaging studies. To capture the emerging fundamental computational properties of this complex system as a whole we present a more parsimonious higher-level computational description in form of a psychological process model – the SIDDM. We leveraged the simplicity of this level of description in a systematic effort to assess how variations in neural network parameters exert their influence on higher level process parameters, by fitting response time distributions generated by the neural network model. This exercise allowed us to show both convergent and divergent biological influences onto higher level processes. Whereas multiple neural mechanisms converge to influence the single decision threshold parameter (albeit under different conditions; see below), distinct neural mechanisms influence distinct SIDDM parameters.

In the past, behavioral analyses of selective response inhibition task performance has primarily been limited to mean RT and mean accuracy (but see Hübner et al., 2010; Whitea et al., 2010a; Noorani and Carpenter, 2012). However, as we demonstrated in Wiecki et al. (a) (see chapter 2 above), these summary statistics are influenced by manipulations of distinct biological mechanisms, which are in turn affected by different cognitive factors (e.g., motivation/reward value and dopamine; conflict and STN; DLPFC speed and manipulations of working memory retrieval speed). Thus, using only the summary statistics often does not permit inference on the underlying causes of performance differences.

The model presented here deconstructs the congruent and incongruent RT distribution into separate cognitive processes that relate to prepotency (v_{pre}), inhibition (v_{inhib}), executive control (v_{exec} and t_{exec}), as well as caution (a) and motor execution speed (t). This model uses the same architecture as Noorani and Carpenter (2012) who showed that this model can capture many patterns of RT distributions commonly observed in antisaccade tasks. As we demonstrated above, this model allows us to adequately fit the RT patterns resulting from our neural network model (see figure 4.3), and allows us to interpret its fundamental computations in terms of a well characterized model. Moreover, by combining these two levels of modeling we derive predictions about which abstract model parameter relates to which neural mechanisms. Specifically, we find that (i) increases in bottom-up saliency by strengthening sensory→striatum as well as sensory→FEF bias weights influence the prepotent drift-rate v_{pre}; (ii) modulating the effectiveness of DLPFC communication with FEF influences t_{exec} and, to a slightly lesser extent v_{exec}; and (iii) DA, STN and fronto→striatum manipulations all influence threshold, with increases in STN strength leading to increases in estimated threshold, and DA or FEF→striatal connectivity leading threshold decrements. Interestingly, the relationship between neural model manipulations and SIDDM parameter appears to be not only monotonic, but linear within sensitive parameter ranges. Threshold reductions in the SIDDM are associated with reduced accuracy and speeded RTs. While increasing v_{pre} also leads to decreased accuracy and faster error RT, correct RT is not influenced. Conversely, increasing the speed or effectiveness of cognitive control by decreasing t_{exec} or increasing v_{exec} increases accuracy, reduces correct RT but has only a minor influence on error RT.

These predictions certainly need to be further validated in empirical studies. Our hope is that this type of explicit formulation of neural mechanisms and their effects on generative parameters that give rise to behavioral observables (RT distributions), will enable better characterization of distinct underlying mechanisms leading to performance deficits in psychiatric conditions. As noted above, increased error rates and faster mean RTs do not allow us to assess the neurobiological source of these deficits. In chapters 6 and 5 below we demonstrate how applying the SIDDM to SRIT data of depressive and Huntington's diseased patients helps identify the source of these deficits.

4.4.1 Multiple mechanisms of decision threshold regulation in fronto-basal-ganglia circuitry at different time scales

Different mechanisms in our neural network influence decision threshold regulation operating at distinct time scales, and modulated by distinct cognitive variables. First, the strength of cortico-striatal projections regulate the ease with which cortical motor plans can be gated by the BG, allowing for speed emphasis in the speed-accuracy trade-off (see figure 4.2). This aspect of our model is quite similar to the model of Lo and Wang (2006) and was subsequently corroborated by Forstmann et al. (2010a). Our multi-level modeling approach converges on the same conclusion but extends this view by showing that decision threshold is also more dynamically regulated on a shorter time-scale by (i) motivational state (changes in DA levels, which are modulated by reinforcement and also facilitate striatal Go signals); and (ii) response conflict and saliency (via the hyperdirect pathway, making it more difficult or Go signals to drive BG gating (Jahfari et al., 2011)).

Our drift-diffusion model analysis also provides a refined view on the computational mechanism of response inhibition. We find that the hyperdirect pathway, implicated in response slowing as a function of conflict (when there is value in the alternative course of action), functions to adjust decision thresholds at the computational level. Specifically, we find that STN efficacy in the neural model is positively correlated with increases in estimated decision threshold (see figure 4.2; (also Ratcliff and Frank, 2012; Jahfari et al., 2011, 2012)). Evidence for conflict-induced decision threshold adjustment via the hyperdirect pathway has been recently described in a reinforcement-based decision making task (Cavanagh et al., 2011). In that study, increases in frontal EEG activity during high conflict decisions were related to increases in decision threshold estimated by the DDM. Intracranial recordings directly within the STN also revealed decision conflict-related activity during the same time period and frequency range as observed over frontal electrodes. Moreover, disruption of STN function with deep brain stimulation led to a reversal of the relationship between frontal EEG and decision threshold, without altering frontal activity itself. These data thus support the notion that frontal-STN communication is involved in decision threshold adjustment as a function of conflict. Similarly, proactive preparation to increase decision threshold in the stop signal task when stop signals are likely is associated with hyperdirect pathway activity (Jahfari et al., 2012).

Notably, the hyperdirect pathway is also implicated in our stop-signal task simulations in which responses have to be inhibited altogether. In that case, STN activity is increased by the salience of the stop-signal detected in rIFG. Thus according to this model, global response inhibition does not imply stopping evidence accumulation, but rather a transient increase in decision threshold, allowing for continued accumulation but difficulty initiating any response. This hypothesis could be tested in the stop-change task, which requires initiation of an alternative response after the stop signal. Preliminary evidence in support of this notion comes from the observed response slowing of the change response (Sharp et al., 2010; Fleming et al., 2010; Chatham et al.). Response slowing is also observed during salient oddball trials (Barcelo et al., 2006; Parmentier et al., 2008). Our model predicts that in both of these cases, neural activity related to accumulation of evidence for the response will proceed as usual, with the same slope as in non-switch or non-oddball trials, just with a higher threshold of response execution.

Chapter 5

A computational cognitive biomarker for cognitive and motor control deficits in Huntington's Disease

This chapter will be submitted for publication and reflects contributions of other authors: **Wiecki T. V.**, Frank M. J. (in prep). A computational cognitive biomarker for cognitive and motor control deficits in Huntington's Disease

Abstract

Huntington's disease (HD) is genetically determined but with variability in symptom onset, leading to uncertainty as to when pharmacological intervention should be initiated. Here we take a computational approach based on neurocognitive phenotyping, computational modeling, and classification, in an effort to provide quantitative predictors of HD before symptom onset. A large sample of patients – consisting of both prodromal individuals carrying the HD mutation (pre-HD), and symptomatic patients after progression to late-stage HD – as well as healthy controls performed the antisaccade task, which requires executive control and response inhibition. While symptomatic HD patients differed substantially from controls in behavioral measures (RT and error rates), there was no such clear behavioral differences in pre-HD. RT distributions and error rates were fit with an accumulator-based model which summarizes the computational processes involved and which are related to identified mechanisms in more detailed neural models of prefrontal cortex and basal ganglia. Classification based on fitted model parameters revealed a key parameter related to executive control differentiated pre-HD from controls, whereas the response inhibition parameter declined only after symptom onset. These findings demonstrate the utility of computational approaches for classification and prediction of brain disorders, and provide clues as to the underlying neural mechanisms.

5.1 Introduction

Huntington's disease (HD) is a debilitating neurodegenerative disease with progressive degradation of motor and cognitive function. From a neurocognitive perspective, HD is a highly intriguing disorder as it has a clearly defined, single genetic mutation in the form of an expanded CAG repeat in the *HTT* gene which predicts with certainty that the disease will develop in an individual. The effects of this mutation on neurobiology have been the subject of intense study with notable progress, although many questions still remain. Indeed, no clinical phase 3 trial to date has been successful for a drug that slows or reverses progression of HD, raising the question of whether the most efficient drug development methods are being leveraged (Kieburtz and Venuto, 2012). A central requirement for success in clinical trials are objective and quantitatve outcome measures that are sensitive to early-stage changes in presymptomatic individuals (pre-HD). Better clinical markers of disease progression could inform when to initiate treatment: too early would increase accumulation of negative side-effects, whereas too late could prevent succesful therapeutic intervention.

TRACK-HD was a large, multi-site, longitudinal study to evaluate various behavioral and imaging measures for their appropriateness in tracking HD progression (Tabrizi et al., 2009). While many measures were sensitive to changes in late-stage HD, a key conclusion was that "these measures are insensitive to change in pre-HD over timescales realistic for clinical trials (Tabrizi et al., 2013) and more sensitive measures are required to capture subtle changes that might be taking place before symptom onset." (Andre et al., 2014). In sum, there is a current lack of clinical markers sensitive to the cognitive changes that occur during the pre-HD stages.

The antisaccade task has been widely used to study executive control and response inhibition of eye movements that has well-studied and dissociable neural mechanisms associated with (i) the prepotency of a pro-saccade response, (ii) the inhibition of that response, and (iii) the executive control needed to dictate the alternative response given the instructed task rule (Wiecki and Frank, 2013; **?**). Notably, several studies have found reliable antisaccade performance deficits in HD patients well before full onset of HD symptoms (e.g. Klöppel et al., 2008; Peltsch et al., 2008; Hicks et al., 2008).

Traditional studies with this task mostly analysed and interpreted behavioral summary statistics such as mean reaction time and accuracy. However, despite the apparent task simplicity, its successful completion involves an intricate interaction within a complex network of brain areas including the frontal cortex and basal ganglia. Indeed, neural circuit modeling and empirical studies suggest that a deficit in any of the involved areas can lead to increased error rates and reaction times, leading to ambiguity in interpretation of observed deficits (Wiecki and Frank, 2013). The emerging field of computational psychiatry (Montague et al., 2011; Maia and Frank, 2011) approaches this problem with the help of computational models that can deconstruct behavioral and neural data into separable generative processes, and to identify whether any of these processes is preferentially altered in mental illness (Wiecki et al., b).

At a mechanistic level, the classical view is that HD arises from selective neurodegeneration within the indirect pathway of the basal ganglia that normally acts to suppress unwanted movements (Aylward et al., 2004; Hobbs et al., 2009; Paulsen et al., 2010; Majid et al., 2011b,a; Tabrizi et al., 2009). In addition to this clearly defined atrophy, there is also more widespread degeneration in frontal cortex (Peltsch et al., 2008; Klöppel et al., 2008; Rao et al., 2014), which could act to impair executive control over action selection Miller and Cohen (2001); Badre (2008); Collins and Frank (2013); Wiecki and Frank (2013).

The aim of the current study was to apply quantitative computational modeling to the TRACK-HD behavioral data set to separate processes thought to relate to selective response inhibition and executive control. We then use machine learning classification to demonstrate that executive control parameter is predictive of HD prior to symptom onset, whereas response inhibition processes are impaired only after motor symptoms are observed..

5.2 Methods

371 subjects performed the antisaccade task, consisting of 123 healthy controls (mean age 46±10 years), 122 presymptomatic gene carriers (pre-HD; mean age 41±8.7 years) that will develop HD later in life, and 125 patients diagnosed with HD

(mean age 49.3±9.8 years). Pre-HD patients were further subdivided into pre-HD-A and pre-HD-B, where pre-HD-B were estimated to be closer than pre-HD-A to progression to HD based on CAG repeat length, indicative of how fast gene carriers progress to late HD (MacMillan and Quarrell, 1996). HD patients were similarly divided into HD-1 and HD-2, indicating relative disease progression with HD-2 group having overall stronger symptom severity.

Several clinical measures were collected. The Unified Huntington's Disease Rating Scale (UHDRS) is the standard assessment tool for HD symptom severity and has two relevant subscores: Total functional capacity (TFC), tracking ability to perform daily events, and the total motor score (TMS) tracking motor abilities specifically (Klempír et al., 2005).

Eye-tracking was used to measure subjects' eye movements. In the antisaccade task subjects had to fixate a central cross on a computer screen. After a fixed delay, a target stimulus appeard on either the left or right side of the fixation cross and subjects had to either saccade towards the target (prosaccade) or to the opposite side (antisaccade). Pro and antisaccades were randomly interleaved. Prosaccade errors were very rare and not analyzed further.

Mean and standard-deviaton (SD) of prosaccade RTs, mean and SD of correct and error antisaccade RTs, as well as accuracy on antisaccade trials were computed as summary statistics.

5.2.1 Distributional analysis

Summary statistics are a useful and easy measure to compute. But while mean and variance can describe a Gaussian distribution perfectly, RT distributions are well known to be quite skewed and non-normal. Thus, summary statistics often fail to capture more nuanced aspects of conflict resolution that are present in the full RT distributions of correct and error trials. Indeed, distributional analysis can help tease apart different processes that can lead to various changes in the RT distributions (due to conflict or other factors), such as a shift in the entire distribution, or preferential changes to the leading edge or the tails of the distribution, and how any such changes are related to increased or decreased accuracy (Ridderinkhof et al., 2005; Noorani and Carpenter, 2012; Ratcliff and McKoon, 2008). Distributional analysis typically involves dividing the RT distribution into quantiles, e.g., the mean of the first 20% of the RT distribution, the second 20%, and so on.

In order to better capture differences in the RT distribution between congruent and incongruent trials, Ridderinkhof et al. (2005) suggested the use of delta-plots. For each subject, the 5 RT quantiles are computed for pro and antisaccade trials separately (only correct antisaccade trials are used). Each quantile is then averaged across pro and antisaccade trials and plotted along the x-axis. To capture conflict-induced slowing, mean RT for each antisaccade quantile is subtracted from mean RT of the corresponding prosaccade quantile and plotted along the y-axis. Thus, the relative slowing for antisaccades compared to prosaccades is captured by a positive y-value in the delta plot. The commonly observed effect is that conflict effects are observed to a greater degree on early RTs, as captured by a decreasing slope of the delta-plot.

5.2.2 Computational modeling

While the delta-plot can reveal behavioral signatures of conflict resolution it does not provide a process level description of how such signatures arise. To this end, we fit a computational model summarizing the three major components to the behavior in the task and which approximate those embedded in more detailed neural models. The model is an extension of a sequential sampling model typically used in two-alternative forced-choice decision making tasks, in which sensory evidence is accumulated up to a response threshold used to initiate motor activity, and where the speed of evidence accumulation is reflected by a “drift rate”. The extended model used here takes into account the dynamics and interactions of prepotent responses, response inhibition, and executive control. As such, the model comprises three single-boundary Wald accumulators: a prepotent (pre), an inhibitory (inhib) and an executive control (exec) accumulator (see figure 5.1). These accumulators race against and interact with each other. Each accumulator is associated with an individual drift-rate (v_{pre}, v_{inhib} and v_{exec}) that determines the speed of integration towards its threshold a. To take into account additional time unrelated to decision processes but summarizing sensory perception and motor execution, we also incorporate a non-decision time parameter t. If the prepotent accumulator reaches its threshold first during an antisaccade trial an error is commited. If the inhibitory accumulator reaches the threshold before the prepotent one, it stops the prepotent accumulator from reaching its threshold. In addition, the executive control accumulator is delayed by a fixed time (t_{exec}) to capture additional time required for rule-retrieval, vector inversion etc. Once it reaches threshold a correct antisaccade is performed. While parameters of the prepotent accumulator (i.e. v_{pre}, a and t) are identified by fitting across both pro and antisaccade trials, all other parameters are fit using only antisaccade trials (as they are irrelevant in prosaccade trials).

As demonstrated in chapter 4, these parameters relate to separately identifiable underlying neurocognitive processes. While v_{exec} and t_{exec} relate to frontal functional connectivity and integration speed, v_{pre} captures cortico-cortical as well as sensory→striatal connectivity. Threshold a on the other hand is influenced by motivational state via tonic dopamine levels and conflict-related processing via the hyperdirect pathway.

As a closed-form solution to this likelihood is difficult to compute we used probability

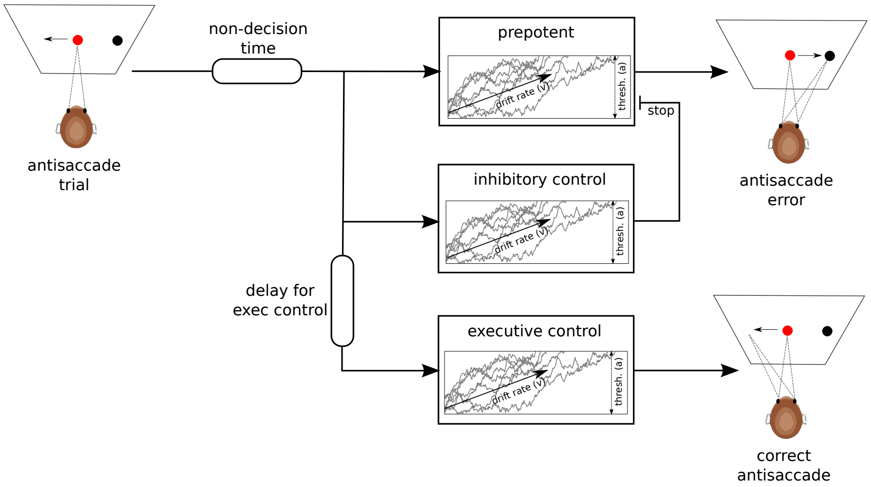

Figure 5.1: Computational process model of the antisaccade task. Depicted is the architecture of accumulators during an antisaccade trial. During prosaccade trials, only the prepotent process is used. See the main text for a description of the model.

density approximation (PDA) introduced by Turner and Sederberg (2013). This likelihood-free method only requires simulation of data from a generative process and approximates a likelhood function using kernel density estimation. We can then easily evaluate the data on the approximated likelihood to compute the summed log probability and find the best fitting parameters using Powell optimization (Powell, 1964) with basin-hopping (Wales and Doye, 1997) to avoid getting stuck in local maxima [1].

5.2.3 Machine Learning

In order to assess the viability of using these methods to classify patients we used machine learning classifiers based on summary behavioral statistics and computational model parameters. The goal was to train classifiers based on a sample of patients and test whether the classifier could discriminate between novel groups of subjects based on behavioral and model parameters. For two-class classification we used logis-

[1] While ideally we would use hierarchical Bayesian estimation of the model parameters (Wiecki et al., c) the small randomness along with the large number of simulations required for a single evaluation of the PDA likelihood function lead to convergence issues and prohibitively long running times.

tic regression with L2-regularization. To optimize the strength of the regularization parameter we ran 10-fold stratified cross-validation which keeps the distribution of labels constant across every split. During cross-validation, the classifier is trained to differentiate 90% of the subjects but tested and evaluated based on its classification accuracy of the previously unseen 10% of subjects. This splitting procedure is repeated 10 times so that all data has been used once to test the classifier. To evaluate the performance of this classifier we ran this cross-validation procedure 200 times on training data and tested the best-performing classifier on held-out test data in a shuffle-split cross-validation with 20% of the data used for testing each time. Classifier performance was then compared using the Area Under the Receiver-Operator-Characteristic Curve (AUC), a measure robust to unequal class sizes. Intuitively, it can be interpreted as the probability of correctly classifying two samples randomly drawn from each of the classes. For multiclass classification we used a Random Forest classifier (Breiman, 2001) that was trained in the same manner.

5.3 Results

5.3.1 Behavior

Standard measures of behavior were more than sufficient to discriminate HD patients from both controls and pre-HD. Specifically, for prosaccade trials, control subjects t(246)=-3.25, $p = 0.001$) as well as pre-HD subjects (t(245)=-3.13, $p = 0.002$) were significantly faster (0.344±0.0806 secs and 0.357±0.0799 secs, respectively) than HD patients (0.398±0.1226 secs; see figure 5.2a). A similar pattern emerged in antisaccade trials where control subjects t(246)=-4.25, $p < 0.001$ as well as pre-HD subjects t(245)=-3.39, $p = 0.001$ were significantly faster (0.344±0.0806 secs and 0.355±0.0866 secs, respectively) than HD patients (0.402±0.1308; see figure 5.2a). Control subjects t(246)=9.68, $p < 0.001$ as well as pre-HD subjects t(245)=8.85, p

< 0.001 were also more accurate (68.4±19.77% and 65.9±19.31%, respectively) than HD patients (41.3±24.06%) on antisaccade trials.

Notably, there was no significant difference between control and pre-HD subjects in mean RT in either prosaccade t(243)=0.15, p = 0.879 or antisaccade t(243)=1.01, p = 0.315 trials, nor in antisaccade accuracy t(243)=-1.00, p = 0.318 (see figure 5.2b). There was, however, a significant trend for pre-HD to demonstrate increase antisaccade RT variability (standard deviation) between pre-HD (0.139±0.0756 secs) and controls (0.122±0.0615 secs), t(243)=1.95, p = 0.052.

5.3.2 Distributional analysis

Delta-plots subtract pro from antisaccade RTs for each quantile along the distribution and show the conflict interference effect (positive deflections) and how it gets resolved over time. The delta-plots for the three different subject groups are shown in figure 5.3. The common pattern of a negative slope Richard Ridderinkhof et al. (2011) is strongly visible in all groups and suggests that conflict is successfully resolved as time progresses. While there are striking differences in the last 3 quantiles between control and HD as well as pre-HD and HD (all p-values < 0.001) there were no differences between controls and pre-HD (all p-values $> .05$).

5.3.3 Computational modeling

Separable effects of response inhibition and executive control

Before describing group differences, it is important to highlight that the model comprises multiple mechanisms by which a correct or incorrect antisaccade is executed. High values of v_{pre} lead to faster prosaccades but also fast antisaccade errors. Both

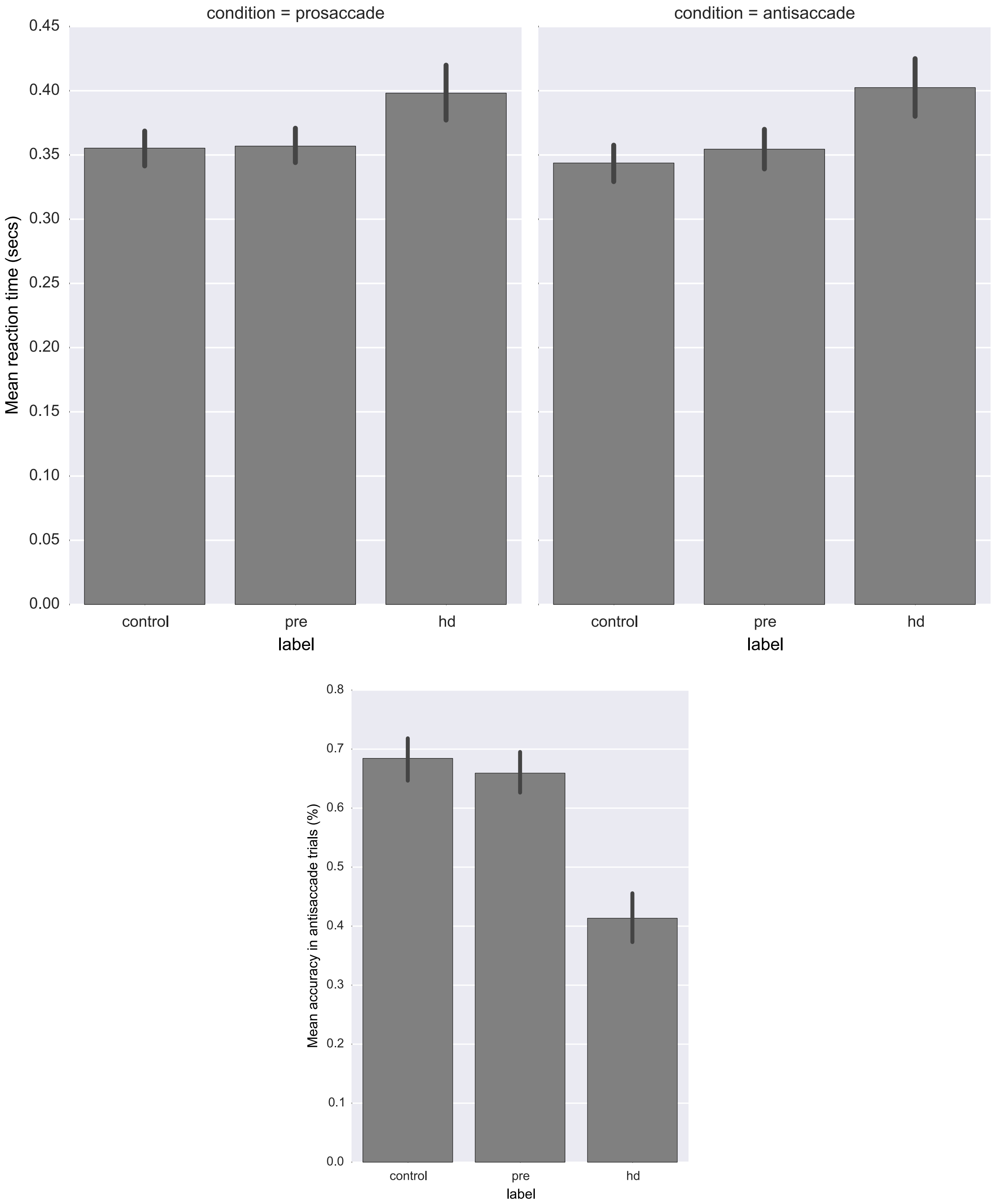

Figure 5.2: **a)** Bar-plots of mean reaction time in seconds across different groups. **b)** Bar-plots of mean accuracy in percent during antisaccade trials across different groups. Error-bars depict 95% confidence intervals.

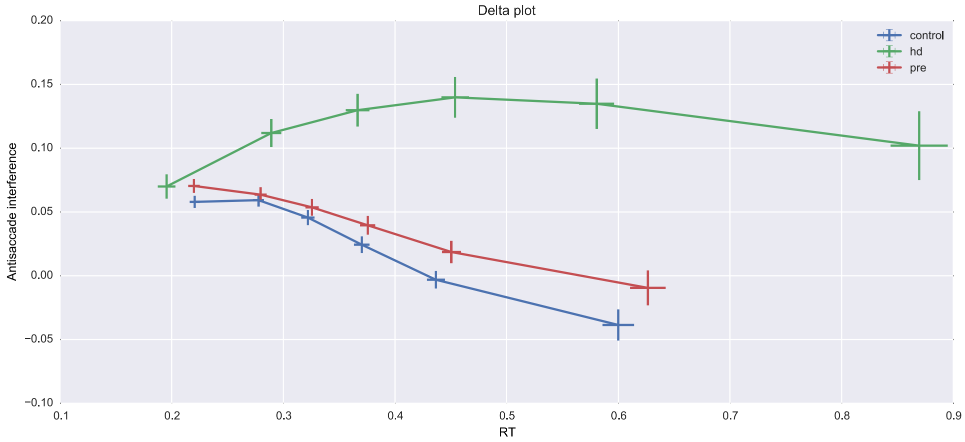

Figure 5.3: Delta-plot showing conflict resolution (negative slope) across time in different groups. Error-bars represent standard errors. See text for details.

the response inhibition parameter v_{stop}, which allows a prepotent saccade to be suppressed, and the executive control parameter v_{exec}, which provides evidence for the controlled antisaccade response, contribute to successful performance (decreased errors). However, high values of v_{exec} lead not only to higher accuracy but faster and less skewed correct antisaccade RTs. In contrast, high values of v_{stop} do not affect antisaccade RTs but rather right-censor the antisaccade error RT distribution (i.e., erroneous pro-saccades will only occur with very fast RTs). Finally, longer t_{exec} time will allow for more time for the prepotent process to reach threshold, and thus will also increase antisaccade errors, but does so by causing a constant shift forward of the whole RT distribution, accounting for the commonly observed pattern of relatively fast errors and delayed correct antisaccade RTs. Thus, each of the model parameters quantify separately identifiable cognitive processes (and putative underlying neural mechanisms). We verified through generative simulations and parameter recovery that indeed these parameters are separately identifiable.

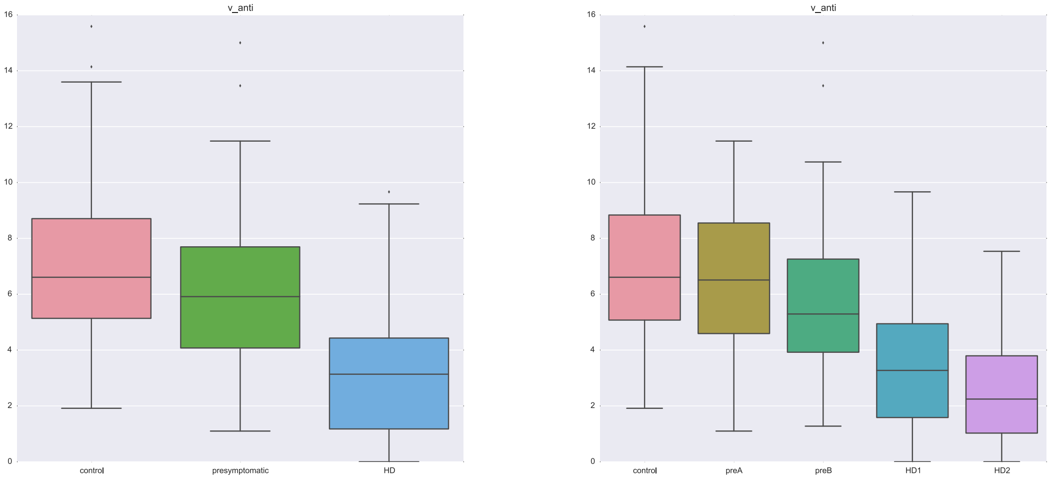

Figure 5.4: **a)** Box-plots of v_{exec} in different groups. **b)** Box-plots of v_{exec} in different subgroups.

Group differences

Unsurprisingly, given the large behavioral differences between symptomatic HD patients and both controls and pre-HD, all model parameters significantly differed between controls and HD as well as between pre-HD and HD (all p-values < 0.01). The more interesting question is whether the refined modeling could help to differentiate pre-HD from controls given that most traditional behavioral analyses revealed no clear differences. Notably, we found that the executive control drift-rate (v_{exec}) was significantly lower t(243)=-2.66, p = 0.008 in pre-HD subjects (6.218±2.6506) compared to controls (7.101±2.5423; see figure 5.4a). This finding suggests subtle executive control deficits in premanifest HD gene carriers. Moreover, visual analysis of changes in executive control drift-rate across subgroups of HD (figure 5.4b) suggests a linear relationship between progression of HD and this parameter, as we assess next.

Dep. Variable:	stage	**R-squared:**	0.393
Model:	OLS	**Adj. R-squared:**	0.383
Method:	Least Squares	**F-statistic:**	38.08
No. Observations:	360	**AIC:**	1116.
Df Residuals:	353	**BIC:**	1143.

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	2.1898	0.298	7.357	0.000	1.604 2.775
v_{exec}	-0.3031	0.027	-11.190	0.000	-0.356 -0.250
v_{pre}	0.1298	0.068	1.905	0.058	-0.004 0.264
a	0.3380	0.124	2.726	0.007	0.094 0.582
t	-0.2199	1.271	-0.173	0.863	-2.719 2.279
t_{exec}	2.7908	0.667	4.181	0.000	1.478 4.103
v_{inhib}	-0.1274	0.029	-4.404	0.000	-0.184 -0.070

Table 5.1: Results of multiple linear regression of model parameters on disease stage where disease stage was coded in a linear way (controls=0, pre-HD-A=1, pre-HD-B=2, HD-1=3, HD-2=4).

Correlations

A multiple linear regression between all model parameters and a linear coding of HD stage (controls=0, pre-HD-A=1, pre-HD-B=2, HD-1=3, HD-2=4) revealed strong correlations between v_{exec}, t_{exec}, v_{inhib} and HD stage. Overall, the model parameters explained 39% of the variance F(4,353)=54.81, $p < 0.001$ (see table 5.1).

While some parameters might show an impairment only after symptoms are evident (e.g., if the mechanisms of motor symptoms are related to the mechanisms producing the reduced model parameter), other parameters might show a more progressive signal even in the stages of pre-HD. We thus assessed for a piecewise linear relationship between parameters and disease stage using a Multivariate Adaptive Regression Spline (MARS) (Friedman, 1991) regression. This iterative algorithm can detect break points in the linear relationship and model them explicitly. The results can be appreciated in figure 5.5. While v_{exec} shows a directly linear relationship, declining from early stages of pre-HD, v_{inhib} seems to only change in later stages once motor

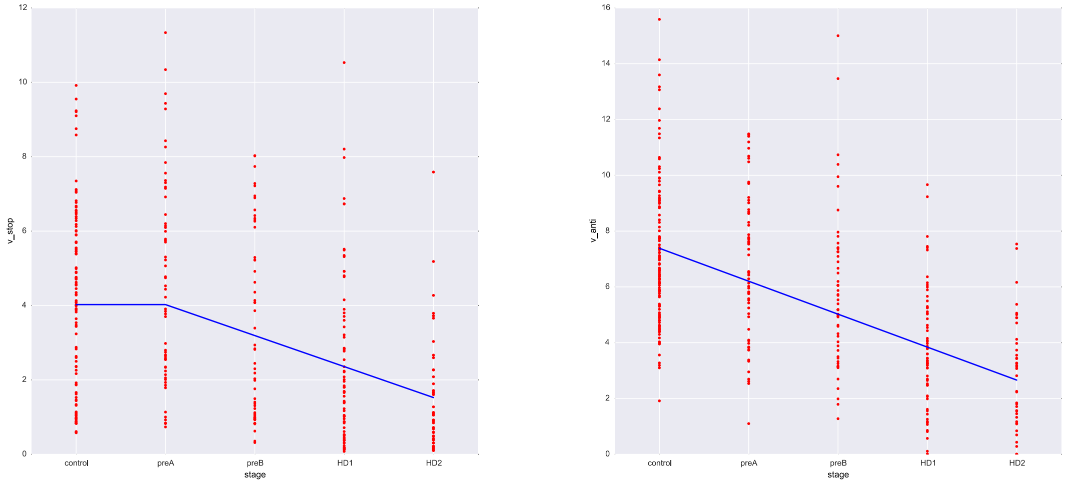

Figure 5.5: Multivariate Adaptive Regression Splines (MARS) estimation of a piecewise linear relationship between v_{stop} **(a)** and v_{exec} **(b)**. See text for more details.

symptoms are present. This fits with our group difference results that showed a significant difference between controls and pre-HD in v_{exec}, but not in v_{inhib}. Interestingly, these results suggest that executive control deficits occur *before* inhibitory control degradation that are only noticable after full HD onset.

Deficits in both v_{exec} and v_{inhib} were also strongly related to patients' TMS motor symptom scores $p < 0.001$ (see figure 5.6 and table 5.2 for a multiple linear regression analysis). Moreover, model parameters were significantly correlated with TFC F(363, 6) = 19.74 and explained 24% of the variance (see table 5.4 for details).

There was no correlation between any of the model parameters and the CAG repeat length in a multlinear regression R=0.02, F(240, 6) = 0.8, p = 0.57.

5.3.4 Machine Learning

We next asked if disease state could be predicted using the model parameters alone, with a focus on clinical applicability. First, we wanted to assess how well each subgroup could be identified given only the model parameters. The confusion matrix in

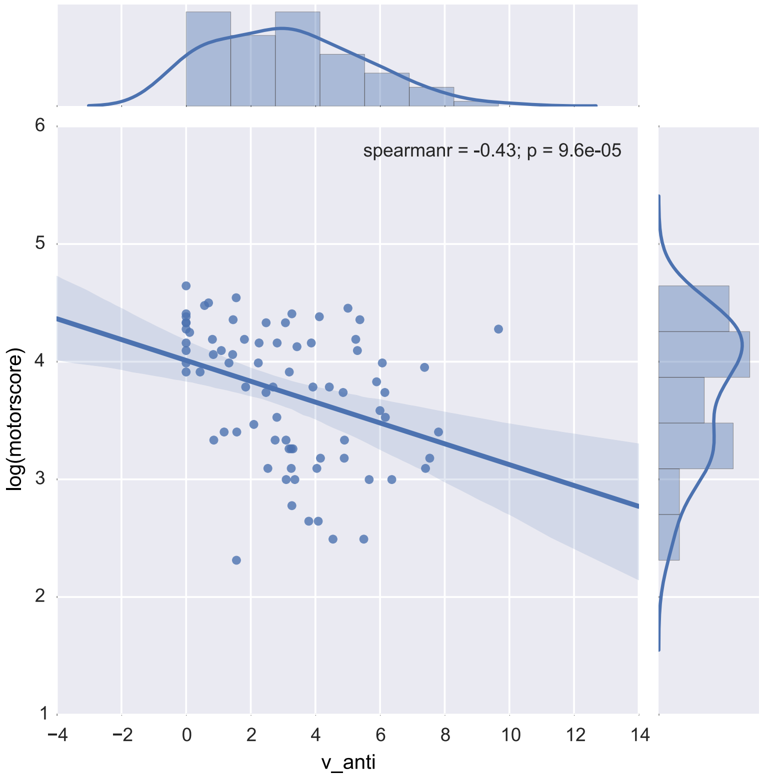

Figure 5.6: Best fitting linear regression line between v_{exec} and log-transformed total motoro score (TMS) on top of raw subject scores.

Dep. Variable:	TMS	**R-squared:**	0.399
Model:	OLS	**Adj. R-squared:**	0.389
Method:	Least Squares	**F-statistic:**	40.14
No. Observations:	370	**AIC:**	2714.
Df Residuals:	363	**BIC:**	2742.

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	14.7502	2.446	6.031	0.000	9.940 19.560
v_{exec}	-2.4764	0.223	-11.109	0.000	-2.915 -2.038
v_{pre}	1.0374	0.554	1.872	0.062	-0.052 2.127
a	3.0337	1.016	2.987	0.003	1.037 5.031
t	-4.3692	10.451	-0.418	0.676	-24.922 16.184
t_{exec}	18.7905	5.513	3.409	0.001	7.950 29.631
v_{inhib}	-1.2665	0.237	-5.334	0.000	-1.733 -0.800

Table 5.2: Results of multiple linear regression of model parameters on total motor score (TMS).

Table 5.3:

Dep. Variable:	tfc	**R-squared:**	0.246
Model:	OLS	**Adj. R-squared:**	0.234
Method:	Least Squares	**F-statistic:**	19.74
No. Observations:	370	**AIC:**	1268.
Df Residuals:	363	**BIC:**	1296.

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	11.7465	0.347	33.890	0.000	11.065 12.428
v_{exec}	0.2365	0.032	7.488	0.000	0.174 0.299
v_{pre}	-0.0971	0.079	-1.236	0.217	-0.251 0.057
a	-0.3199	0.144	-2.223	0.027	-0.603 -0.037
t	0.3705	1.481	0.250	0.803	-2.542 3.283
t_{exec}	-2.1401	0.781	-2.740	0.006	-3.676 -0.604
v_{inhib}	0.1398	0.034	4.154	0.000	0.074 0.206

Table 5.4: Results of multiple linear regression of model parameters on total functional capacity (TFC).

figure 5.7 shows the results of training a random forest and testing its multiclass predictions on held-out data (i.e, predicting patient group status in subjects for whom the training procedure had not seen). The classifier achieves an accuracy of 40% which is modestly above chance (i.e. 33% due to class imbalances).

The next clinical setting we consider is whether a classifier can discriminate between controls and non-symptomatic subjects carrying the CAG repeat mutation. This application could be of interest if any signal picked up by the classifier could help identify pre-HD subjects that might be closer to converting to symptom onset. We compare classifier performance when trained on behavioral summary data (mean and SD RT in pro and antisaccade trials as well as accuracy in antisaccade trials), versus when it is trained on the discriminative model parametes v_{exec}, versus when it is trained on the standard UHDRS assessment score consisting of TMS and TFC. The AUC of the classifiers on held-out data can be appreciated in figure 5.8 All classifiers were significantly better than chance (all p-values < 0.05). As can be seen, UHDRS provides the highest level accuracy ($p < 0.001$) followed by v_{exec}, followed by all model parameters, and finally the summary scores ($p < 0.001$) which operate close to chance.

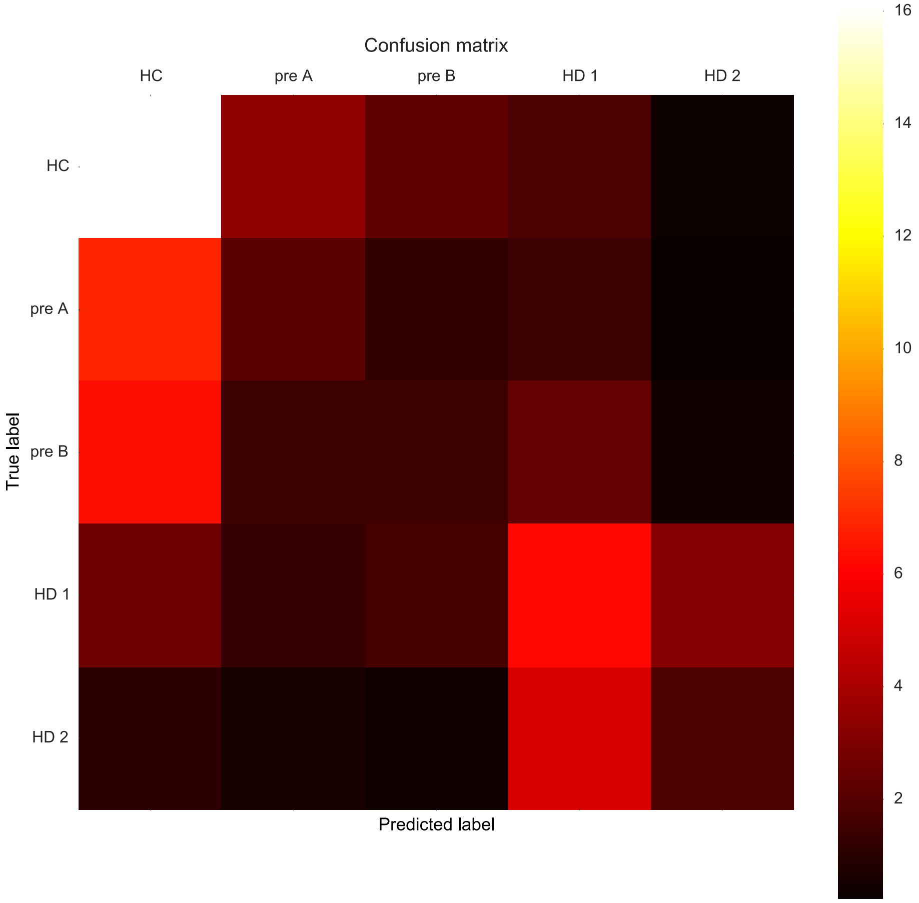

Figure 5.7: Confusion matrix showing true class labels as well as class labels predicted.

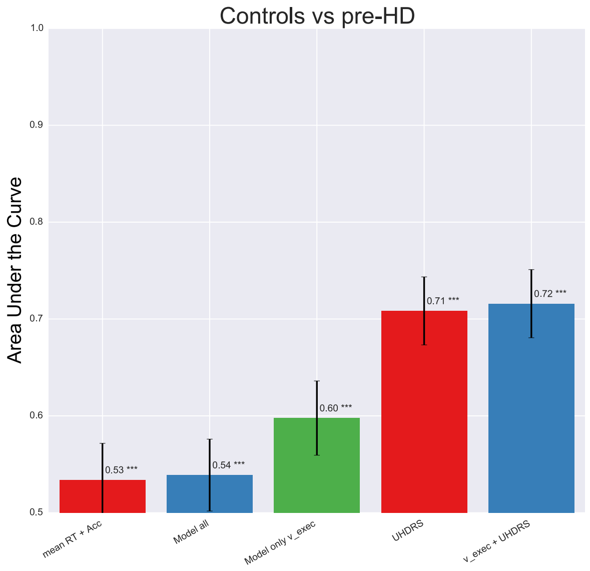

Figure 5.8: Bar-plot comparing Area under the ROC Curve (AUC) of a logistic regression classifier trained on different data to predict HC and pre-HD. Error-bars represent standard deviation.

Next we evaluated how well the above classifer was able to discriminate pre-HD-B subjects from controls. As can be seen in figure 5.9, the success of the classifer was improved when considering pre-HD-B subjects, particularily when the classifiers were trained on both model parameters and UHDRS scores. Summary statistics did not seem to sensitive to this pattern and were had significantly lower AUC than all other classifiers (all p-values < 0.0001). Interestingly, combining v_{exec} with UHDRS scores also leads to higher accuracy than using UHDRS alone.

In the case where we know if a patient has the CAG repeat mutation it is relevant to classify how close a pre-HD individual is to progressing to manifest HD. We thus trained a classifier to predict subgroups pre-HD-A and pre-HD-B. As can be seen in figure 5.10, our previously identified parameter v_{exec} results in the highest accuracy. However, significance testing only relevealed a trend ($p = 0.089$) when comparing v_{exec} to UHDRS scores and no significant difference when comparing accuracy using

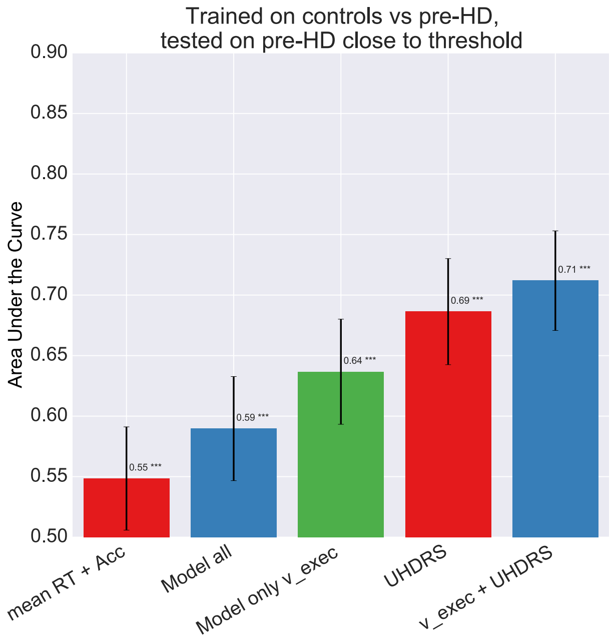

Figure 5.9: Bar-plot comparing Area under the ROC Curve (AUC) of a logistic regression classifier trained to predict controls from pre-HD subjects but evaluated on its performance on predicting pre-HD-B. Error-bars represent standard deviation.

all model parameters to UHDRS (p = 0.28). All parameters, v_{exec} and UHDRS significantly outperformed RT summary measures (all p-values < 0.001). All classifiers were significantly different from chance (all p-values < 0.001). While the combination of v_{exec} and UHDRS scores suggest a slight improvement, this difference was not significant (p = 0.12).

5.4 Discussion

We demonstrated that computational methods based on the antisaccade behavioral data are useful in detecting subtle differences between non-symptomatic HD patients and controls, and between different stages of pre-HD. As in earlier reports, manifest HD patients had longer, more variable RTs as well as increased error rates in antisaccade trials (Klöppel et al., 2008; Peltsch et al., 2008; Hicks et al., 2008). This

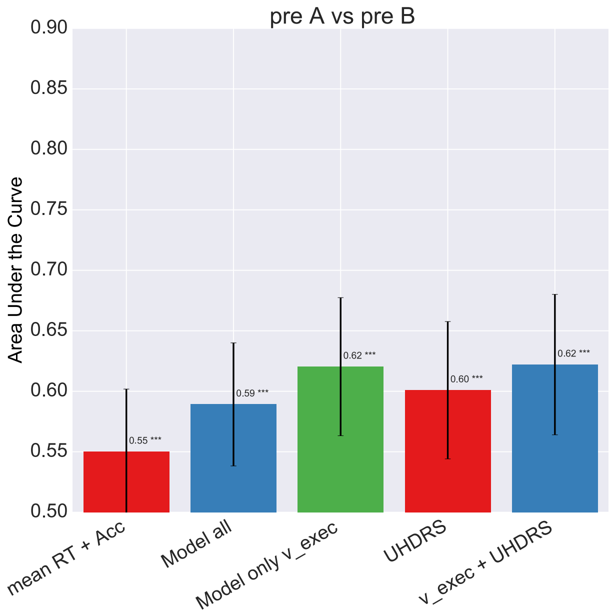

Figure 5.10: Bar-plot comparing Area under the ROC Curve (AUC) of a logistic regression classifier trained on different data to predict pre-HD-A and pre-HD-B. Error-bars represent standard deviation.

result was echoed by our analysis using delta-plots. We then fit a computational model inspired by Noorani and Carpenter (2012) that decomposes the behavior on the antisaccade task into cognitive processes that quantify prepotent response tendencies, speed of inhibitory control to stop the prepotent response when it is maladaptive, and speed and onset time of executive control to initiate volitional saccades. The HD group was associated with differences in every model parameter, suggesting wide-spread neurodegeneration in this group. In contrast, the pre-HD group was selectively associated with deficits in executive control parameter, accompanied by skewed correct antisaccade trials

The pre-HD stage has been mostly been attributed to response inhibition deficits assumed to result from indirect pathway degeneration (Aylward et al., 2004; Hobbs et al., 2009; Paulsen et al., 2010; Majid et al., 2011b,a; Tabrizi et al., 2009; Majid et al., 2013). The indirect pathway of the BG has been suggested to provide a selective NoGo signal that suppresses maladaptive response tendencies (Frank, 2005; ?; Kravitz et al., 2012). Only in later stages, once motor-symptoms set in, other areas become impacted, such as other BG nuclei (subthalamic nucleus and substantia nigra), the thalamus, as well as cerebellum, cortex, and brainstem (Johnson et al., 2001; Kassubek et al., 2005; MacMillan and Quarrell, 1996). Contrary to this theory, our modeling results suggest that the early deficits observed in selective response inhibition tasks such as the antisaccade task result from executive control deficits rather than reduced response inhibition per-se. This result could suggest that it might not be indirect pathway degeneration that occurs in the early, pre-HD stages but rather frontal or fronto-striatal degradations. Our elaborated neural model of these tasks identify a pathway from prefontal cortex to striatum that is involved in executive control to facilitate an adaptive rule-based response Wiecki and Frank (2013). This theory is corroborated by a diffusion tensor imaging study that found reductions in white matter fibers projecting from the FEF to the caudate body of the BG in pre-HD individuals (Klöppel et al., 2008). The amount of this degradation, as well as the UHDRS motor score (Peltsch et al., 2008), are associated with increased RT variability in voluntarily guided saccades, consistent with our findings and with a reduction in drift-rate (Wagenmakers et al., 2007; Ratcliff and McKoon, 2008). Furthermore, some evidence suggests that pre-HD is actually associated with *increased* indirect pathway activity (Milnerwood et al., 2010), perhaps needed to counteract prepotent response tendencies when executive control is weakened. A recent study by Rao et al. (2014) also suggests that deficits in inhibitory control tasks like the stop-signal task are related to reduced activation of frontal areas such as the pre-supplementary motor cortex (pre-SMA) and dorsal anterior cingulate cortex (dACC).

A second explanation of our finding is that it is indeed caused by indirect pathway degradation but in parts of the BG responsible for executive control which could in principle be affected in earlier disease stages than parts of the BG responsible for motor control. The BG has traditionally been associated with gating motor commands (Mink, 1996). However, more recently it was shown that it also is involved in higher cognitive processing such as working memory updating (Frank et al., 2001; McNab and Klingberg, 2008; Baier et al., 2010; Chatham et al., 2014). Anatomically, the BG is known to form loops that originate in cortex, innervate the BG, and connect back up to the cortex via the thalamus in highly structured circuits (Alexander et al., 1986). Dorso-lateral PFC (DLPFC) is associated with executive control (e.g. Miller and Cohen, 2001; Chambers et al., 2009) and consistently activated in antisaccade trials (Wegener et al., 2008; Funahashi et al., 1993; Johnston and Everling, 2006). Notably, DLPFC innervates anatomical regions of the BG distinct from certain motor areas relevant for saccade generation (including FEF (Munoz and Everling, 2004), SEF (Schlag-Rey et al., 1997) and pre-SMA (Congdon et al., 2009; Aron et al., 2007a; Isoda and Hikosaka, 2007)). This alternative account thus suggests that indirect pathway degradations first happen in the BG areas innervated by DLPFC and only later progresses to areas innervated by motor cortex. However at this time, no clear mechanism is known which would lead to this progression within the BG.

These results might also be relevant for clinical and pharmaceutical research. Currently, there are no clinically proven therapies that could reverse the cognitive decline associated with the late stages of this disease. Thus, as with other neuronal disorders like Alzheimer's disease, focus in the clinic shifted towards early intervention to slow the progression which requires detection of subtle cognitive changes before the symptoms become visible neurologically.

Unfortunately, neither summary statistics nor delta-plots showed significant differences between control subjects and pre-HD individuals. Strikingly, however, our computational modeling analysis did show a significant difference in the drift-rate parameter for executive control (v_{exec}). Moreover, when splitting patients into subgroups a linear relationship between v_{exec} and the progressive stages from early pre-HD to late HD emerged. Other model parameters associated with inhibitory control v_{inhib}, delay of executive control, prepotent response bias, response caution and motor execution were only affected in HD patients suggesting non-linear degradation of the various cognitive processes involved in the antisaccade task.

The computational approach provided several advantages. The model allowed us to detect an effect between controls and pre-HD. Moreover, the affacted parameter allows for a more cognitive interpretation of the results. Our classification results show that the model parameters, specifically the above identified v_{exec} parameter can provide higher classification accuracy than RT summary statistics, albeit not by a huge margin. The accuracies overall were not higher than the current clinical standard UHDRS. This result, however, is not surprising given that UHDRS is a key metric used in classifying subject subgroups that we used to evaluate the classifier. Moreover, the classifier was more successful in specifically discriminating pre-HD-B patients from controls, suggesting that it could potentially detect patients that are closer to converting to symptom onset. This hope awaits further data after more patients have converted to be tested. Moreover, in a clinical setting we would likely use a battery of various cognitive tasks that could increase classification accuracy. The fact that data from a single task is competitive with UHDRS in certain circumstances is thus encouraging.

Ultimately, the hope is to identify measures that are more sensitive than TFC and TMS which are of limited clinical use to track disease progression in pre-HD (Tabrizi et al., 2013). As v_{exec} showed correlations with these measures it could be such a clinical marker but it would require more validation and further analysis on longitudinal data to establish it as such.

5.5 Acknowledgements

We are grateful to Sara Tabrizi, Chrystalina Antoniades, Chris Kennard, Beth Borowsky, Monica Lewis, and Mina Creathorn for providing the antisaccade data from the TRACK-IN-HD study and valuable discussions.

Chapter 6

A Computational Analysis of Flanker Interference in Depression

This chapter will be submitted for publication and reflects contributions of other authors:

Wiecki T. V., Dillon D., Pizagalli A., EMBARC Research Group (in prep). A Computational Analysis of Flanker Interference in Depression. *Clinical Psychological Science.*

6.1 Abstract

Background. Depression is associated with poor executive function, butcounterintuitivelyit can lead to highly accurate performance on certain cognitively demanding tasks. The psychological and neural mechanisms responsible for this paradoxical finding are unclear. To address this issue, we applied a drift diffusion model (DDM) to flanker task data from depressed and healthy adults participating in the multi-site Establishing Moderators and Biosignatures of Antidepressant Response for Clinical Care for Depression (EMBARC) study.

Methods. One hundred unmedicated, depressed adults and forty healthy controls completed a flanker task. We investigated the effect of flanker interference on accuracy and response time, and used the DDM to examine group differences in cognitive processes recruited by the task. Findings were interpreted in the context of neural network simulations that relate model parameters from the DDM to the function of cortico-striatal circuitry, which is negatively affected in depression.

Results. Consistent with prior reports, depressed participants responded more slowly but also more accurately than controls on incongruent trials. These data were explained by the DDM, which indicated that although executive control was slow in depressed participants, this was more than offset by decreased prepotent response bias. Model parameters indexing the speed of executive control and prepotency were negatively correlated with anhedonia.

Conclusions. Executive control was delayed in depression but this was counterbalanced by reduced prepotent response bias, illustrating how participants with executive function deficits can nevertheless perform accurately in a cognitive control task. Neural network simulations suggest that these results reflect tonically reduced striatal dopamine in depression.

6.2 Introduction

How does depression affect higher-order cognition? Given its association with maladaptive rumination (Nolen-Hoeksema, 1991) and abnormal frontal lobe function (Wagner et al., 2006), one might expect depression to be associated with uniform deficits in executive function, which refers to the exertion of cognitive control in order to achieve goals in the face of obstacles. Indeed, a meta-analysis found broadly negative effects of Major Depressive Disorder (MDD) on executive function (Snyder, 2013). Incorporating data from 113 studies, the meta-analysis linked MDD to impaired performance on tasks tapping inhibition, set-shifting, and working memory updating. Thus, a strong negative relationship between depression and executive function seems well-established.

However, a close reading of the literature reveals a puzzling pattern that complicates this picture: several studies document positive effects of depression and sad mood on performance in tasks that would seem to depend on executive function. For instance, Snyder and Kaiser (2014) reported that although anxiety impaired selection from amongst competing response options in three language tasks, increased depression facilitated selection once variance associated with anxiety was controlled. As a second example, Au et al. (2003) assessed the effects of sad, positive, and neutral moods on decision-making during financial trading. Across two experiments, sad mood was associated with accurate decisions and conservative allocation strategies, leading to financial gains. By contrast, positive mood was linked to inaccurate decisions coupled with aggressive allocations, leading to poor outcomes: while participants in sad moods profited, those in positive moods incurred net losses. Although sad mood and depression are clearly not equivalent, the fact that excessive sadness is one of two cardinal symptoms of depression (Association, 2013) makes these results surprising; one might have expected a negative effect of sad mood on complex financial decisions, which surely involve executive function.

Finally, studies that have employed the Eriksen flanker task (Eriksen and Eriksen, 1974) have also yielded counterintuitive findings. Several versions of the flanker task exist, but they all share a common structure: participants must report the identity of a centrally presented stimulus that is surrounded by flankers, which can call for either the same response as the central stimulus (congruent condition) or the opposite response (incongruent condition). For example, in the arrow flanker task participants report the direction (left or right) of a central arrow that is flanked by arrows pointing in the same direction (congruent: <<<<< or >>>>) or the opposite direction (incongruent: <<><< or >><>>). Response time (RT) and accuracy are typically lower in the incongruent condition due to interference introduced by the misleading flankers, and resisting this interference is considered evidence of intact executive function.

Against this backdrop, results from two flanker studies are striking (Dubal et al., 2000; Dubal and Jouvent, 2004). In these studies, severely anhedonic undergraduates responded more more accurately (but also more slowly) on incongruent trials than did healthy participants, suggesting that executive function was intact but delayed in the anhedonic group. Because anhedonia is the other cardinal symptom of MDD (Association, 2013), alongside excessive sadness, these data accentuate the paradox: MDD has negative effects on executive function, but its two defining symptomsanhedonia and sadnessare associated with accurate performance on cognitive control tasks. How can these results be explained?

To date, answers to this question have appealed to cognitive styles. Depressed individualsand healthy individuals in sad moodsappear to adopt a deliberative, analytical stance towards information processing (Andrews et al., 2007; PW and JA, 2009). When a task calls for rapid decisions based on intuition, this is counterproductive and accuracy suffers (e.g. Ambady and Gray, 2002). But when fast responses are likely to produce errors, the careful, thorough approach associated with depressed mood can support accurate responding.

Unfortunately, this answer raises a second question: why is depression associated with a systematic information processing style? As yet there is no clear answer, with psychological accounts ranging from a desire to avoid the negative emotions triggered by errors (e.g. Robinson and Meier, 2007), to the operation of an evolutionarily-evolved mechanism that promotes focused attention in order to solve the problems that caused depressed mood in the first place (PW and JA, 2009). These accounts are intriguing, but they are somewhat difficult to test and they have not been directly related to brain function. In the current study, we seek to address these limitations by using the drift diffusion model (DDM; Ratcliff & McKoon, 2008). The DDM can identify specific cognitive processes that support performance in the flanker task and that are influenced by depression (Pe et al., 2013b; Hübner et al., 2010; Whitea et al., 2010b). Furthermore, because the DDM has been studied in the context of neural network simulations of cortico-striatal-thalamic circuits (Frank & Ratcliff, 2013; Wiecki & Frank, in prep), its use permits inferences about abnormal brain function in depression. This work is part of a larger effort to advance psychiatric research by focusing on individual and group differences in the computations performed by different brain systems (Maia & Frank, 2011; Montague et al., 2011; Wiecki & Frank, in press). Our goal here was to determine if the DDM could uncover changes in basic cognitive functions that would explain slow but accurate performance in depression, and that could also be related to the growing literature on the neuroscience of depression.

6.3 Method

The data described here were collected in a multi-site study examining predictors of antidepressant treatment response in unipolar depression, entitled Establishing Moderators and Biosignatures of Antidepressant Response for Clinical Care for Depression (EMBARC) (http://clinicaltrials.gov/show/NCT01407094). The four sites are Columbia University Medical Center in New York, Massachusetts General Hospital and McLean Hospital in Massachusetts, the University of Texas Southwestern Medical Center, and the University of Michigan. Participants with unipolar depression complete several behavioral, self-report, and physiological assessments prior to enrolling in a double-blind, placebo-controlled clinical trial, designed to identify biomarkers of response to the selective serotonin reuptake inhibitor sertraline. Data collection is ongoing and the blind is unbroken, thus we do not consider treatment response here. Instead, we present an analysis of flanker task data from the first 100 depressed participants enrolled in the study and 40 healthy adults who served as controls.

6.3.1 Participant recruitment, eligibility criteria, and payment

Participants were recruited using flyers and posters, and by research coordinators who visited local clinics. All participants provided informed consent following procedures approved by the site IRBs. Adults aged 18-65 of all races and ethnicities were invited to participate. Eligible depressed participants met DSM-IV criteria for nonpsychotic MDD, as assessed via the SCID-I/P (MB et al., 2002), and scored 14 or above on the self-report version of the 16 item Quick Inventory of Depression Symptomatology (QIDS-SR16; Rush et al., 2003). Based on published norms, this QIDS-SR16 score corresponds to moderate depression (Rush et al., 2003). Exclusion criteria included: lifetime psychotic depressive, schizophrenic, bipolar, schizoaffective, or other Axis I psychotic disorder; current primary diagnosis of obsessive compulsive disorder; meeting DSM-IV criteria for either substance dependence in the six months prior (excluding nicotine), or substance abuse in the past two months; actively suicidal or requiring immediate hospitalization; or presence of any unstable medical conditions that would likely require hospitalization during the duration of the study. Critically, no depressed participant was being treated with antidepressant medication when the data described here were collected.

Data from two depressed individuals were excluded due to difficulty following instructions and technical problems, leaving a sample of 98 depressed participants (New York: n = 21; Massachusetts: n = 10; Texas: n = 44; Michigan: n = 23). Ten healthy controls who did not meet criteria for any Axis I disorder were also tested at each site. Participants were paid $50 for completing the testing session, which included additional tasks not described here.

6.3.2 Questionnaires

Participants in the EMBARC study complete several questionnaires directed at a variety of topics, including personality traits, social functioning, and medical history. In addition to the QIDS-SR16, we concentrate on data from the Snaith Hamilton Pleasure Scale (SHAPS Snaith et al., 1995), a measure of anhedonia. We focus on the SHAPS because performance on the flanker task is sensitive to anhedonia (Dubal et al., 2000; Dubal and Jouvent, 2004).

6.3.3 Flanker task

We used a flanker task with an individually-titrated response window (Holmes et al. 2010). Participants completed a 30-trial practice session that included 15 congruent trials and 15 incongruent trials. The flanking arrows were first presented alone (duration: 100 ms) and were then joined by the central arrow (50 ms)--the total stimulus duration was thus 150 ms. Participants were asked to indicate whether the center arrow pointed left or right by pressing a button, and accuracy and RT were recorded.

Participants next completed five blocks of 70 trials (46 congruent, 24 congruent), for a total of 350 trials (230 congruent, 120 incongruent). To ensure adequate task difficulty, a response deadline was established for each block that corresponded to the 85th percentile of the RT distribution from incongruent trials in the preceding block; in the first block, the practice RT distribution was used for this purpose. Stimulus presentation was followed by a fixation cross (1400 ms). If the participant did not respond by the response deadline, a screen reading TOO SLOW! was presented next (300 ms). Participants were told that if they saw this screen, they should speed up.

If a response was made before the deadline, the TOO SLOW! screen was omitted and the fixation cross remained onscreen for the 300 ms interval. Finally, each trial ended with presentation of the fixation cross for an additional 200-400 ms. Thus, total trial time varied between 2050-2250 ms. The sequence of congruent and incongruent trials was established with optseq2 (http://surfer.nmr.mgh.harvard.edu/optseq/) and was identical across participants.

6.3.4 Quality control

Quality control checks were used to exclude datasets characterized by unusually poor performance. First, for each participant we defined outlier trials as those in which either the raw RT was less than 150 ms or the log-transformed RT exceeded the participants mean±3SD, computed separately for congruent and incongruent stimuli. Second, we excluded datasets with: 35 or more RT outliers (i.e., greater than 10% of trials), fewer than 200 outlier-free congruent trials, fewer then 90 outlier-free incongruent trials, or lower than 50% correct for congruent or incongruent trials. Data from 92 depressed and 37 healthy participants passed these checks and constitute the final sample. Trials characterized by RT outliers were excluded from all analyses.

6.3.5 Analysis of flanker interference effects on accuracy and RT

To investigate effects of flanker interference on accuracy and RT, we computed two linear mixed models using the lme4 package (version 1.0.5) in the R software environment (R Core team, 2013). In the first model, RT was the dependent variable. We expected the depressed group to respond more slowly than controls, particularly in response to incongruent stimuli, and depression has been linked to altered error responses (Chiu & Deldin, 2007). Therefore, we entered a Group x Stimulus x Accuracy interaction and Site as independent variables. In the second model, accuracy was the dependent variable and the independent variables were the Group x Stimulus interaction and Site. Because accuracy was scored as 0 or 1, logistic regression was used for this model. Participant was entered as a random effect in both models.

6.3.6 Computational modeling

Our version of the DDM is an adaptation of the Linear Approach to Threshold with Ergodic Rate model developed for use with the anti-saccade task (Noorani & Carpenter, 2013). As shown in figure 6.1, the model consists of three, single-boundary drift-diffusion (Wald) accumulators that integrate noisy evidence over time with a certain drift-rate: the higher the drift-rate, the faster the accumulation. A response is registered when the drift-process crosses a threshold. While congruent trials only require that a single prepotent accumulator reaches threshold in order to commit a response, incongruent trials are modeled as a race between two accumulators: a prepotent unit that always responds in agreement with the flanking arrows (figure 6.1, top), and an executive control unit that responds according to the central arrow (figure 6.1, bottom). Accumulation of the executive control unit is delayed by a constant time-offset (figure 6.1, bottom left) that simulates additional processes such as the retrieval and application of rules (Wiecki & Frank, 2013). This offset is necessary to model the commonly observed slowing on correct incongruent RTs. The unit that crosses its threshold first determines whether the model commits an error (figure 6.1, top right) or makes the correct response (figure 6.1, bottom right). There is also a third inhibitory control accumulator that acts as a brake, stopping the prepotent accumulator when its threshold is reached (figure 6.1, middle). Thus, the model has the following parameters: a single threshold setting for each accumulator; drift-rates for the prepotent, inhibitory, and executive control accumulators; a delay time to onset for the executive control unit; and a constant, non-decision time capturing motor execution (figure 6.1, upper left).

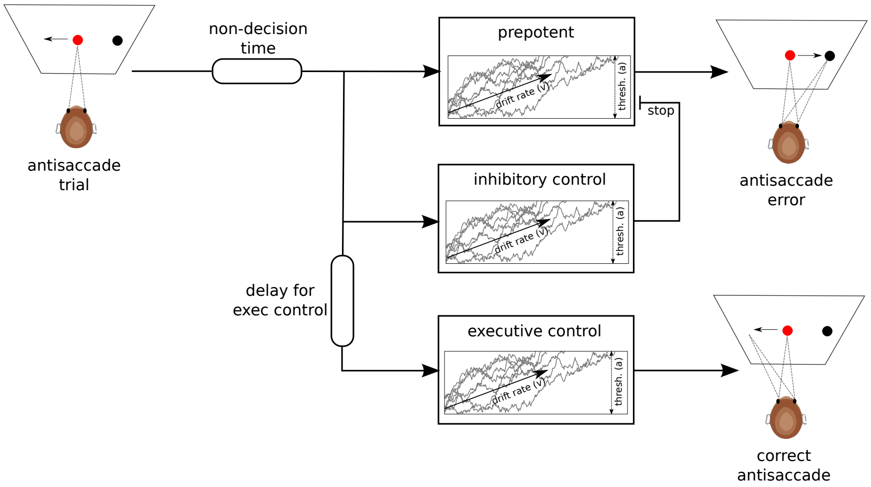

Figure 6.1: Computational model, adapted from LATER model for application to the flanker task.

In this model, accuracy on incongruent trials depends on whether the prepotent or executive control drift-process crosses its threshold first, and RT corresponds to the passage time of the winning accumulator. The model is able to capture the commonly observed pattern of fast error RTs and slower correct RTs on incongruent trials (Noorani & Carpenter, 2013; Ridderinkhof et al., 2010). Intuitively, higher prepotent drift-rate will lead to faster congruent trials as well as more fast errors in incongruent trials, due to the race more often being won by the prepotent accumulator. Higher inhibitory control drift-rates will counteract this by stopping prepotent responses earlier and allowing the executive control process to reach threshold. Increases of executive control drift-rate will lead to shorter correct RTs on incongruent trials.

To find the best-fitting parameters, we used Powell-optimization (Powell, 1964) with basin hopping (Wales & Doye, 1997) to avoid local maxima. Model fit was evaluated by probability density approximation (Turner & Sederberg, 2014), which uses kernel density estimation of samples generated from the above-described process and does not require a closed-form solution of the likelihood function. Weakly informative priors were placed on model parameters to constrain extreme model fits. The model was fit to each participant's full distribution of RT data from congruent and incongruent trials simultaneously, with threshold settings and prepotent drift-rate shared between the two stimulus types.

6.4 Results

6.4.1 Demographics and clinical measures

There was no group difference (ts < 1.1, ps > 0.27) in age (controls: 36.22±14.32; depressed: 39.16±12.99) or years of education (controls: 15.77±4.52; depressed: 15.06±2.43). QIDS-SR16 scores were higher in depressed participants (18.48±2.87) versus controls (1.46±1.30), t(125.04) = -46.29, $p < 0.001$. The mean QIDS-SR16 score in the depressed group indicates moderate depression.

6.4.2 Flanker interference effects

RT. Controls (figure 6.2a) and depressed participants (figure 6.2b) responded more quickly on correct congruent trials versus correct incongruent trials, consistent with flanker interference. Both groups showed the opposite pattern when making errors, generating faster RTs on incorrect incongruent trials versus incorrect congruent trials. This pattern led to a Stimulus x Accuracy interaction, Z = 22.82, $p < 0.001$.

The model also returned a Group x Accuracy interaction, Z = 3.28, p = 0.001, and a Group x Stimulus interaction, Z = 2.05, p = 0.04. Follow-up contrasts linked the Group x Accuracy interaction to a difference on correct trials: when responding correctly, depressed participants were slower than controls, $\chi2(1) = 6.19$, p = 0.03. There was no difference on error trials (p = 0.42). The Group x Stimulus interaction reflected a marginal difference on incongruent trials, with depressed participants responding more slowly than controls, $\chi2(1) = 4.38$, p = 0.07. Depressed participants were also slower on congruent trials, but this difference was not significant (p = 0.24).

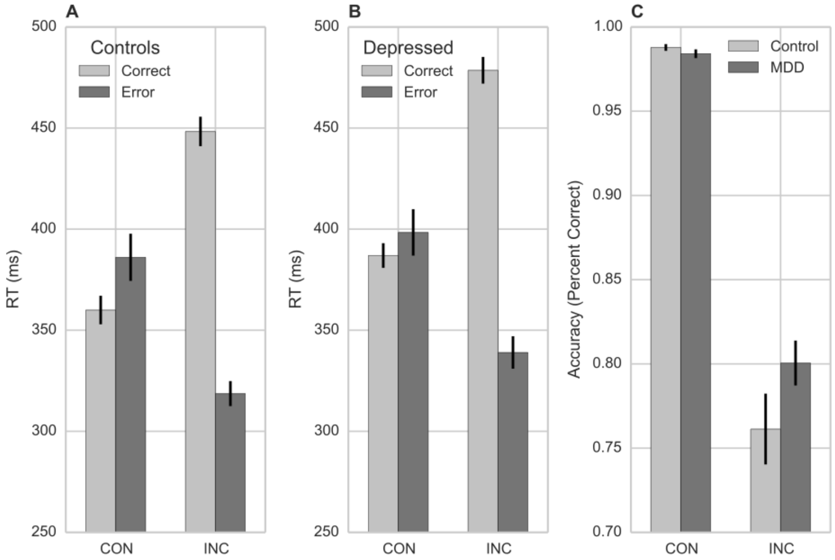

Figure 6.2: Flanker interference effects on (A) RT in controls, (B) RT in depressed participants, and (C) accuracy in both groups.

Thus, depressed participants responded more slowly than controls, with a significant difference emerging for correct responses and a marginal difference for responses to incongruent stimuli. Slow responses on incongruent trials are commonly observed in depressed samples, and they are consistent with executive function deficits (Snyder, 2013).

Accuracy. As shown in figure 6.2c, both groups were more accurate when responding to congruent versus incongruent stimuli, consistent with flanker interference. However, depressed participants were more accurate than controls on incongruent trials, leading to a Group x Stimulus interaction, $Z = 3.86$, $p < 0.001$. Follow-up linear contrasts confirmed a Group effect on incongruent trials, $\chi 2(1) = 13.39$, $p < 0.001$, but not congruent trials ($p = 0.76$). This result echoes reports of better accuracy in sad and anhedonic samples (Au et al., 2013; Dubal et al., 2004).

Model parameter & Healthy Controls & Depressed Participants
Non-decision time & 212 ±34 & 207±59
Prepotent drift-rate* & 7.00±1.45 & 6.37±1.28
Inhibitory drift-rate & 9.76±1.95 & 9.67±2.18
Executive control: drift-rate* & 10.38±2.61 & 9.28±2.80
Executive control: delay to onset & 131.23±25.27 & 138.97±34.99
Threshold & 1.05±0.33 & 1.14±0.44

Table 6.1: Mean (± SD) best fitting parameter values from the drift diffusion model (ms). *Depressed $<$ Controls, p $<$ 0.05.

6.4.3 Computational modeling

Best-fitting parameter values from the DDM are presented in Table 6.1. Independent t-tests revealed that the executive control drift-rate on incongruent trials was lower in depressed relative to healthy participants, t(77) = 2.04, p = 0.04, consistent with sluggish executive function in depression. However, the prepotent drift-rate was also lower in depressed participants, t(77) = 2.40, p = 0.02. This finding is intriguing because if the prepotent bias were weak enough, it could potentially fully offset the executive control deficit, leading to the pattern of slow but accurate responses seen in the data.

To test this hypothesis, we conducted three simulations that involved generating hypothetical RT distributions. Our aim was to isolate the effects of certain parameters on the data. Next, we conducted our first simulation by setting all parameters to the best-fitting values for the controls, as returned by the DDM, and then adjusted only the executive control drift rate to the best-fitting value for the depressed participants. As shown in figure 6.3, this resulted in prolonged incongruent RT but no difference in accuracy. Thus, this simulation did not adequately recapitulate the actual RT data from depressed participants. In the second simulation, we returned the executive control drift rate to the control value but set the prepotent drift rate to the best-fitting value for depressed participants. As can be seen, while this modulation accounts for the increase in accuracy it failed to capture the increased RT in correct incongruent trials. In the third simulation, we set both the executive control and prepotent drift rates to the best-fitting values for the depressed group, leaving all other parameters settings optimized for the controls. As shown in figure 6.3, this yielded the pattern actually observed in the depressed participants: responding was slower overall, but the error rate on incongruent trials is strongly reduced. Thus, this sequence of simulations demonstrates that if prepotent response bias is decreased, highly accurate performance can be observed even if executive control is sluggish. Informally, these results can be conceptualized as showing that poor cognitive control is less problematic if what must be controlled is weaker, at least in the context of the paradigm used here.

6.4.4 Correlation with anhedonia

As shown in figure 6.4, when data from both groups were considered together, we found significant Pearson correlations between anhedonia, as assessed by the SHAPS, and both these drift-rates (prepotent: r[122] = 0.23, $p < 0.007$); executive control: r[122] = 0.28, $p < 0.001$).

6.5 Discussion

This study produced three results. First, responding was slow but accurate in depressed participants. Second, the DDM pointed to sluggish executive control and reduced prepotent response bias in the MDD group, and simulations highlighted that particular combination of parameters as necessary to recapitulate the behavioral results from incongruent trials in the depressed group. Third, executive control and prepotent drift-rates were negatively correlated with anhedonia across groups.

As demonstrated, computational models have the ability to provide a cognitive-level

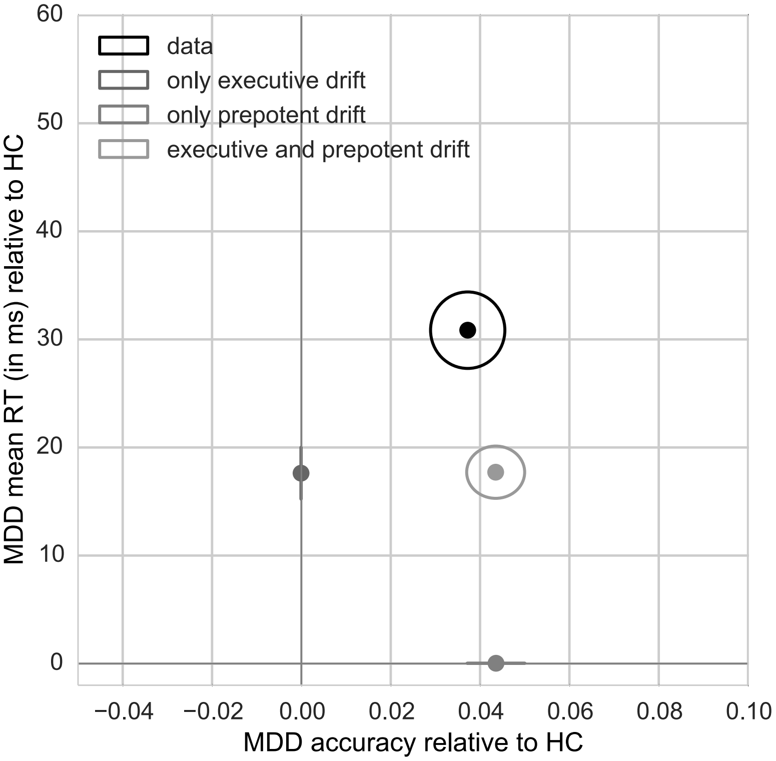

Figure 6.3: Using RT simulations to isolate the effects of particular cognitive processes. Points indicate mean RT of correct incongruent trials (x-axis) and accuracy on incongruent trials (y-axis) of MDD subjects relative to HCs. Ellipses indicate standard error of the mean. Actual RT distributions ('data') generated by the depressed and control groups shows MDD subjects to be slower and more accurate on incongruent trials. Manipulation of the executive control drift rate ('only executive drift') led to slowing in the MDD group but no change in accuracy. Manipulation of the prepotent drift rate ('only prepotent drift') resulted in an increase in accuracy in the MDD group but no change in correct incongruent RT. Simultaneous adjustment of the executive control and prepotent drift rates ('executive and prepotent drift') yielded data that closely captures the specific pattern of increased accuracy and prolonged incongruent RT in depressed participants.

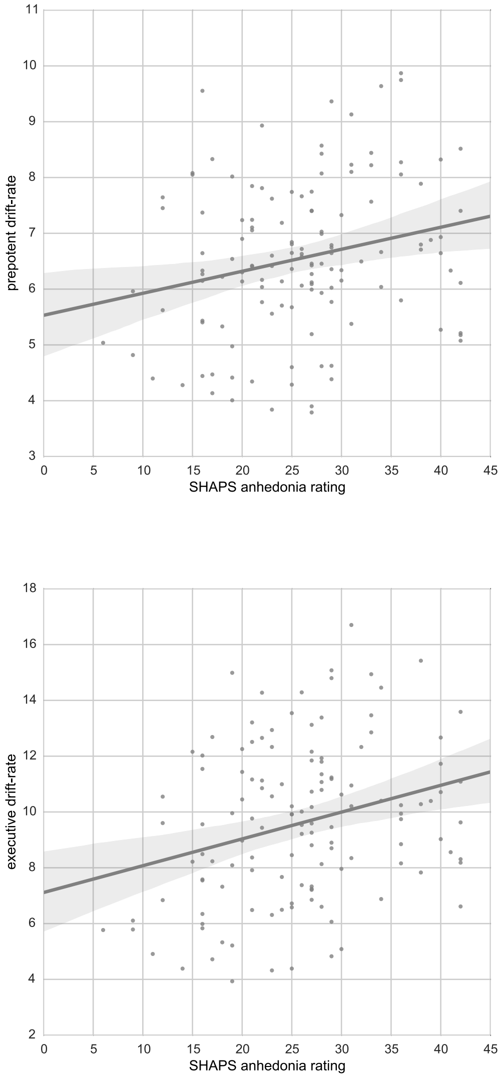

Figure 6.4: Self-reported anhedonia was correlated with the prepotent drift rate (left) and executive control drift rate (right panel) across the two groups. Shaded regions show the 95% confidence interval.

description of performance differences observed in behavioral tasks like the flanker task. Specifically, we provide a plausible explanation for the conundrum that while depressed patients show cognitive deficits in many tasks, they appear to have increased accuracy in the Flanker task. By deconstruction the behavior into cognitive processes like prepotency, inhibition and executive control we confirm that depressed patients do show executive control deficits; however, this deficit that would lead to more errors is more than offset by a simultaneous decrease in prepotent response bias.

Further, by relating cognitive model parameters to underlying neurobiological processes we will hypothesize that tonically reduced striatal dopamine in depression could explain our findings.

6.5.1 Reduced striatal dopamine in depression

A promising neural explanation for reduced drift-rates in the depressed group involves dopaminergic innervation of the striatum, the input structure for the basal ganglia. The basal ganglia gate action plans stored in frontal cortex (Alexander & Crutcher, 1990; Brown et al., 2004; Frank, 2005; Frank et al., 2005; Mink, 1996). Specifically, the selective activation of basal ganglia neurons in the Go and NoGo pathways acts to facilitate or suppress action plans, making their execution more or less likely, respectively (Chevalier & Deniau, 1990; Mink, 1996). This balance between facilitation and suppression is modulated by dopamine, which excites Go neurons but inhibits NoGo neurons, thus increasing the probability that a given action will be executed (Frank, 2005). Conversely, low concentrations of striatal dopamine disinhibit indirect NoGo neurons and result in weak activation of Go neurons, leading to overall response slowing (Wiecki & Frank 2010; Wiecki et al., 2009). Critically, this reduced gating speed can affect habitual actions (parameterized by the prepotent drift-rate) and volitional action (parameterized by the executive control drift-rate) to a similar degree (Wiecki & Frank 2013). Thus, low striatal dopamine may account for reduced prepotent and executive control drift-rates in the depressed group. This is consistent with data from paradigms focused on motivation and reward processing, which have highlighted abnormal striatal dopamine concentration and function in depression (Dillon et al., 2014; Treadway & Zald, 2011). To resolve the paradox highlighted in the Introduction, tonically reduced striatal dopamine may explain how deficient executive function and preserved accuracy can coexist in depression.

However, there is an obvious limitation with this admittedly speculative proposal: it is not clear that reduced prepotent response bias should always offset slow executive control, as it did in this study. In some cases reduced executive control might dominate, yielding responses that are both slow and inaccurate. From a neurobiological perspective, such a pattern might emerge if abnormalities in frontal regions important for retrieving task rules and biasing gating via connections with Go vs. NoGo circuitry are more pronounced than aberrations in striatal dopamine concentrations (Wiecki & Frank, 2013). It is unclear what factors would lead to balanced versus imbalanced deficits in executive function and prepotent response bias, but their identification should be a priority. Otherwise, mixed findings across case/control studies are likely because the proportion of individuals whose neural profile matches one of these two alternatives (i.e., balanced versus imbalanced) may vary substantially across different depressed samples.Furthermore, the fact that correlations between anhedonia and the executive control and prepotent drift-rates emerged across both groups is conceptually consistent with prior findings (Dubal et al., 2000; Dubal & Jouvent, 2004) and underscores the fact that meaningful individual differences in these neurocognitive processes extend beyond clinical samples. In particular, although anhedonia is a marker of psychopathology, variation in hedonic capacity is evident in the healthy population (Meehl, 2001). The current results suggest that hypohedonic individuals are likely to show reduced prepotent response bias and slow executive control.

6.6 Limitations

This study benefited from a large, unmedicated depressed sample collected at four sites and the use of computational tools to isolate specific cognitive processes. However, important limitations must be mentioned. First, negative effects of depression are strongest in unconstrained tasks. When participants are told what to do and when to do it, effects of depression are typically weakened (Dillon and Pizzagalli, 2013; Ehring et al., 2010). The flanker task features clear instructions and few response options, and participants need not spontaneously generate plans or explore novel options. Consequently, it may be less sensitive to depression than tasks with those attributes.

Second, the use of brief stimulus durations and individually-titrated response deadlines may have limited our ability to detect effects of depression because they provide little opportunity for mind-wandering, minimizing the impact of rumination on performance. Because rumination is a robust correlate of depression and a sign of poor executive control, future flanker studies might benefit from longer stimulus durations and more lax response deadlines.

Finally, while the computational model used here has been validated on the related antisaccade task by Noorani & Carpenter (2013), other models have been successfully applied to the flanker task (e.g., Hbner et al. 2010; White et al., 2011). The relationship between these models is not well-established, and they might suggest negative effects of depression on different parameters (e.g., response threshold). Ultimately, studying relationships between these models and the underlying neurobiology may prove helpful for adjudicating between them, because the neurobiology of depression may render certain models more plausible than others.

6.7 Conclusions

Depressed participants responded more slowly but also more accurately than controls in the flanker task, extending prior studies that have found similar patterns in smaller depressed samples (Holmes & Pizzagalli, 2010; Siegle et al., 2004). Because depression impairs executive function, highly accurate performance has been difficult to explain. The current study used computational modeling to provide new insight. Specifically, reduced prepotent response bias offset slow executive control in our depressed sample. Data from neural network simulations (Wiecki & Frank, 2013) and the larger literature indicate that both these abnormalities may reflect tonically reduced striatal dopamine. The fact that anhedonia was negatively correlated with the prepotent and executive control drift-rates across healthy and depressed participants suggests that similar performance on cognitive control tasks may be found in hypohedonic individuals who do not meet a clinical diagnosis.

Chapter 7

Limitations and future directions

One central quantitative limitation of chapters 4, **??** and 6 are the divergence from the hierarchical Bayesian estimation presented and applied in chapters 1 and 3. The reasons for this short-coming are purely technical. While a closed-form solution that is relatively easy to evaluate exists for the DDM, no such formula could be established for the SIDDM. As such, to evaluate this likelihood we had to revert to Monte-Carlo simulation which is computationally more expensive and introduces approximation noise. These factors severely complicated the use of already computationally costly MCMC sampling algorithms. We thus had to revert to non-hierarchical MAP optimization to fit the models. Recent progress in approximate Bayesian computation (ABC) (see Turner and Van Zandt (2012) for a tutorial), like the use of kernel approximations to reduce the number of required likelihood evaluations (Meeds and Welling) appear as promising methods to remedy this short-coming.

We had also placed hope in the use of clustering methods like Bayesian non-parametrics that simultaneously determine the clustering of data points as well as the number of clusters from the data, as described in appendix A. While these methods continue to look very promising in regards to establishing new disease boundaries based on cognitive functional profiles we had limited success on the real-world data sets. It is likely that even though the number of subjects in our data sets was comparatively high for psychology studies, they are still orders of magnitude too small for meaningful inference of the distribution of cognitive functional profiles of the healthy and clinical population. To remedy this short-coming we need to design and carry out large clinical studies that test thousands of subjects with a wide array of mental disorders on cognitive tasks from various domains.

In conclusion, while computational psychiatry is still in its infancy, the statistical tools described in this thesis, in combination with the appropriate data sets, show great promise to move psychiatry away from subjective questionnaire based disease classification towards a quantitative medicine that diagnoses and treats dysfunctions of the neurocircuitry rather than symptoms.

Appendix A

Mathematical details: Computational Psychiatry

The following serves as a reference for the mathematical details of the methods motivated above.

A.1 Parameters used in simulation study

The below table contains the group means of the parameters used to create subjects of two groups. Each individual subject was created by adding normally distributed noise of $\sigma = .1$ to the group mean.

Parameter	Group 1	Group 2
non-decision time	.3	.25
drift-rate	1	1.2
threshold	2	2.2

A.2 Drift-Diffusion Model

Mathematically, the DDM is defined by a stochastic differential equation called the Wiener process with drift:

$$dW \sim \mathcal{N}(v, \sigma^2) \tag{A.1}$$

where v represents the drift-rate and σ the variance. As we often only observe the response times of subjects we are interested in the wiener first passage time (wfpt) – the time it takes W to cross one of two boundaries. Assuming two absorbing boundaries of this process and through some fairly sophisticated math (see e.g. Smith, 2000) it is possible to analytically derive the time this process will first pass one of the two boundaries (i.e. the wiener first passage time; wfpt). This probability distribution[1] then serves as the likelihood function for the DDM.

A.3 Bayesian Inference

A.3.1 Hierarchical Bayesian modeling

Bayesian methods require specification of a generative process in form of a likelihood function that produced the observed data x given some parameters θ. By specifying our prior belief we can use Bayes formula to invert the generative model and make inference on the probability of parameters θ:

$$P(\theta|x) = \frac{P(x|\theta) * P(\theta)}{P(x)} \tag{A.2}$$

[1] the wfpt will not be a distribution rather than a single value because of the stochasticity of the wiener process

Where $P(x|\theta)$ is the likelihood and $P(\theta)$ is the prior probability. Computation of the marginal likelihood $P(x)$ requires integration (or summation in the discrete case) over the complete parameter space Θ:

$$P(x) = \int_{\Theta} P(x, \theta) \, \mathrm{d}\theta \tag{A.3}$$

Note that in most scenarios this integral is analytically intractable. Sampling methods like Markov-Chain Monte Carlo (MCMC) (Gamerman and Lopes, 2006) circumvent this problem by providing a way to produce samples from the posterior distribution. These methods have been used with great success in many different scenarios (Gelman et al., 2003) and will be discussed in more detail below.

A hierarchical model has a particular benefit to cognitive modeling where data is often scarce. We can construct a hierarchical model to more adequately capture the likely similarity structure of our data. As above, observed data points of each subject $x_{i,j}$ (where $i = 1, \ldots, S_j$ data points per subject and $j = 1, \ldots, N$ for N subjects) are distributed according to some likelihood function $f|\theta$. We now assume that individual subject parameters θ_j are normal distributed around a group mean with a specific group variance ($\lambda = (\mu, \sigma)$ with hyperprior G_0) resulting in the following generative description:

$$\mu, \sigma \sim G_0() \tag{A.4}$$

$$\theta_j \sim \mathcal{N}(\mu, \sigma^2) \tag{A.5}$$

$$x_{i,j} \sim f(\theta_j) \tag{A.6}$$

See figure A.1 for the corresponding graphical model description.

Another way to look at this hierarchical model is to consider that our fixed prior on θ

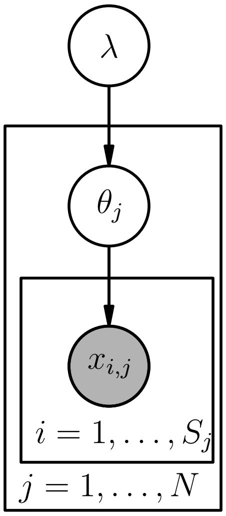

Figure A.1: Graphical notation of a hierarchical model. Circles represent continuous random variables. Arrows connecting circles specify conditional dependence between random variables. Shaded circles represent observed data. Finally, plates around graphical nodes mean that multiple identical, independent distributed random variables exist.

from formula (A.2) is actually a random variable (in our case a normal distribution) parameterized by λ which leads to the following posterior formulation:

$$P(\theta, \lambda|x) = \frac{P(x|\theta) * P(\theta|\lambda) * P(\lambda)}{P(x)} \tag{A.7}$$

Note that we can factorize $P(x|\theta)$ and $P(\theta|\lambda)$ due to their conditional independence. This formulation also makes apparent that the posterior contains estimation of the individual subject parameters θ_j and group parameters λ.

A.3.2 Empirical Bayesian Approximation

Empirical Bayes can be regarded as an approximation of equation (A.7). To derive this approximation consider $P(\theta|x)$ which we can calculate by integrating over $P(\lambda)$:

$$P(\theta|x) = \frac{P(x|\theta)}{P(x)} \int P(\theta|\lambda)P(\lambda)\,\mathrm{d}\lambda \tag{A.8}$$

Now, if the true distribution $P(\theta|\lambda)$ is sharply peaked, the integral can be replaced with the point estimate of its peak $\lambda^{\star}$:

$$P(\theta|x) \simeq \frac{P(x|\theta)P(\theta|\lambda^{\star})}{P(x|\lambda^{\star})} \tag{A.9}$$

Note, however, that $\lambda^{\star}$ depends itself on $P(\theta|x)$. One algorithm to solve this interdependence is Expectation Maximization (EM) (Dempster et al., 1977). EM is an iterative algorithm that alternates between computing the expectation of $P(\theta|x)$ (this can be easily done by Laplace Approximation (Azevedo-filho and Shachter, 1994)) and then maximizing the prior point estimate $\lambda^{\star}$ based on the current values obtained by the expectation step. This updated point estimate is then used in turn to recompute the expectation. The algorithm is run until convergence or some other criterion in reached. This approach is used for example by Huys et al. (2012b) to fit their reinforcement learning models.

A.3.3 Markov-Chain Monte-Carlo

As mentioned above, the posterior is often intractable to compute analytically. While Empirical Bayes provides a useful approximation, an alternative approach is to estimate the full posterior by drawing samples from it. One way to achieve this is to construct a Markov-Chain that has the same equilibrium distribution as the posterior (Gamerman and Lopes, 2006). Algorithms of this class are called Markov-Chain Monte Carlo (MCMC) samplers.

One common and widely applicable algorithm is Metropolis-Hastings (Chib and Greenberg, 1995; Andrieu et al., 2003). Assume we wanted to generate samples θ from the posterior $p(\theta|x)$. In general, we can not sample from $p(\theta|x)$ directly. Metropolis-Hastings instead generates samples θ^t from a proposal distribution $q(\theta^t|\theta^{t-1})$ where the next position θ^t only depends on the previous position at θ^{t-1} (i.e. the Markov-property). For simplicity we will assume that this proposal distribution is symmetrical; i.e. $q(\theta^t|\theta^{t-1}) = q(\theta^{t-1}|\theta^t)$. A common choice for the proposal distribution is the Normal distribution, formally:

$$\theta^t \sim \mathcal{N}(\theta^{t-1}, \sigma^2) \tag{A.10}$$

The proposed jump to θ^t is then accepted with probability α:

$$\alpha = \min(1, \frac{p(\theta^t|x)}{p(\theta^{t-1}|x)}) \tag{A.11}$$

In other words, the probability of accepting a jump depends on the probability ratio of the proposed jump position θ^t to the previous position θ^{t-1}. Critically, in this probability ratio, the intractable integral in the denominator (i.e. $p(x) = \int p(x, \theta)\, \mathrm{d}\theta$) cancels out. This can be seen by applying Bayes formula (A.2):

$$\frac{p(\theta^t|x)}{p(\theta^{t-1}|x)} = \frac{\frac{p(x|\theta^t)p(\theta^t)}{p(x)}}{\frac{p(x|\theta^{t-1})p(\theta^{t-1})}{p(x)}} = \frac{p(x|\theta^t)p(\theta^t)}{p(x|\theta^{t-1})p(\theta^{t-1})} \tag{A.12}$$

Thus, to calculate the probability of accepting a jump we only have to evaluate the likelihood and prior, *not* the intractable posterior.

Note that θ^0 has to be initialized at some position and can not directly be sampled from the posterior. From this initial position, the Markov chain will explore other parts of the parameter space and only gradually approach the posterior region. The first samples generated are thus not from the true posterior and are often discarded as “burn-in”. Note moreover that once the algorithm reaches a region of high probability it will continue to explore lower probability regions in the posterior, albeit with lower frequency. This random-walk behavior is due to the probability ratio α which allows Metropolis-Hastings to also sometimes accept jumps from a high probability position to a low probability position.

Another common algorithm is Gibbs sampling that iteratively updates each individual random variable conditional on the other random variables set to their last sampled value (e.g Frey and Jojic, 2005). Starting at some configuration θ^0, the algorithm makes T iterations over each random variable θ_i. At each iteration t each random variable is sampled conditional on the current $(t-1)$ value of all other random variables that it depends on:

$$\theta_i^t \sim p(\theta_i^{(t)} | \theta_{i \neq j}^{(t-1)}) \tag{A.13}$$

Critically, $\theta_{i \neq j}^{(t-1)}$ are treated as constant. The sampled value of $\theta_i^{(t)}$ will then be treated as fixed while sampling the other random variables.

Note that while Gibbs sampling never rejects a sample (which often leads to faster convergence and better mixing), in contrast to Metropolis-Hastings, it does require sampling from the conditional distribution which is not always tractable.

A.4 Likelihood free methods

Several likelihood-free methods have emerged in the past (for a review, see Turner and Van Zandt (2012)). Instead of an analytical solution of the likelihood function, these methods require a sampling process that can simulate a set of data points from a generative model for each θ. We will call the simulated data y and the observed data x. Approximate Bayesian Computation (ABC) relies on a distance measure $\rho(x, y)$ that compares how similar the simulated data y is to the observed data x

(commonly, this distance measure relies on summary statistics). We can then use the Metropolis-Hastings algorithm introduced before and change the acceptance ration α (A.11) to use $\rho(x,y)$ instead of a likelihood function.

$$\alpha = \begin{cases} \min(1, \frac{p(\theta^t)}{p(\theta^{t-1})}) & \text{if } \rho(x,y) \leq \epsilon_0 \\ 0 & \text{if } \rho(x,y) \geq \epsilon_0 \end{cases} \tag{A.14}$$

where ϵ_0 is an acceptance threshold. Large ϵ_0 will result in higher proposal acceptance probability but a worse estimation of the posterior while small ϵ_0 will lead to better posterior estimation but slower convergence.

An alternative approach to ABC is to construct a synthetic likelihood function based on summary statistics (Wood, 2010). Specifically, we sample N_r multiple data sets $y_{1,\ldots,N_r}$ from the generative process. We then compute summary statistics $s_{1,\ldots,N_r}$ for each simulated data set[2]. Based on these summary statistics we then construct the synthetic likelihood function to evaluate θ (see figure A.2 for an illustration):

$$p(x|\theta) \simeq \mathcal{N}(S(x); \mu_\theta, \Sigma_\theta) \tag{A.15}$$

This synthetic likelihood function based on summary statistics can then be used as a drop-in replacement for e.g. the Metropolis-Hastings algorithm outlined above.

A.5 Model Comparison

Computational models often allow formulation of several plausible accounts of cognitive behavior. One way to differentiate between these various plausible hypotheses as expressed by alternative models is model comparison: which of several alternative

[2] The summary statistics must (i) be sufficient and (ii) normally distributed

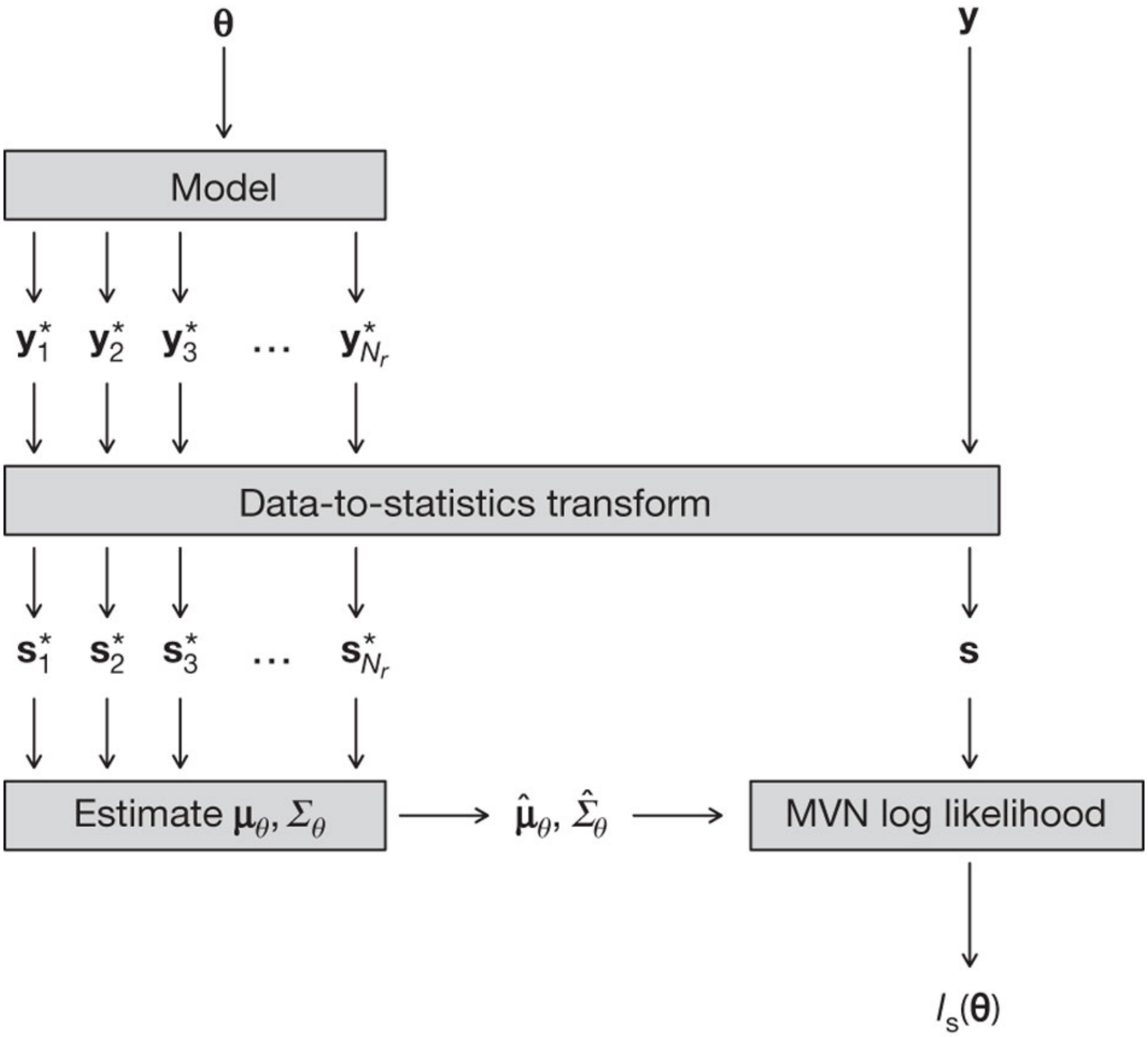
θ
y
Model
$\mathbf{y}_1^*$ $\mathbf{y}_2^*$ $\mathbf{y}_3^*$ … $\mathbf{y}_{N_r}^*$
Data-to-statistics transform
$\mathbf{s}_1^*$ $\mathbf{s}_2^*$ $\mathbf{s}_3^*$ … $\mathbf{s}_{N_r}^*$
s
Estimate $\mu_\theta, \Sigma_\theta$
$\hat{\mu}_\theta, \hat{\Sigma}_\theta$
MVN log likelihood
$l_s(\theta)$

Figure A.2: Construction of a synthetic likelihood. To evaluate parameter vector θ, N_r data sets $y_{1,\ldots,N_r}$ are sampled from the generative model. On each sampled data set summary statistics $s_{1,\ldots,N_r}$ are computed. Based on these summary statistics a multivariate normally distribution is approximated with mean μ_θ and covariance matrix Σ_θ. The likelihood is approximated by evaluating summary statistics of the actual data on the log normal distribution with the estimated μ_θ and Σ_θ. Reproduced from (Wood, 2010).

models provides the best explanation of the data? In the following we review various methods and metrics to compare hierarchical models. The most critical property for model comparison is that model complexity gets penalized because more complex models have greater degrees of freedom and could thus overfit data. Several model comparison measures have been devised.

A.5.1 Deviance Information Criterion

The Deviance Information Criterion (DIC) is a measure which trades off model complexity and model fit (Spiegelhalter et al., 2002b). Several similar measures exist such as Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). However, both these measures use the number of parameters as a proxy for model complexity. While a reasonable approximation to the complexity of non-hierarchical models, the relationship between model parameters (some of which are latent) and complexity in hierarchical models is more intricate. The DIC measure instead infers the number of parameters from the posterior. The DIC is computed as follows:

$$\mathrm{DIC} = \bar{D} + pD \tag{A.16}$$

where

$$pD = \bar{D} - \hat{D} \tag{A.17}$$

$\bar{D}$ is the posterior mean of the deviance (i.e. $-2 * \log(\text{likelihood})$) and $\hat{D}$ is a point estimate of the deviance obtained by substituting in the posterior means. Loosely, $\bar{D}$ represents how well the model fits the data on average while $\hat{D}$ captures the deviance at the best fitting parameter combination. pD then acts as a measure related to the posterior variability and used as a proxy for the effective number of parameters.

Complex models with many parameters will tend to have higher posterior variability and thus result in increased pD penalization.

Note that the only parameters that affect $\hat{D}$ directly in our hierarchical model (equation A.7) are the subject parameters θ_i. Thus, DIC estimates model fit based on how well individual subjects explain the observed data.

A.5.2 BIC

The Bayesian Information Criterion (BIC) is defined as follows:

$$\text{BIC} = -2 * \log p(x|\hat{\theta}^{ML}) + k * \log(n) \tag{A.18}$$

where k is the number of free parameters, n is the number of data points, x is the observed data and $\log p(x|k)$ is the likelihood of the parameters given the data (Schwarz, 1978).

While BIC can not directly be applied to hierarchical models (as outlined above), it is possible to integrate out individual subject parameters (e.g. Huys et al., 2012b):

$$\log p(x|\hat{\theta}^{ML}) = \sum_i \log \int p(x_i|h)p(h|\hat{\theta}^{ML})\,\mathrm{d}h \tag{A.19}$$

where x_i is the data belonging to the ith subject. The resulting score is called integrated BIC.

Since the subject parameters are integrated out, integrated BIC estimates how well the group parameters are able to explain the observed data.

A.5.3 Bayes Factor

Another measure to compare two models is the Bayes Factor (BF) (Kass and Raftery, 1995). It is defined as the ratio between the marginal model probabilities of the two models:

$$BF = \frac{p(x|M_1)}{p(x|M_2)} = \frac{\int p(\theta_1|M_1)p(x|\theta_1, M_1)\,\mathrm{d}\theta_1}{\int p(\theta_2|M_2)p(x|\theta_2, M_2)\,\mathrm{d}\theta_2} \tag{A.20}$$

The magnitude of this ratio informs the degree one should belief in one model compared to the other.

As BF integrates out subject *and* group parameters this model comparison measure should be used when different classes of models are to be compared in their capacity to explain observed data.

A.6 Mixture Models

A.6.1 Gaussian Mixture Models

Mixture models infer k number of clusters in a data set. The assumption of normally distributed clusters leads to a Gaussian Mixture Model (GMM) with a probability density function as follows:

$$p(x|\pi, \mu_{1,\ldots,K}, \sigma_{1,\ldots,K}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k, \sigma_k^2) \tag{A.21}$$

Each observed data point x_i can be created by drawing a sample from the normal distribution selected by the unobserved indicator variable z_i which itself is distributed according to a multinomial distribution π:

$$\mu_k, \sigma_k \sim \mathrm{G}_0() \tag{A.22}$$

$$z_i \sim \pi \tag{A.23}$$

$$x_i \sim \mathcal{N}(\mu_{z_i}, \sigma^2_{z_i}) \tag{A.24}$$

where the base measure G_0 defines the prior for μ_k and σ_k. To simplify the inference it is often advisable to use a conjugate prior for these paramters. For example, the normal distribution is the conjugate prior for a normal distribution with known variance:

$$\mu_k \sim \mathcal{N}(\mu_0, \sigma_0) \tag{A.25}$$

In a similar fashion, we can assign the mixture weights a symmetric Dirichlet prior:

$$\pi \sim \mathrm{Dir}(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}) \tag{A.26}$$

Note that the GMM assumes a mixture distribution on the level of the observed data x_i. However, in our relevant case of a multi-level hierarchical model we need to place the mixture at the level of the latent subject parameters instead of the observed data. As before, we use the subject index $j = 1, \ldots, N$.

$$\mu_k, \sigma_k \sim \mathrm{G}_0() \tag{A.27}$$

$$\pi \sim \mathrm{Dir}(\alpha) \tag{A.28}$$

$$z_j \sim \mathrm{Categorical}(\pi) \tag{A.29}$$

$$\theta_j \sim \mathcal{N}(\mu_{z_j}, \sigma^2_{z_j}) \tag{A.30}$$

$$x_{i,j} \sim \mathrm{f}(\theta_j) \tag{A.31}$$

Where f denotes the likelihood function.

Interestingly, the famous K-Means clustering algorithm is identical to a Gaussian Mixture Model (GMM) in the limit $\sigma^2 \to 0$ (Kulis et al., 2012). K-Means is an expectation maximization (EM) algorithm that alternates between an expectation step during which data points are assigned to their nearest cluster centroids and a maximization step during which new cluster centroids are estimated. This algorithm is repeated until convergence is reached (i.e. no points are reassigned to new clusters).

A.6.2 Dirichlet Process Gaussian Mixture Models

Dirichlet processes Gaussian mixture models (DPGMMs) belong to the class of Bayesian non-parametrics (Antoniak, 1974). They can be viewed as a variant of GMMs with the critical difference that they assume an infinite number of potential mixture components (see Gershman and Blei (2012) for a review). Such mixture models can infer sub-groups when the data is heterogeneous as is generally the case in patient populations. While the mindset describing these methods was their application towards the SSM, their applicability is much more general than that. For example, the case-studies described above which used, among others, RL models to identify differences between HC and psychiatric patients could easily be embedded into this hierarchical Bayesian mixture model framework we outlined here. Such a combined model would estimate model parameters and identify subgroups simultaneously. There are multiple benefits to such an approach. First, computational models fitted via hierarchical Bayesian estimation provide a tool to accurately describe the neurocognitive functional profile of individuals. Second, the mixture model approach is ideally suited to deal with the heterogeneity in patients but also healthy controls (Fair et al., 2012). Third, by testing psychiatric patients with a range of diagnoses (as opposed to most previous research studies that only compare patients with a single diagnosis, e.g. SZ, to controls) we might be able to identify shared pathogenic cascades as suggested by Buckholtz and Meyer-Lindenberg (2012).

$$p(x|\pi, \mu_{1,\dots,\infty}, \sigma_{1,\dots,\infty}) = \sum_{k=1}^{\infty} \pi_k \mathcal{N}(x_i|\mu_k, \sigma_k^2) \tag{A.32}$$

As above, we specify our generative mixture model:

$$\mu_k, \sigma_k \sim \mathrm{G}_0() \tag{A.33}$$

$$z_i \sim \mathrm{Categorical}(\pi) \tag{A.34}$$

$$x_i \sim \mathcal{N}(\mu_{z_i}, \sigma_{z_i}^2) \tag{A.35}$$

with the critical difference of replacing the hyperprior π with the *stick breaking process* (Sethuraman, 1991):

$$\pi \sim \mathrm{StickBreaking}(\alpha) \tag{A.36}$$

The stick-breaking process is a realization of a Dirichlet process (DP). Specifically, $\pi = \{\pi_k\}_{k=1}^{\infty}$ is an infinite sequence of mixture weights derived from the following

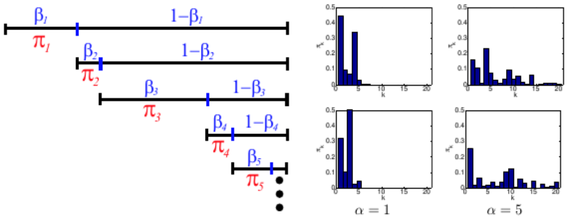

Figure A.3: Left: Stick-breaking process. At each iteration (starting from the top) a π is broken off with relative length $\sim$ Beta$(1, \alpha)$. Right: Histogram over different realizations of the stick-breaking process. As can be seen, higher values of hyperprior α lead to a more spread out distribution. Taken from Eric Sudderth's PhD thesis.

process:

$$\beta_k \sim \text{Beta}(1, \alpha) \tag{A.37}$$

$$\pi_k \sim \beta_k * \prod_{l=1}^{k-1}(1 - \beta_l) \tag{A.38}$$

with $\alpha > 0$. See figure A.3 for a visual explanation.

The Chinese Restaurant Process (CRP) – named after the apparent infinite seating capacity in Chinese restaurants – allows for a more succinct model formulation. Consider that customers z_i are coming into the restaurant and are seated at table k with probability:

$$p(z_i = k | z_{1,\ldots,n-1}, \alpha, K) = \frac{n_k + \alpha/K}{n - 1 + \alpha}$$

where $k = 1 \ldots K$ is the table and n_k is the number of customers already sitting at table k (see figure A.4 for an illustration). It can be seen that in the limit as $K \to \infty$ this expression becomes:

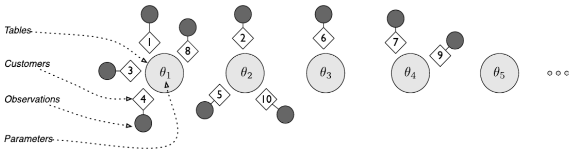

Figure A.4: Illustration of the Chinese Restaurant Process. Customers are seated at tables with parameters θ. The more customers are already seated at a table, the higher the probability that future customers are seated at the same table (i.e. clustering property). Taken from Gershman and Blei (2012).

$$p(z_i = k | z_{1,\ldots,n-1}, \alpha) = \frac{n_k}{n - 1 + \alpha}$$

Thus, as customers are social, the probability of seating customer z_i to table k is proportional the number of customers already sitting at that table. This desirable clustering property is also known as the "rich get richer".

Note that for an individual empty table k at which no customer has been seated (i.e. $n_k = 0$) the probability of seating a new customer to that table goes to 0 in the limit as $K \to \infty$. However, at the same time the number of empty tables approaches infinity. Consider that we have so far seated L customers to tables and the set $\mathbf{Q}$ contains all empty tables such that there are $|\mathbf{Q}| = K - L$ empty tables in the restaurant. The probability of seating a customer z_i at an empty table becomes:

$$p(z_i \in \mathbf{Q} | \mathbf{z}_{1,\ldots,n-1}, \alpha) = \frac{\alpha}{n - 1 + \alpha}$$

As can be seen, the probability of starting a new table is proportional to the concentration parameter α. Intuitively, large values of the dispersion parameter α lead to more clusters being used.

Thus, while the Stick-Breaking process sampled mixture weights from which we had to infer cluster assignments, the CRP allows for direct sampling of cluster assignments.

The resulting model can then be written as:

$$\mu_k, \sigma_k \sim \mathrm{G}_0() \tag{A.39}$$

$$z_{1,\ldots,N} \sim \mathrm{CRP}(\alpha) \tag{A.40}$$

$$x_i \sim \mathcal{N}(\mu_{z_i}, \sigma^2_{z_i}) \tag{A.41}$$

Finally, in a hierarchical group model we would need to place the infinite mixture on the subject level rather than the observed data level:

$$\mu_k, \sigma_k \sim \mathrm{G}_0() \tag{A.42}$$

$$z_j \sim \mathrm{CRP}(\alpha) \tag{A.43}$$

$$\theta_j \sim \mathcal{N}(\mu_{z_j}, \sigma^2_{z_j}) \tag{A.44}$$

$$x_{i,j} \sim \mathrm{F}(\theta_j) \tag{A.45}$$

See figure A.5 for a graphical model description.

Note that while the potential number of clusters is infinite, any realization of this process will always lead to a finite number of clusters as we always have finite amounts of data. However, this method allows the addition (or subtraction) of new clusters as new data becomes available.

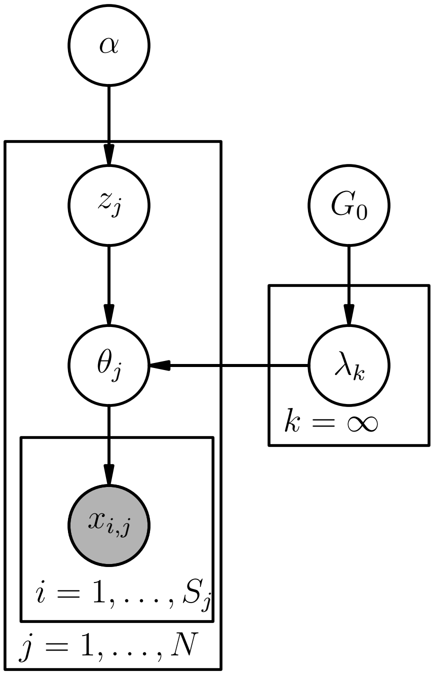

Figure A.5: Graphical model representation of the hierarchical Dirichlet process mixture model. Group parameters $\lambda_k = (\mu_k, \sigma_k)$. See text for details.

Appendix B

Mathematical details: Neural network model of response inhibition

B.1 Software

The model and the Python scripts are available at `http://ski.clps.brown.edu/BG_Projects/`.

B.2 Implementation details

Like the original Frank (2006) model, this model is implemented in the Emergent neural modeling software framework (Aisa et al., 2008), which can be downloaded here:

`http://grey.colorado.edu/emergent/index.php/Main_Page`.

Emergent measures simulator time in cycles. Here, we convert this time to ms by multiplying cycles by 4 to roughly match behavioral and electrophysiological data.

Emergent uses point neurons with excitatory, inhibitory, and leak conductances contributing to an integrated membrane potential, which is then thresholded and transformed via an $\frac{x}{x+1}$ sigmoidal function to produce a rate code output communicated to other neurons (discrete spiking can also be used, but produces noisier results).

The membrane potential V_m is a function of ionic conductances g with reversal (driving) potentials E as follows:

$$\triangle V_m(t) = \tau \sum_c g_c(t)\overline{g_c}(E_c - V_m(t)) \tag{B.1}$$

with 3 channels (c) corresponding to: e excitatory input; l leak current; and i inhibitory input. Following electrophysiological convention, the overall conductance is decomposed into a time-varying component $g_c(t)$ computed as a function of the dynamic state of the model, and a constant $\overline{g_c}$ that controls the relative influence of the different conductances. The equilibrium potential can be written in a simplified form by setting the excitatory driving potential (E_e) to 1 and the leak and inhibitory driving potentials (E_l and E_i) of 0:

$$V_m^\infty = \frac{g_e\overline{g_e}}{g_e\overline{g_e} + g_l\overline{g_l} + g_i\overline{g_i}} \tag{B.2}$$

which shows that the neuron is computing a balance between excitation and the opposing forces of leak and inhibition. This equilibrium form of the equation can be understood in terms of a Bayesian decision making framework (O'Reilly and Munakata, 2000).

The excitatory net input/conductance $g_e(t)$ or η_j is computed as the proportion of open excitatory channels as a function of sending activations times the weight values:

$$\eta_j = g_e(t) = \langle x_i w_{ij} \rangle = \frac{1}{n} \sum_i x_i w_{ij} \tag{B.3}$$

The inhibitory conductance can either be computed by the kWTA function described in the next section or by modeling inhibitory interneurons. Leak is a constant. Activation communicated to other cells (y_j) is a thresholded (Θ) sigmoidal function of the membrane potential with gain parameter γ:

$$y_j(t) = \frac{1}{\left(1 + \frac{1}{\gamma[V_m(t) - \Theta]_+}\right)} \tag{B.4}$$

where $[x]_+$ is a threshold function that returns 0 if $x0$ and x if $x0$. To avoid dividing by 0 we assume $y_j(t) = 0$ if it returns 0. This activation is subject to scaling factors (wt_scale.abs and wt_scale.rel) which modify how much impact the projections have on the post-synaptic neurons.

B.3 Inhibition within and between layers

Inhibition between layers (i.e. for GABAergic projections between BG layers and striatal inhibitory interneurons) is achieved via simple unit inhibition, where the inhibitory current g_i for the unit is determined from the net input of the sending unit. For *within* layer lateral inhibition (used here in premotor cortex), Leabra uses a kWTA (k-Winners-Take-All) function to achieve inhibitory competition among neurons within each layer (area). The kWTA function computes a uniform level of inhibitory current for all neurons in the layer, such that the k + 1th most excited unit within a layer is generally below its firing threshold, while the kth is typically above threshold. Activation dynamics similar to those produced by the kWTA function have been shown to result from simulated inhibitory interneurons that project both feedforward and feedback inhibition (O'Reilly and Munakata, 2000). Thus, although the kWTA function is somewhat biologically implausible in its implementation (e.g., requiring global information about activation states and using sorting mechanisms), it provides a computationally effective approximation to biologically plausible inhibitory dynamics. kWTA is computed via a uniform level of inhibitory current for all neurons in the layer as follows:

$$g_i = g^{\Theta}_{k+1} + q(g^{\Theta}_{k} - g^{\Theta}_{k+1}) \tag{B.5}$$

where $0q1$ (0.25 default) is a parameter Θ for setting the inhibition between the upper bound of g_k and Θ . These boundary inhibition values are the lower bound of g_{k+1} computed as a function of the level of inhibition necessary to keep a unit right at threshold:

$$g_i = g^{\Theta}_{k+1} + q(g^{\Theta}_{k} - g^{\Theta}_{k+1}) \tag{B.6}$$

In the basic version of the kWTA function, which is relatively rigid about the kWTA constraint and is therefore used for output layers, g^{Θ}_{k} and g^{Θ}_{k+1} are set to the threshold inhibition value for the kth and k+1th most excited neurons, respectively. Thus, the inhibition is placed exactly to allow k neurons to be above threshold, and the remainder below threshold. For this version, the q parameter is almost always .25, allowing the kth unit to be sufficiently above the inhibitory threshold.

The premotor cortex uses the average-based kWTA version, g^{Θ}_{k} is the average g^{Θ}_{i} value for the top k most excited neurons, and g^{Θ}_{k+1} is the average of g^{Θ}_{i} for the remaining $n-k$ neurons. This version allows for more flexibility in the actual number of neurons active depending on the nature of the activation distribution in the layer and the value of the q parameter (which is typically .6), and is therefore used for hidden layers.

Hysterisis and Accommodation

$$I_a(t) = g_a(t)\bar{g_a}(V_m(t) - E_a) \tag{B.7}$$

$$I_h(t) = g_h(t)\bar{g_h}(V_m(t) - E_h) \tag{B.8}$$

E_h is excitatory; E_a inhibitory.

g_a and g_h are time-varying functions that depend on previous activity, integrated over different time periods.

$$g_a(t) = \begin{cases} g_a(t-1) + dt_{g_a}(1 - g_a(t-1)); & \text{if}(b_a(t) = \Theta_a) \\ g_a(t-1) + dt_{g_a}(0 - g_a(t-1)); & \text{if}(b_a(t) = \Theta_d) \end{cases} \tag{B.9}$$

B.4 Computation of conflict

dACC activity is the Hopfield energy of pre-SMA:

$$\text{dACC}_{\text{act}} = \text{FEF}_{\text{left}_{\text{act}}} * \text{FEF}_{\text{right}_{\text{act}}} \tag{B.10}$$

Bibliography

Désirée S Aichert, Nicola M Wöstmann, Anna Costa, Christine Macare, Johanna R Wenig, Hans-Jürgen Möller, Katya Rubia, and Ulrich Ettinger. Associations between trait impulsivity and prepotent response inhibition. *Journal of clinical and experimental neuropsychology*, 00(00), August 2012. ISSN 1744-411X. doi: 10.1080/13803395.2012.706261. URL http://www.ncbi.nlm.nih.gov/pubmed/22888795.

Brad Aisa, Brian Mingus, and Randy O'Reilly. The emergent neural modeling system. *Neural Networks*, 21(8):1146–1152, Oct 2008. URL http://www.ncbi.nlm.nih.gov/pubmed/18684591.

G.E. Alexander, M.R. DeLong, and P.L. Strick. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, 9:357–381, 05 1986. URL http://www.ncbi.nlm.nih.gov/pubmed/3085570.

William H Alexander and Joshua W Brown. Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, 14(10):1338–1344, Oct 2011. URL http://www.ncbi.nlm.nih.gov/pubmed/21926982.

N Ambady and HM Gray. On being sad and mistaken: mood effects on the accuracy of thin-slice judgments. *Journal of personality and social psychology*, 83:947–961, 2002. URL http://psycnet.apa.org/journals/psp/83/4/947/.

Ralph Andre, Rachael I Scahill, Salman Haider, and Sarah J Tabrizi. Biomarker development for Huntington's disease. *Drug discovery today*, 00(00):2–9, March 2014. ISSN 1878-5832. doi: 10.1016/j.drudis.2014.03.002. URL http://www.ncbi.nlm.nih.gov/pubmed/24632006.

PW Andrews, SH Aggen, and GF Miller. The functional design of depression's influence on attention: A preliminary test of alternative control-process mechanisms.

Evolutionary ..., 5:584–604, 2007. URL http://psycnet.apa.org/psycinfo/2008-06860-008.

C Andrieu, N De Freitas, A Doucet, and MI Jordan. An introduction to MCMC for machine learning. *Machine learning*, 2003. URL http://www.springerlink.com/index/xh62794161k70540.pdf.

Pilar Andrs. Frontal cortex as the central executive of working memory: time to revise our view. *Cortex; a journal devoted to the study of the nervous system and behavior*, 39:871–896, 10 2003. URL http://www.ncbi.nlm.nih.gov/pubmed/14584557.

CE Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The annals of statistics*, 1974. URL http://www.jstor.org/stable/10.2307/2958336.

A. R. Aron. The neural basis of inhibition in cognitive control. *The Neuroscientist : a review journal bringing neurobiology, neurology and psychiatry*, 13(3):214–228, June 2007. ISSN 1073-8584. doi: 10.1177/1073858407299288. URL http://dx.doi.org/10.1177/1073858407299288.

A. R. Aron and F. Verbruggen. Stop the presses: dissociating a selective from a global mechanism for stopping. *Psychological science : a journal of the American Psychological Society / APS*, 19(11):1146–1153, November 2008. ISSN 1467-9280. doi: 10.1111/j.1467-9280.2008.02216.x. URL http://dx.doi.org/10.1111/j.1467-9280.2008.02216.x.

Adam R. Aron. From reactive to proactive and selective control: developing a richer model for stopping inappropriate responses. *Biological Psychiatry*, 69(12):e55–e68, Jun 2011. URL http://www.ncbi.nlm.nih.gov/pubmed/20932513.

Adam R Aron and Russell A Poldrack. Cortical and subcortical contributions to stop signal response inhibition: role of the subthalamic nucleus. *The Journal of* *neuroscience : the official journal of the Society for Neuroscience*, 26:2424–33, 03 2006. URL http://www.ncbi.nlm.nih.gov/pubmed/16510720.

Adam R Aron, Paul C Fletcher, Ed T Bullmore, Barbara J Sahakian, and Trevor W Robbins. Stop-signal inhibition disrupted by damage to right inferior frontal gyrus in humans. *Nature neuroscience*, 6:115–116, 01 2003. URL http://www.ncbi.nlm.nih.gov/pubmed/12536210.

Adam R Aron, Stephen Monsell, Barbara J Sahakian, and Trevor W Robbins. A componential analysis of task-switching deficits associated with lesions of left and right frontal cortex. *Brain : a journal of neurology*, 127(7):1561–1573, 06 2004. URL http://www.ncbi.nlm.nih.gov/pubmed/15090477.

Adam R Aron, Tim E Behrens, Steve Smith, Michael J Frank, and Russell A Poldrack. Triangulating a cognitive control network using diffusion-weighted magnetic resonance imaging (mri) and functional mri. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27(14):3743–3752, 04 2007a. URL http://www.ncbi.nlm.nih.gov/pubmed/17409238.

Adam R. Aron, Sarah Durston, Dawn M. Eagle, Gordon D. Logan, Cathy M. Stinear, and Veit Stuphorn. Converging evidence for a fronto-basal-ganglia network for inhibitory control of action and cognition. *J. Neurosci.*, 27(44):11860–11864, October 2007b. doi: 10.1523/JNEUROSCI.3644-07.2007. URL http://dx.doi.org/10.1523/JNEUROSCI.3644-07.2007.

American Psychiatric Association. *Diagnostic and statistical manual of mental disorders (5th ed.)*. Arlington, VA, American Psychiatric Publishing, 2013.

Gary Aston-Jones and Jonathan D Cohen. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annual review of neuroscience*, 28:403–450, 07 2005. URL http://www.ncbi.nlm.nih.gov/pubmed/16022602.

K Au, F Chan, D Wang, and I Vertinsky. Mood in foreign exchange trading: Cognitive processes and performance. *...and Human Decision Processes*, 91: 322–328, 2003. URL http://www.sciencedirect.com/science/article/pii/S0749597802005101.

E H Aylward, B F Sparks, K M Field, V Yallapragada, B D Shpritz, A Rosenblatt, J Brandt, L M Gourley, K Liang, H Zhou, R L Margolis, and C A Ross. Onset and rate of striatal atrophy in preclinical Huntington disease. *Neurology*, 63(1):66–72, July 2004. ISSN 1526-632X. URL http://www.ncbi.nlm.nih.gov/pubmed/15249612.

Adriano Azevedo-filho and Ross D Shachter. Laplace s Method Approximations for Probabilistic Inference in Belief Networks with Continuous Variables. pages 28–36, 1994.

J. C. Badcock, P. T. Michie, L. Johnson, and J. Combrinck. Acts of control in schizophrenia: dissociating the components of inhibition. *Psychological medicine*, 32(2):287–297, February 2002. ISSN 0033-2917. URL http://view.ncbi.nlm.nih.gov/pubmed/11866323.

David Badre. Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. *Trends in cognitive sciences*, 12, May 2008. URL http://www.ncbi.nlm.nih.gov/pubmed/18403252.

Bernhard Baier, Hans-Otto Karnath, Marianne Dieterich, Frank Birklein, Carolin Heinze, and Notger G. Muller. Keeping memory clear and stable–the contribution of human basal ganglia and prefrontal cortex to working memory. *J*, 30(29):9788–9792, 7 2010. URL http://www.jneurosci.org/cgi/content/abstract/30/29/9788.

Francisco Barcelo, Carles Escera, Maria J Corral, and Jose A Periez. Task switching and novelty processing activate a common neural network for cognitive control.

Journal of cognitive neuroscience, 18, Oct 2006. URL http://www.ncbi.nlm.nih.gov/pubmed/17014377.

E Becker and M Rinck. Sensitivity and response bias in fear of spiders. *Cognition and Emotion*, 2004. URL http://www.tandfonline.com/doi/abs/10.1080/02699930341000329.

Stefan Behnel, Robert Bradshaw, Craig Citro, Lisandro Dalcin, Dag Sverre Seljebotn, and Kurt Smith. Cython: The Best of Both Worlds. *Computing in Science & Engineering*, 13(2):31–39, March 2011. ISSN 1521-9615. doi: 10.1109/MCSE.2010.118. URL http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=5582062&contentType=Journals+&+Magazines.

M. A. Bellgrove, C. D. Chambers, A. Vance, N. Hall, M. Karamitsios, and J. L. Bradshaw. Lateralized deficit of response inhibition in early-onset schizophrenia. *Psychological medicine*, 36(4):495–505, April 2006. ISSN 0033-2917. doi: 10.1017/S0033291705006409. URL http://dx.doi.org/10.1017/S0033291705006409.

H Bernheimer, W Birkmayer, O Hornykiewicz, K Jellinger, and F Seitelberger. Brain dopamine and the syndromes of Parkinson and Huntington Clinical, morphological and neurochemical correlations. *Journal of the Neurological Sciences*, 20(4):415–455, December 1973. ISSN 0022510X. doi: 10.1016/0022-510X(73)90175-5. URL http://dx.doi.org/10.1016/0022-510X(73)90175-5.

Sebastian Bitzer, Hame Park, Felix Blankenburg, and Stefan J Kiebel. Perceptual decision making: drift-diffusion model is equivalent to a Bayesian model. *Frontiers in human neuroscience*, 8:102, January 2014. ISSN 1662-5161. doi: 10.3389/fnhum.2014.00102. URL http://www.frontiersin.org/Journal/10.3389/fnhum.2014.00102/abstracthttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3935359&tool=pmcentrez&rendertype=abstract.

M. M. Botvinick, T. S. Braver, D. M. Barch, C. S. Carter, and J. D. Cohen. Conflict monitoring and cognitive control. *Psychological Review*, 108:624–652, December 2001.

Matthew M. Botvinick, Jonathan D. Cohen, and Cameron S. Carter. Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences*, 8 (12):539–546, 11 2004. URL http://www.ncbi.nlm.nih.gov/pubmed/15556023.

Marcel Brass, Jan Derrfuss, Birte Forstmann, and D Yves von Cramon. The role of the inferior frontal junction area in cognitive control. *Trends in cognitive sciences*, 9, 2005. URL http://www.ncbi.nlm.nih.gov/pubmed/15927520.

T S Braver, D M Barch, J R Gray, D L Molfese, and A Snyder. Anterior cingulate cortex and response conflict: effects of frequency, inhibition and errors. *Cerebral cortex (New York, N.Y. : 1991)*, 11:825–836, 09 2001. URL http://www.ncbi.nlm.nih.gov/pubmed/11532888.

A Breier, L Kestler, C Adler, I Elman, N Wiesenfeld, A Malhotra, and D Pickar. Dopamine d2 receptor density and personal detachment in healthy subjects. *The American journal of psychiatry*, 155:1440–1442, 10 1998. URL http://www.ncbi.nlm.nih.gov/pubmed/9766779.

Leo Breiman. Random forests. *Machine learning*, pages 1–33, 2001. doi: 10.1023/A:1010933404324. URL http://link.springer.com/article/10.1023/A:1010933404324.

Kay H. Brodersen, Lorenz Deserno, Florian Schlagenhauf, Zhihao Lin, Will D. Penny, Joachim M. Buhmann, and Klaas E. Stephan. Dissecting psychiatric spectrum disorders by generative embedding. *NeuroImage. Clinical*, 4:98–111, November 2013. ISSN 22131582. doi: 10.1016/j.nicl.2013.11.002. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3863808&tool=pmcentrez&rendertype=abstracthttp://linkinghub.elsevier.com/retrieve/pii/S2213158213001502.

Jeff M Bronstein, Michele Tagliati, Ron L Alterman, Andres M Lozano, Jens Volkmann, Alessandro Stefani, Fay B Horak, Michael S Okun, Kelly D Foote, Paul Krack, Rajesh Pahwa, Jaimie M Henderson, Marwan I Hariz, Roy a Bakay, Ali Rezai, William J Marks, Elena Moro, Jerrold L Vitek, Frances M Weaver, Robert E Gross, and Mahlon R DeLong. Deep brain stimulation for Parkinson disease: an expert consensus and review of key issues. *Archives of neurology*, 68 (2):165, February 2011. ISSN 1538-3687. doi: 10.1001/archneurol.2010.260. URL http://www.ncbi.nlm.nih.gov/pubmed/20937936.

Joshua Brown, Daniel Bullock, and Stephen Grossberg. How laminar frontal cortex and basal ganglia circuits interact to control planned and reactive saccades. *Neural Networks*, 17:471–510, 04 2004. URL http://www.ncbi.nlm.nih.gov/pubmed/15109680.

Scott D Brown and Andrew Heathcote. The simplest complete model of choice response time: linear ballistic accumulation. *Cognitive psychology*, 57(3):153–78, November 2008. ISSN 1095-5623. doi: 10.1016/j.cogpsych.2007.12.002. URL http://www.ncbi.nlm.nih.gov/pubmed/18243170.

Silvia A. Bunge and Jonathan D. Wallis, editors. *Neuroscience of rule-guided behavior*. Oxford University Press, January 2008.

Borís Burle, Camille-Aimé Possamaï, Franck Vidal, Michel Bonnet, and Thierry Hasbroucq. Executive control in the Simon effect: an electromyographic and distributional analysis. *Psychological research*, 66(4):324–36, November 2002. ISSN 0340-0727. doi: 10.1007/s00426-002-0105-6. URL http://www.ncbi.nlm.nih.gov/pubmed/12466929.

Weidong Cai, Caitlin L. Oldenkamp, and Adam R. Aron. A proactive mechanism for selective suppression of response tendencies. *The Journal of Neuroscience*, 31(16): 5965–5969, Apr 2011. URL http://www.ncbi.nlm.nih.gov/pubmed/21508221.

B. J. Casey, J. T. Nigg, and S. Durston. New potential leads in the biology and treatment of attention deficit-hyperactivity disorder. *Current opinion in neurology*, 20(2):119–124, April 2007. ISSN 1350-7540. doi: 10.1097/WCO.0b013e3280a02f78. URL http://dx.doi.org/10.1097/WCO.0b013e3280a02f78.

James F Cavanagh, Thomas V Wiecki, Michael X Cohen, Christina M Figueroa, Johan Samanta, Scott J Sherman, and Michael J Frank. Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature neuroscience*, 14:1462–1467, Sep 2011. URL http://www.ncbi.nlm.nih.gov/pubmed/21946325.

James F Cavanagh, Laura Zambrano-Vazquez, and John J B Allen. Theta lingua franca: a common mid-frontal substrate for action monitoring processes. *Psychophysiology*, 49(2):220–38, February 2012. ISSN 1540-5958. doi: 10.1111/j.1469-8986.2011.01293.x. URL http://www.ncbi.nlm.nih.gov/pubmed/22091878.

S. R. Chamberlain, N. Del Campo, J. Dowson, U. Müller, L. Clark, T. W. Robbins, and B. J. Sahakian. Atomoxetine improved response inhibition in adults with attention deficit/hyperactivity disorder. *Biological psychiatry*, 62(9):977–984, November 2007. ISSN 0006-3223. doi: 10.1016/j.biopsych.2007.03.003. URL http://dx.doi.org/10.1016/j.biopsych.2007.03.003.

S. R. Chamberlain, A. Hampshire, U. Müller, K. Rubia, N. Del Campo, K. Craig, R. Regenthal, J. Suckling, J. P. Roiser, J. E. Grant, E. T. Bullmore, T. W. Robbins, and B. J. Sahakian. Atomoxetine modulates right inferior frontal activation during inhibitory control: a pharmacological functional magnetic resonance imaging study. *Biological psychiatry*, 65(7):550–555, April 2009. ISSN 1873-2402. doi: 10.1016/j.biopsych.2008.10.014. URL http://dx.doi.org/10.1016/j.biopsych.2008.10.014.

Samuel R. Chamberlain, Naomi A. Fineberg, Andrew D. Blackwell, Trevor W.

Robbins, and Barbara J. Sahakian. Motor inhibition and cognitive flexibility in obsessive-compulsive disorder and trichotillomania. *Am J Psychiatry*, 163(7):1282–1284, July 2006. doi: 10.1176/appi.ajp.163.7.1282. URL http://dx.doi.org/10.1176/appi.ajp.163.7.1282.

Samuel R Chamberlain, Lara Menzies, Adam Hampshire, John Suckling, Naomi A Fineberg, Natalia del Campo, Mike Aitken, Kevin Craig, Adrian M Owen, Edward T Bullmore, Trevor W Robbins, and Barbara J Sahakian. Orbitofrontal dysfunction in patients with obsessive-compulsive disorder and their unaffected relatives. *Science*, 321, Jul 2008. URL http://www.ncbi.nlm.nih.gov/pubmed/18635808.

Christopher D. Chambers, Hugh Garavan, and Mark A. Bellgrove. Insights into the neural basis of response inhibition from cognitive and clinical neuroscience. *Neuroscience & Biobehavioral Reviews*, 33(5):631–646, May 2009. ISSN 01497634. doi: 10.1016/j.neubiorev.2008.08.016. URL http://dx.doi.org/10.1016/j.neubiorev.2008.08.016.

C. Chatham, M. Frank, and D. Badre. Corticostriatal output gating during selection from working memory. *Neuron*, 81(4):930–942, Jan 2014.

C.H. Chatham, E.C. Claus, A. Kim, C. Curran, M.T. Banich, and Y Munakata. Cognitive control reflects context monitoring, not stopping, in response inhibition. *PloS one*.

Christopher H Chatham, Eric D Claus, Albert Kim, Tim Curran, Marie T Banich, and Yuko Munakata. Cognitive control reflects context monitoring, not motoric stopping, in response inhibition. *PloS one*, 7, 2012. URL http://www.ncbi.nlm.nih.gov/pubmed/22384038.

S Chib and E Greenberg. Understanding the metropolis-hastings algorithm. *The*

American Statistician, 1995. URL http://amstat.tandfonline.com/doi/abs/10.1080/00031305.1995.10476177.

Ivar A H Clemens, Maaike De Vrijer, Luc P J Selen, Jan A M Van Gisbergen, and W Pieter Medendorp. Multisensory processing in spatial orientation: an inverse probabilistic approach. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 31(14):5365–77, April 2011a. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.6472-10.2011. URL http://www.jneurosci.org/content/31/14/5365.short.

Ivar A H Clemens, Maaike De Vrijer, Luc P J Selen, Jan A M Van Gisbergen, and W Pieter Medendorp. Multisensory processing in spatial orientation: an inverse probabilistic approach. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 31(14):5365–77, April 2011b. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.6472-10.2011. URL http://www.jneurosci.org/content/31/14/5365.short.

Jeffrey Cockburn and Clay B Holroyd. Focus on the positive: computational simulations implicate asymmetrical reward prediction error signals in childhood attention-deficit/hyperactivity disorder. *Brain research*, 1365:18–34, December 2010. ISSN 1872-6240. doi: 10.1016/j.brainres.2010.09.065. URL http://www.ncbi.nlm.nih.gov/pubmed/20875804.

Jessica R Cohen and Russell A Poldrack. Automaticity in motor sequence learning does not impair response inhibition. *Psychonomic bulletin & review*, 15, Feb 2008. URL http://www.ncbi.nlm.nih.gov/pubmed/18605489.

Anne G E Collins and Michael J Frank. How much of reinforcement learning is working memory, not reinforcement learning? a behavioral, computational, and neurogenetic analysis. *The European journal of neuroscience*, 35, Apr 2012. URL http://www.ncbi.nlm.nih.gov/pubmed/22487033.

Anne G. E. Collins and Michael J. Frank. Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review*, 120(1):190–229, Jan 2013. URL http://www.ncbi.nlm.nih.gov/pubmed/23356780.

Eliza Congdon, R. Todd Constable, Klaus P. Lesch, and Turhan Canli. Influence of slc6a3 and comt variation on neural activation during response inhibition. *Biological Psychology*, 81(3):144–152, July 2009. ISSN 03010511. doi: 10.1016/j.biopsycho.2009.03.005. URL http://dx.doi.org/10.1016/j.biopsycho.2009.03.005.

Daniel Cressey. Psychopharmacology in crisis. *Nature*, June 2011. ISSN 1476-4687. doi: 10.1038/news.2011.367. URL http://www.nature.com/news/2011/110614/full/news.2011.367.html.

C. E. Curtis and M. DEsposito. Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences*, 7:415–423, January 2003.

Chambers Christopher D., Bellgrove Mark A., Gould Ian C., English Therese, Garavan Hugh, McNaught Elizabeth, Kamke Marc, and Mattingley Jason B. Dissociable mechanisms of cognitive control in prefrontal and premotor cortex. *Journal of neurophysiology*, 98(6):3638–3647, Dec 2007. URL http://www.ncbi.nlm.nih.gov/pubmed/17942624.

Jeffrey W Dalley, Barry J Everitt, and Trevor W Robbins. Impulsivity, compulsivity, and top-down cognitive control. *Neuron*, 69(4), Feb 2011. URL http://www.ncbi.nlm.nih.gov/pubmed/21338879.

Nathaniel D. Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12):1704–1711, 11 2005. URL http://www.ncbi.nlm.nih.gov/pubmed/16286932.

Wang DeLiang, Arnj Kristjansson, and Ken Nakayama. Efficient visual search without top-down or bottom-up guidance. *Perception & Psychophysics*, 67(2):239–253, January 2005.

AP Dempster, NM Laird, and DB Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. ...*, 39(1):1–38, 1977. URL http://www.jstor.org/stable/10.2307/2984875.

S. Deneve. Bayesian spiking neurons i: Inference. *Neural Comput*, 20(1):91–117, 2008.

Jan Derrfuss, Marcel Brass, and D Yves von Cramon. Cognitive control in the posterior frontolateral cortex: evidence from common activations in task coordination, interference control, and working memory. *NeuroImage*, 23, Oct 2004. URL http://www.ncbi.nlm.nih.gov/pubmed/15488410.

Jan Derrfuss, Marcel Brass, Jane Neumann, and D. Yves von Cramon. Involvement of the inferior frontal junction in cognitive control: meta-analyses of switching and stroop studies. *Human brain mapping*, 25(1):22–34, May 2005. URL http://www.ncbi.nlm.nih.gov/pubmed/15846824?ordinalpos.

DG Dillon and DA Pizzagalli. Evidence of successful modulation of brain activation and subjective experience during reappraisal of negative emotion in unmedicated depression. *Psychiatry Research: Neuroimaging*, 212:99–107, 2013. URL http://www.sciencedirect.com/science/article/pii/S0925492713000036.

Bradley B. Doll, W. Jake Jacobs, Alan G. Sanfey, and Michael J. Frank. Instructional control of reinforcement learning: a behavioral and neurocomputational investigation. *Brain Research*, 1299:74–94, Nov 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19595993.

S Dubal, a Pierson, and R Jouvent. Focused attention in anhedonia: a P3 study. *Psychophysiology*, 37(5):711–4, September 2000. ISSN 0048-5772. URL http://www.ncbi.nlm.nih.gov/pubmed/11037048.

Stéphanie Dubal and Roland Jouvent. Time-on-task effect in trait anhedonia. *European psychiatry : the journal of the Association of European Psychiatrists*, 19(5): 285–91, August 2004. ISSN 0924-9338. doi: 10.1016/j.eurpsy.2004.04.007. URL http://www.ncbi.nlm.nih.gov/pubmed/15276661.

Dawn M Eagle, Christelle Baunez, Daniel M Hutcheson, Olivia Lehmann, Aarti P Shah, and Trevor W Robbins. Stop-signal reaction-time task performance: role of prefrontal cortex and subthalamic nucleus. *Cerebral cortex*, 18, Jan 2008. URL http://www.ncbi.nlm.nih.gov/pubmed/17517682.

T Ehring, B Tuschen-Caffier, and J Schnülle. Emotion regulation and vulnerability to depression: spontaneous versus instructed use of emotion suppression and reappraisal. *Emotion*, 10:563–572, 2010. URL http://psycnet.apa.org/journals/emo/10/4/563/.

M Eimer. S-R compatibility and response selection. *Acta Psychologica*, 90(1-3):301–313, November 1995. ISSN 00016918. doi: 10.1016/0001-6918(95)00022-M. URL http://dx.doi.org/10.1016/0001-6918(95)00022-M.

B Elvevå g and T E Goldberg. Cognitive impairment in schizophrenia is the core of the disorder. *Critical reviews in neurobiology*, 14(1):1–21, January 2000. ISSN 0892-0915. URL http://www.ncbi.nlm.nih.gov/pubmed/11253953.

Erik E Emeric, Melanie Leslie, Pierre Pouget, and Jeffrey D Schall. Performance monitoring local field potentials in the medial frontal cortex of primates: supplementary eye field. *Journal of neurophysiology*, 104(3):1523–37, September 2010. ISSN 1522-1598. doi: 10.1152/jn.01001.2009. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2944693&tool=pmcentrez&rendertype=abstract.

B. A. Eriksen and C. W. Eriksen. Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception and Psychophysics*, 16:143–149, January 1974.

Jonathan St B T Evans. In two minds: dual-process accounts of reasoning. *Trends in cognitive sciences*, 7(10):454–459, 01 2005. URL http://www.ncbi.nlm.nih.gov/pubmed/14550493.

S Everling and D P Munoz. Neuronal correlates for preparatory set associated with pro-saccades and anti-saccades in the primate frontal eye field. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 20:387, 01 2000. URL http://www.ncbi.nlm.nih.gov/pubmed/10627615.

Stefan Everling, Michael C. Dorris, Raymond M. Klein, and Douglas P. Munoz. Role of primate superior colliculus in preparation and execution of anti-saccades and pro-saccades. *J. Neurosci.*, 19(7):2740–2754, April 1999. URL http://www.jneurosci.org/cgi/content/abstract/19/7/2740.

D. a. Fair, D. Bathula, M. a. Nikolas, and J. T. Nigg. Distinct neuropsychological subgroups in typically developing youth inform heterogeneity in children with ADHD. *Proceedings of the National Academy of Sciences*, April 2012. ISSN 0027-8424. doi: 10.1073/pnas.1115365109. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1115365109.

M. Falkenstein, J. Hohnsbein, J. Hoormann, and L. Blanke. Effects of cross-modal divided attention on late erp components: Ii. error processing in choice reaction tasks. *Electroencephalography and Clinical Neurophysiology*, 78:447–55, January 1991.

Stephen V Faraone, Roy H Perlis, Alysa E Doyle, Jordan W Smoller, Jennifer J Goralnick, Meredith A Holmgren, and Pamela Sklar. Molecular genetics of attention-deficit/hyperactivity disorder. *Biol Psychiatry*, 57(11):1313–1323, Jun 2005. URL http://dx.doi.org/10.1016/j.biopsych.2004.11.024.

I. Farley, K. Price, E McCullough, J. Deck, W Hordynski, and O Hornykiewicz. Norepinephrine in chronic paranoid schizophrenia: above-normal levels in limbic forebrain. *Science*, 200(4340):456–458, April 1978. ISSN 0036-8075. doi: 10.1126/science.644310. URL http://www.sciencemag.org/content/200/4340/456.abstract.

William Feller. *An Introduction to Probability Theory and Its Applications, Vol. 1, 3rd Edition*. Wiley, 1968. ISBN 0471257087. URL http://www.amazon.com/Introduction-Probability-Theory-Applications-Edition/dp/0471257087.

Stephen M Fleming, Charlotte L Thomas, and Raymond J Dolan. Overcoming status quo bias in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, 107, Mar 2010. URL http://www.ncbi.nlm.nih.gov/pubmed/20231462.

Kristen A Ford and Stefan Everling. Neural activity in primate caudate nucleus associated with pro- and antisaccades. *Journal of neurophysiology*, 102(4):2334–2341, Oct 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19692516.

Birte U Forstmann, Gilles Dutilh, Scott Brown, Jane Neumann, D Yves von Cramon, K Richard Ridderinkhof, and Eric-Jan Wagenmakers. Striatum and presma facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences of the United States of America*, 105, Nov 2008. URL http://www.ncbi.nlm.nih.gov/pubmed/18981414.

Birte U Forstmann, Alfred Anwander, Andreas Schafer, Jane Neumann, Scott Brown, Eric-Jan Wagenmakers, Rafal Bogacz, and Robert Turner. Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. *Proceedings of the National Academy of Sciences of the United States of America*, pages 1–5online, Aug 2010a. URL http://www.ncbi.nlm.nih.gov/pubmed/20733082.

Birte U. Forstmann, Scott Brown, Gilles Dutilh, Jane Neumann, and Eric-Jan Wagenmakers. The neural substrate of prior information in perceptual decision making:

a model-based analysis. *Frontiers in human neuroscience*, 4:1–12 (online), 2010b. URL http://www.ncbi.nlm.nih.gov/pubmed/20577592.

Johan Frank, Michael J.and Samanta, Ahmed A. Moustafa, and Scott J. Sherman. Hold your horses: Impulsivity, deep brain stimulation, and medication in parkinsonism. *Science*, 318(5854):1309–1312, 11 2007a. URL http://www.ncbi.nlm.nih.gov/pubmed/17962524.

M. J. Frank. Dynamic dopamine modulation in the basal ganglia: A neurocomputational account of cognitive deficits in medicated and non-medicated Parkinsonism. *Journal of Cognitive Neuroscience*, 17:51–72, January 2005.

M. J. Frank, B. Loughry, and R. C. O'Reilly. Interactions between the frontal cortex and basal ganglia in working memory: A computational model. *Cognitive, Affective, and Behavioral Neuroscience*, 1:137–160, January 2001.

M. J. Frank, L. C. Seeberger, and R. C. O'Reilly. By carrot or by stick: Cognitive reinforcement learning in Parkinsonism. *Science*, 306(5703):1940–1943, January 2004.

Michael Frank and David Badre. Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: Computational analysis. *Cerebral Cortex*, page online, Jun 2011. URL http://cercor.oxfordjournals.org/content/early/2011/06/21/cercor.bhr114.abstract.

Michael J Frank. Hold your horses: a dynamic computational role for the subthalamic nucleus in decision making. *Neural networks : the official journal of the International Neural Network Society*, 19:1120–1136, 10 2006. URL http://www.ncbi.nlm.nih.gov/pubmed/16945502.

Michael J Frank, Amy Santamaria, Randall C O'Reilly, and Erik Willcutt. Testing computational models of dopamine and noradrenaline dysfunction in attention deficit/hyperactivity disorder. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, 32:1583–1599, 06 2007b. URL http://www.ncbi.nlm.nih.gov/pubmed/17164816.

Michael J Frank, Anouk Scheres, and Scott J Sherman. Understanding decision-making deficits in neurological conditions: insights from models of natural action selection. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 362:1641–1654, 08 2007c. URL http://www.ncbi.nlm.nih.gov/pubmed/17428775.

Michael J Frank, Bradley B Doll, Jen Oas-Terpstra, and Francisco Moreno. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience*, 12(8):1062–1068, Aug 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19620978.

BJ Frey and N Jojic. A comparison of algorithms for inference and learning in probabilistic graphical models. *Pattern Analysis and Machine Intelligence, . . .*, 2005. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1471706.

Jerome H. Friedman. Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1):1–67, March 1991. ISSN 2168-8966. URL http://projecteuclid.org/euclid.aos/1176347963.

K J Friston, L Harrison, and W Penny. Dynamic causal modelling. *NeuroImage*, 19 (4):1273–302, August 2003. ISSN 1053-8119. URL http://www.ncbi.nlm.nih.gov/pubmed/12948688.

M. Fukumoto-Motoshita, M. Matsuura, T. Ohkubo, H. Ohkubo, N. Kanaka, E. Matsushima, M. Taira, T. Kojima, and T. Matsuda. Hyperfrontality in patients with schizophrenia during saccade and antisaccade tasks: a study with fmri. *Psychiatry and clinical neurosciences*, 63(2):209–217, April 2009. ISSN 1440-1819.

doi: 10.1111/j.1440-1819.2009.01941.x. URL http://dx.doi.org/10.1111/j.1440-1819.2009.01941.x.

S. Funahashi, M. V. Chafee, and P. S. Goldman-Rakic. Prefrontal neuronal activity in rhesus monkeys performing a delayed anti-saccade task. *Nature*, 365:753–756, 11 1993. URL http://www.ncbi.nlm.nih.gov/pubmed/8413653.

D Gamerman and HF Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. 2006. URL http://books.google.com/books?hl=en&lr=&id=yPvECi_L3bwC&oi=fnd&pg=PR13&dq=gamerman+bayesian&ots=NisYwWffgg&sig=WI3OIzsaVzc-mRbr_DceissL7jk.

H. Garavan, R. Hester, K. Murphy, C. Fassbender, and C. Kelly. Individual differences in the functional neuroanatomy of inhibitory control. *Brain Research*, 1105(1): 130–142, August 2006. ISSN 00068993. doi: 10.1016/j.brainres.2006.03.029. URL http://dx.doi.org/10.1016/j.brainres.2006.03.029.

W. J. Gehring, B. Goss, M. G. H. Coles, D. E. Meyer, and E. Donchin. A neural system for error detection and compensation. *Psychological Science*, 4(6):385–390, January 1993.

A Gelman, JB Carlin, HS Stern, and DB Rubin. *Bayesian data analysis*. 2003. URL http://books.google.com/books?hl=en&lr=&id=TNYhnkXQSjAC&oi=fnd&pg=PP1&dq=Gelman+Carlin+Stern+04&ots=5H4S8DAwH3&sig=W9fgGzxMiklkMGA2fnwQACTz8BY.

Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472, 1992.

Samuel J. Gershman and David M. Blei. A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12, February 2012. ISSN 00222496. doi: 10.1016/j.jmp.2011.08.004. URL http://linkinghub.elsevier.com/retrieve/pii/S002224961100071X.

James M Gold, James A Waltz, Kristen J Prentice, Sarah E Morris, and Erin A Heerey. Reward processing in schizophrenia: a deficit in the representation of value. *Schizophrenia bulletin*, 34:835–847, 08 2008. URL http://www.ncbi.nlm.nih.gov/pubmed/18591195.

James M Gold, James A Waltz, Tatyana M Matveeva, Zuzana Kasanova, Gregory P Strauss, Ellen S Herbener, Anne G E Collins, and Michael J Frank. Negative Symptoms and the Failure to Represent the Expected Reward Value of Actions. 69(2):129–138, 2012.

Joshua I Gold and Michael N Shadlen. Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, 36:299–308, 10 2002. URL http://www.ncbi.nlm.nih.gov/pubmed/12383783.

Joshua I Gold and Michael N Shadlen. The neural basis of decision making. *Annual review of neuroscience*, 30, 2007. URL http://www.ncbi.nlm.nih.gov/pubmed/17600525.

Jesse Heymann Goldberg, Michael Alan Farries, and Michale S Fee. Integration of cortical and pallidal inputs in the basal ganglia-recipient thalamus of singing birds. *Journal of neurophysiology*, 108(5):1403–29, June 2012. ISSN 1522-1598. doi: 10.1152/jn.00056.2012. URL http://www.ncbi.nlm.nih.gov/pubmed/22673333.

Ian Greenhouse, Caitlin L Oldenkamp, Adam R Aron, and San Diego. Stopping a response has global or non-global effects on the motor system depending on preparation. *Journal of neurophysiology*, October 2011. ISSN 1522-1598. doi: 10.1152/jn.00704.2011. URL http://www.ncbi.nlm.nih.gov/pubmed/22013239.

K. Gurney, T. J. Prescott, and P. Redgrave. A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biological Cybernetics*, 84:401–410, 06 2001. URL http://www.ncbi.nlm.nih.gov/pubmed/11417052.

Suzanne N. Haber. The primate basal ganglia: parallel and integrative networks. *Journal of Chemical Neuroanatomy*, 26(4):317–330, 01 2003. URL http://www.ncbi.nlm.nih.gov/pubmed/14729134.

T D Hälbig, W Tse, P G Frisina, B R Baker, E Hollander, H Shapiro, M Tagliati, W C Koller, and C W Olanow. Subthalamic deep brain stimulation and impulse control in Parkinson's disease. *European journal of neurology : the official journal of the European Federation of Neurological Societies*, 16(4):493–7, April 2009. ISSN 1468-1331. doi: 10.1111/j.1468-1331.2008.02509.x. URL http://www.ncbi.nlm.nih.gov/pubmed/19236471.

P E Hallett. Primary and secondary saccades to goals defined by instructions. *Vision research*, 18:1270–1296, 02 1979. URL http://www.ncbi.nlm.nih.gov/pubmed/726270.

Adam Hampshire, Samuel R Chamberlain, Martin M Monti, John Duncan, and Adrian M Owen. The role of the right inferior frontal gyrus: inhibition and attentional control. *NeuroImage*, 50:1313–1319, Jan 2010. URL http://www.ncbi.nlm.nih.gov/pubmed/20056157.

D G Harden and A A Grace. Activation of dopamine cell firing by repeated l-dopa administration to dopamine-depleted rats: its potential role in mediating the therapeutic response to l-dopa treatment. *J Neurosci*, 15(9):6157–66, January 1995.

M. S. Harris, J. L. Reilly, M. S. Keshavan, and J. A. Sweeney. Longitudinal studies of antisaccades in antipsychotic-naive first-episode schizophrenia. *Psychological medicine*, 36(4):485–494, April 2006. ISSN 0033-2917. doi: 10.1017/S0033291705006756. URL http://dx.doi.org/10.1017/S0033291705006756.

WR Hess, S Bürgi, and V Bucher. Motorische Funktion des Tectal- und Tegmentalgebietes. *Mschr Psychiat Neurol*, 112:1–52, 1946.

Stephen L Hicks, Matthieu P a Robert, Charlotte V P Golding, Sarah J Tabrizi, and Christopher Kennard. Oculomotor deficits indicate the progression of Huntington's disease. *Progress in brain research*, 171(08):555–8, January 2008. ISSN 1875-7855. doi: 10.1016/S0079-6123(08)00678-X. URL http://www.ncbi.nlm.nih.gov/pubmed/18718352.

Takatoshi Hikida, Kensuke Kimura, Norio Wada, Kazuo Funabiki, and Shigetada Nakanishi. Distinct roles of synaptic transmission in direct and indirect striatal pathways to reward and aversive behavior. *Neuron*, 66:896–907, 2010.

O. Hikosaka. Role of basal ganglia in initiation of voluntary movements. In M. A. Arbib and S. Amari, editors, *Dynamic Interactions in Neural Networks: Models and Data*, pages 153–167. Springer-Verlag, Berlin, January 1989.

O. Hikosaka. *GABAergic output of the basal ganglia*, volume 160 of *Progress in Brain Research*, pages 209–226. 2007. doi: 10.1016/S0079-6123(06)60012-5. URL http://dx.doi.org/10.1016/S0079-6123(06)60012-5.

O. Hikosaka and M. Isoda. Brain mechanisms for switching from automatic to controlled eye movements. *Progress in brain research*, 171:375–382, 2008. ISSN 1875-7855. doi: 10.1016/S0079-6123(08)00655-9. URL http://dx.doi.org/10.1016/S0079-6123(08)00655-9.

O. Hikosaka and R. H. Wurtz. Saccadic eye movements following injection of lidocaine into the superior colliculus. *Experimental brain research. Experimentelle Hirnforschung. Expérimentation cérébrale*, 61(3):531–539, 1986. ISSN 0014-4819. URL http://view.ncbi.nlm.nih.gov/pubmed/3082658.

O. Hikosaka, Y. Takikawa, and R. Kawagoe. Role of the basal ganglia in the control of purposive saccadic eye movements. *Physiological Reviews*, 80(3):953–978, Jul 2000. URL http://www.ncbi.nlm.nih.gov/pubmed/10893428.

Okihide Hikosaka, Kae Nakamura, and Hiroyuki Nakahara. Basal ganglia orient eyes to reward. *Journal of neurophysiology*, 95(2):567–584, Feb 2006. URL http://www.ncbi.nlm.nih.gov/pubmed/16424448.

Nicola Z Hobbs, Susie M D Henley, Edward J Wild, Kelvin K Leung, Chris Frost, Roger A Barker, Rachael I Scahill, Josephine Barnes, Sarah J Tabrizi, and Nick C Fox. Automated quantification of caudate atrophy by local registration of serial MRI: evaluation and application in Huntington's disease. *NeuroImage*, 47(4):1659–65, October 2009. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2009.06.003. URL http://www.ncbi.nlm.nih.gov/pubmed/19523522.

Clay B Holroyd and Michael G H Coles. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological review*, 109:679–709, 10 2002. URL http://www.ncbi.nlm.nih.gov/pubmed/12374324.

Sien Hu and Chiang-Shan R Li. Neural processes of preparatory control for stop signal inhibition. *Human brain mapping*, 000, October 2011. ISSN 1097-0193. doi: 10.1002/hbm.21399. URL http://www.ncbi.nlm.nih.gov/pubmed/21976392.

Yanping Huang and Rajesh P N Rao. Reward optimization in the primate brain: a probabilistic model of decision making under uncertainty. *PloS one*, 8(1):e53344, January 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0053344. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3551910&tool=pmcentrez&rendertype=abstract.

Ronald Hübner, Marco Steinhauser, and Carola Lehle. A dual-stage two-phase model of selective attention. *Psychological review*, 117(3):759–84, July 2010. ISSN 1939-1471. doi: 10.1037/a0019471. URL http://www.ncbi.nlm.nih.gov/pubmed/20658852.

V. C. Huddy, A. R. Aron, M. Harrison, T. R. E. Barnes, T. W. Robbins, and E. M.

Joyce. Impaired conscious and preserved unconscious inhibitory processing in recent onset schizophrenia. *Psychological Medicine*, 39(06):907–916, 2009. doi: 10.1017/S0033291708004340. URL http://dx.doi.org/10.1017/S0033291708004340.

M F Huerta, L A Krubitzer, and J H Kaas. Frontal eye field as defined by intracortical microstimulation in squirrel monkeys, owl monkeys, and macaque monkeys. II. Cortical connections. *The Journal of comparative neurology*, 265(3):332–61, November 1987. ISSN 0021-9967. URL http://www.ncbi.nlm.nih.gov/pubmed/2447132.

Scott A Huettel and Gregory McCarthy. What is odd in the oddball task? Prefrontal cortex is activated by dynamic changes in response strategy. *Neuropsychologia*, 42(3):379–86, January 2004. ISSN 0028-3932. URL http://www.ncbi.nlm.nih.gov/pubmed/14670576.

Samuel B Hutton and Ulrich Ettinger. The antisaccade task as a research tool in psychopathology: a critical review. *Psychophysiology*, 43(3):302–13, May 2006. ISSN 0048-5772. doi: 10.1111/j.1469-8986.2006.00403.x. URL http://www.ncbi.nlm.nih.gov/pubmed/16805870.

Quentin J. M. Huys, Neir Eshel, Elizabeth O'Nions, Luke Sheridan, Peter Dayan, and Jonathan P. Roiser. Bonsai Trees in Your Head: How the Pavlovian System Sculpts Goal-Directed Choices by Pruning Decision Trees. *PLoS Computational Biology*, 8(3):e1002410, March 2012a. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002410. URL http://dx.plos.org/10.1371/journal.pcbi.1002410.

Quentin J M Huys, Neir Eshel, Elizabeth O'Nions, Luke Sheridan, Peter Dayan, and Jonathan P Roiser. Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS computational biology*, 8, 2012b. URL http://www.ncbi.nlm.nih.gov/pubmed/22412360.

Kai Hwang, Katerina Velanova, and Beatriz Luna. Strengthening of top-down frontal cognitive control networks underlying the development of inhibitory control: a functional magnetic resonance imaging effective connectivity study. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 30(46): 15535–45, November 2010. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.2825-10.2010. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2995693&tool=pmcentrez&rendertype=abstract.

Steven E Hyman. Revolution Stalled. 4(155):1–5, 2012a.

Steven E Hyman. Revolution Stalled. 4(155):1–5, 2012b.

T. Insel, B. Cuthbert, M. Garvey, R. Heinssen, D. S. Pine, K. Quinn, C. A. Sanislow, and P. W. Wang. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, 2010. URL http://works.bepress.com/charles_sanislow/2/.

Masaki Isoda and Okihide Hikosaka. Switching from automatic to controlled action by monkey medial frontal cortex. *Nature neuroscience*, 10(2):240–248, 01 2007. URL http://www.ncbi.nlm.nih.gov/pubmed/17237780.

Masaki Isoda and Okihide Hikosaka. Role for subthalamic nucleus neurons in switching from automatic to controlled eye movement. *The Journal of Neuroscience*, 28: 7209–7218, 07 2008. URL http://www.ncbi.nlm.nih.gov/pubmed/18614691.

Sara Jahfari, Lourens Waldorp, Wery P M van den Wildenberg, H Steven Scholte, K Richard Ridderinkhof, and Birte U Forstmann. Effective connectivity reveals important roles for both the hyperdirect (fronto-subthalamic) and the indirect (fronto-striatal-pallidal) fronto-basal ganglia pathways during response inhibition. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 31 (18):6891–9, May 2011. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.5253-10.2011. URL http://www.ncbi.nlm.nih.gov/pubmed/21543619.

Sara Jahfari, Frederick Verbruggen, Michael J Frank, Lourens J Waldorp, Lorenza Colzato, K Richard Ridderinkhof, and Birte U Forstmann. How preparation changes the need for top-down control of the basal ganglia when inhibiting premature actions. *The Journal of neuroscience*, 32, Aug 2012. URL http://www.ncbi.nlm.nih.gov/pubmed/22875921.

Gerhard Jocham, Tilmann A Klein, and Markus Ullsperger. Dopamine-mediated reinforcement learning signals in the striatum and ventromedial prefrontal cortex underlie value-based choices. *The Journal of neuroscience*, 31, Feb 2011. URL http://www.ncbi.nlm.nih.gov/pubmed/21289169.

Katherine Johnson, Ross Cunnington, Robert Iansek, John Bradshaw, Nellie Georgiou, and Edmond Chiu. Movement-related potentials in Huntington's disease: movement preparation and execution. *Experimental Brain Research*, 138(4):492–499, June 2001. ISSN 00144819. doi: 10.1007/s002210100733. URL http://link.springer.com/10.1007/s002210100733.

Kevin Johnston and Stefan Everling. Monkey dorsolateral prefrontal cortex sends task-selective signals directly to the superior colliculus. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 26(48):12471–8, November 2006. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.4101-06.2006. URL http://www.ncbi.nlm.nih.gov/pubmed/17135409.

Robert E. Kass and Adrian E. Raftery. Bayes factors and model uncertainty. Technical Report 571, Carnegie Mellon University Dept of Statistics, Pittsburgh, PA 15213, January 1993.

Robert E. Kass and Adrian E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795, June 1995. ISSN 0162-1459. doi: 10.1080/01621459.1995.10476572. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572.

Jan Kassubek, Freimut D Juengling, Daniel Ecker, and G Bernhard Landwehrmeyer. Thalamic atrophy in Huntington's disease co-varies with cognitive performance:

a morphometric MRI analysis. *Cerebral cortex (New York, N.Y. : 1991)*, 15(6): 846–53, June 2005. ISSN 1047-3211. doi: 10.1093/cercor/bhh185. URL http://www.ncbi.nlm.nih.gov/pubmed/15459079.

Karl Kieburtz and Charles Venuto. TRACK-HD: both promise and disappointment. *Lancet neurology*, 11(1):24–5, January 2012. ISSN 1474-4465. doi: 10.1016/S1474-4422(11)70285-X. URL http://www.ncbi.nlm.nih.gov/pubmed/22137355.

Jirí Klempír, Olga Klempírova, Natasa Spacková, Jana Zidovská, and Jan Roth. Unified Huntington's disease rating scale: clinical practice and a critical approach. *Functional neurology*, 21(4):217–21, 2005. ISSN 0393-5264. URL http://www.ncbi.nlm.nih.gov/pubmed/17367582http://europepmc.org/abstract/MED/17367582.

Stefan Klöppel, Bogdan Draganski, Charlotte V Golding, Carlton Chu, Zoltan Nagy, Philip a Cook, Stephen L Hicks, Christopher Kennard, Daniel C Alexander, Geoff J M Parker, Sarah J Tabrizi, and Richard S J Frackowiak. White matter connections reflect changes in voluntary-guided saccades in pre-symptomatic Huntington's disease. *Brain : a journal of neurology*, 131(Pt 1):196–204, January 2008. ISSN 1460-2156. doi: 10.1093/brain/awm275. URL http://www.ncbi.nlm.nih.gov/pubmed/18056161.

Nils Kolling, Timothy Behrens, Rogier Mars, and Matthew Rushworth. Neural mechanisms of foraging. *Science*, 336(6077):95–98, 4 2012. URL http://www.sciencemag.org/content/336/6077/95.abstract.

S Kornblum, T Hasbroucq, and A Osman. Dimensional overlap: cognitive basis for stimulus-response compatibility–a model and taxonomy. *Psychological review*, 97 (2):253–270, 06 1990. URL http://www.ncbi.nlm.nih.gov/pubmed/2186425.

Alexxai Kravitz, Benjamin Freeze, Philip Parker, Kenneth Kay, Myo Thwin, Karl Deisseroth, and Anatol Kreitzer. Regulation of Parkinsonian motor behaviours by optogenetic control of basal ganglia circuitry. *Nature*, 466:622–626, Jul 2010. URL http://www.nature.com/nature/journal/vaop/ncurrent/full/nature09159.html.

Alexxai V Kravitz, Lynne D Tye, and Anatol C Kreitzer. Distinct roles for direct and indirect pathway striatal neurons in reinforcement. *Nature neuroscience*, Apr 2012. URL http://www.ncbi.nlm.nih.gov/pubmed/22544310.

J Kruschke. *Doing Bayesian data analysis: A tutorial introduction with R and BUGS*. Academic Press / Elsevier, 2010. ISBN 9780123814852. URL http://books.google.com/books?hl=en&lr=&id=ZRMJ-CebFm4C&oi=fnd&pg=PP2&dq=kruschke+doing+bayesian+data+analysis&ots=DsCCPI6uAW&sig=05YmmOLJ8DbRLwvtXxwWyIg0Eq0.

Andrea A. Kuhn, David Williams, Andreas Kupsch, Patricia Limousin, Marwan Hariz, Gerd-Helge Schneider, Kielan Yarrow, and Peter Brown. Event-related beta desynchronization in human subthalamic nucleus correlates with motor performance. *Brain*, 127(4):735–746, April 2004. doi: 10.1093/brain/awh106. URL http://dx.doi.org/10.1093/brain/awh106.

Brian Kulis, Michael I Jordan, Jordan Eecs, and Berkeley Edu. Revisiting k-means : New Algorithms via Bayesian Nonparametrics. 2012.

David LaBerge. A recruitment theory of simple behavior. *Psychometrika*, 27(4): 375–396, December 1962. ISSN 0033-3123. doi: 10.1007/BF02289645. URL http://www.springerlink.com/index/10.1007/BF02289645.

M. D. Lee and E.-J. Wagenmakers. *Bayesian Modeling for Cognitive Science: A Practical Course.* Cambridge University Press., 2013.

Lauren a Leotti and Tor D Wager. Motivational influences on response inhibition measures. *Journal of experimental psychology. Human perception and performance*, 36(2):430–47, April 2010. ISSN 1939-1277. doi: 10.1037/a0016802. URL http://www.ncbi.nlm.nih.gov/pubmed/20364928.

H. C. Leung and W. Cai. Common and differential ventrolateral prefrontal activity during inhibition of hand and eye movements. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27(37):9893–9900, September 2007. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.2837-07.2007. URL http://dx.doi.org/10.1523/JNEUROSCI.2837-07.2007.

Dennis Victor Lindley. *Introduction to Probability and Statistics from Bayesian Viewpoint. Part 2: inference.* CUP Archive, 1965.

Chung-Chuan Lo and Xiao-Jing Wang. Cortico-basal ganglia circuit mechanism for a decision threshold in reaction time tasks. *Nature neuroscience*, 9(7):956–963, 06 2006. URL http://www.ncbi.nlm.nih.gov/pubmed/16767089.

Chung-Chuan Lo, Leanne Boucher, Martin Pare, Jeffrey D. Schall, and Xiao-Jing Wang. Proactive inhibitory control and attractor dynamics in countermanding action: a spiking neural circuit model. *The Journal of Neuroscience*, 29(28):9059–9071, Jul 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19605643.

G. D. Logan. On the ability to inhibit simple thoughts and actions: Ii. stop-signal studies of repetition priming. *Journal of Experimental Psychology*, 11(4):675–691, January 1985.

G. D. Logan and W. B. Cowan. On the ability to inhibit thought and action: A theory of an act of control. *Psychological Review*, 91(3):295–327, January 1984.

Gordon D. Logan, Russell J. Schachar, and Rosemary Tannock. Impulsivity and inhibitory control. *Psychological Science*, 8(1):60–64, 1997. doi: 10.1111/j.1467-9280.1997.tb00545.x. URL http://dx.doi.org/10.1111/j.1467-9280.1997.tb00545.x.

M T Lu, J B Preston, and P L Strick. Interconnections between the prefrontal cortex and the premotor areas in the frontal lobe. *The Journal of comparative neurology*, 341(3):375–92, March 1994. ISSN 0021-9967. doi: 10.1002/cne.903410308. URL http://www.ncbi.nlm.nih.gov/pubmed/7515081.

J MacMillan and O Quarrell. The neurobiology of Huntington's disease. *MAJOR PROBLEMS IN NEUROLOGY*, 31:317–358, 1996.

Tiago V Maia and Michael J Frank. From reinforcement learning models to psychiatric and neurological disorders. *Nature neuroscience*, 14(2):154–62, February 2011. ISSN 1546-1726. doi: 10.1038/nn.2723. URL http://www.ncbi.nlm.nih.gov/pubmed/21270784.

D S Adnan Majid, Adam R Aron, Wesley Thompson, Sarah Sheldon, Samar Hamza, Diederick Stoffers, Dominic Holland, Jody Goldstein, Jody Corey-Bloom, and Anders M Dale. Basal ganglia atrophy in prodromal Huntington's disease is detectable over one year using automated segmentation. *Movement disorders : official journal of the Movement Disorder Society*, 26(14):2544–51, December 2011a. ISSN 1531-8257. doi: 10.1002/mds.23912. URL http://www.ncbi.nlm.nih.gov/pubmed/21932302.

D S Adnan Majid, Diederick Stoffers, Sarah Sheldon, Samar Hamza, Wesley K Thompson, Jody Goldstein, Jody Corey-Bloom, and Adam R Aron. Automated structural imaging analysis detects premanifest Huntington's disease neurodegeneration within 1 year. *Movement disorders : official journal of the Movement Disorder Society*, 26(8):1481–8, July 2011b. ISSN 1531-8257. doi: 10.1002/mds.23656. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3136652&tool=pmcentrez&rendertype=abstract.

D S Adnan Majid, Weidong Cai, Jody Corey-Bloom, and Adam R Aron. Proactive selective response suppression is implemented via the basal ganglia. *The Journal* *of neuroscience : the official journal of the Society for Neuroscience*, 33(33):13259–69, August 2013. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.5651-12.2013. URL http://www.ncbi.nlm.nih.gov/pubmed/23946385.

GM Manguno-Mire, JI Constans, and JH Geer. Anxiety-related differences in affective categorizations of lexical stimuli. *Behaviour research and therapy*, 2005. URL https://www.sciencedirect.com/science/article/pii/S000579670400049Xhttp://www.sciencedirect.com/science/article/pii/S000579670400049X.

Elise L Mansfield, Frini Karayanidis, Sharna Jamadar, Andrew Heathcote, and Birte U Forstmann. Adjustments of response threshold during task switching: A model-based functional magnetic resonance imaging study. *The Journal of neuroscience*, 31(41), Oct 2011. URL http://www.ncbi.nlm.nih.gov/pubmed/21994385.

Dora Matzke and Eric-Jan Wagenmakers. Psychological interpretation of the ex-gaussian and shifted wald parameters: a diffusion model analysis. *Psychonomic bulletin & review*, 16, Oct 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19815782.

First MB, Spitzer RL, Gibbon M, and Williams JBW. *Structured Clinical Interview for DSM-IV-TR Axis I Disorders*. New York, Research Version, (SCID-I/P) Biometrics Research, New York State Psychiatric Institute, patient edition, 2002.

Jennifer E McDowell, Gregory G Brown, Martin Paulus, Antigona Martinez, Sara E Stewart, David J Dubowitz, and David L Braff. Neural correlates of refixation saccades and antisaccades in normal and schizophrenia subjects. *Biological psychiatry*, 51(3):216–223, 02 2002. URL http://www.ncbi.nlm.nih.gov/pubmed/11839364.

Fiona McNab and Torkek Klingberg. Prefrontal cortex and basal ganglia control access to working memory. *Nature Neuroscience*, 11(1):103–107, 12 2008. URL http://www.ncbi.nlm.nih.gov/pubmed/18066057.

Edward Meeds and Max Welling. GPS-ABC : Gaussian Process Surrogate Approximate Bayesian Computation. pages 1–17.

L. Menzies, S. Achard, S. R. Chamberlain, N. Fineberg, C. H. Chen, N. del Campo, B. J. Sahakian, T. W. Robbins, and E. Bullmore. Neurocognitive endophenotypes of obsessive-compulsive disorder. *Brain : a journal of neurology*, 130(Pt 12):3223–3236, December 2007. ISSN 1460-2156. doi: 10.1093/brain/awm205. URL http://dx.doi.org/10.1093/brain/awm205.

E K Miller and J D Cohen. An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24:167–202, 2001. URL http://www.ncbi.nlm.nih.gov/pubmed/11283309.

Austen J Milnerwood, Clare M Gladding, Mahmoud A Pouladi, Alexandra M Kaufman, Rochelle M Hines, Jamie D Boyd, Rebecca W Y Ko, Oana C Vasuta, Rona K Graham, Michael R Hayden, Timothy H Murphy, and Lynn A Raymond. Early increase in extrasynaptic NMDA receptor signaling and expression contributes to phenotype onset in Huntington's disease mice. *Neuron*, 65(2):178–90, January 2010. ISSN 1097-4199. doi: 10.1016/j.neuron.2010.01.008. URL http://www.sciencedirect.com/science/article/pii/S0896627310000139.

J. W. Mink. The basal ganglia: Focused selection and inhibition of competing motor programs. *Progress in Neurobiology*, 50:381–425, 03 1996. URL http://www.ncbi.nlm.nih.gov/pubmed/9004351.

A Miyake, N P Friedman, M J Emerson, A H Witzki, A Howerter, and T D Wager. The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. *Cognitive psychology*, 41:49–100, 09 2000. URL http://www.ncbi.nlm.nih.gov/pubmed/10945922.

P. Read Montague, Peter Dayan, and Terrence J. Sejnowski. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of Neuroscience*, 16(5):1936–1947, 01 1996. URL http://www.ncbi.nlm.nih.gov/pubmed/8774460.

PR Read Montague, RJ Raymond J. Dolan, KJ Karl J. Friston, and Peter Dayan. Computational psychiatry. *Trends in Cognitive Sciences*, 16(1):1–9, December 2011. ISSN 13646613. doi: 10.1016/j.tics.2011.11.018. URL http://linkinghub.elsevier.com/retrieve/pii/S1364661311002518http://www.sciencedirect.com/science/article/pii/S1364661311002518.

J. R. Monterosso, A. R. Aron, X. Cordova, J. Xu, and E. D. London. Deficits in response inhibition associated with chronic methamphetamine abuse. *Drug and alcohol dependence*, 79(2):273–277, August 2005. ISSN 0376-8716. doi: 10.1016/j.drugalcdep.2005.02.002. URL http://dx.doi.org/10.1016/j.drugalcdep.2005.02.002.

S. Morein-Zamir, N. A. Fineberg, T. W. Robbins, and B. J. Sahakian. Inhibition of thoughts and actions in obsessive-compulsive disorder: extending the endophenotype? *Psychological medicine*, pages 1–10, July 2009. ISSN 1469-8978. doi: 10.1017/S003329170999033X. URL http://dx.doi.org/10.1017/S003329170999033X.

Sharon Morein-Zamir and Alan Kingstone. Fixation offset and stop signal intensity effects on saccadic countermanding: a crossmodal investigation. *Experimental brain research. Experimentelle Hirnforschung. Expérimentation cérébrale*, 175(3):453–62, November 2006. ISSN 0014-4819. doi: 10.1007/s00221-006-0564-x. URL http://www.ncbi.nlm.nih.gov/pubmed/16783558.

Martijn J Mulder, Dienke Bos, Juliette M H Weusten, Janna van Belle, Sarai C van Dijk, Patrick Simen, Herman van Engeland, and Sarah Durston. Basic impairments in regulating the speed-accuracy tradeoff predict symptoms of attention-deficit/hyperactivity disorder. *Biological psychiatry*, 68(12):1114–9, December 2010a. ISSN 1873-2402. doi: 10.1016/j.biopsych.2010.07.031. URL http://www.ncbi.nlm.nih.gov/pubmed/20926067.

Martijn J. Mulder, Dienke Bos, Juliette M.H. H Weusten, Janna van Belle, Sarai C. van Dijk, Patrick Simen, Herman van Engeland, and Sarah Durston. Basic impairments in regulating the speed-accuracy tradeoff predict symptoms of attention-deficit/hyperactivity disorder. *Biological psychiatry*, 68(12):1114–9, December 2010b. ISSN 1873-2402. doi: 10.1016/j.biopsych.2010.07.031. URL http://linkinghub.elsevier.com/retrieve/pii/S0006322310008255http://www.ncbi.nlm.nih.gov/pubmed/20926067.

Yuko Munakata, Seth A. Herd, Christopher H. Chatham, Brendan E. Depue, Marie T. Banich, and Randall C. O'Reilly. A unified framework for inhibitory control. *Trends in Cognitive Sciences*, 15(10):453–459, Oct 2011. URL http://www.ncbi.nlm.nih.gov/pubmed/21889391.

Douglas P. Munoz and Stefan Everling. Look away: the anti-saccade task and the voluntary control of eye movement. *Nature Reviews Neuroscience*, 5(3):218–228, Mar 2004. URL http://www.ncbi.nlm.nih.gov/pubmed/14976521.

Kae Nakamura and Okihide Hikosaka. Role of dopamine in the primate caudate nucleus in reward modulation of saccades. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 26(20):5360–5369, 05 2006. URL http://www.ncbi.nlm.nih.gov/pubmed/16707788.

A Nambu, H Tokuno, I Hamada, H Kita, M Imanishi, T Akazawa, Y Ikeuchi, and N Hasegawa. Excitatory cortical inputs to pallidal neurons via the subthalamic nucleus in the monkey. *Journal of Neurophysiology*, 84(1):289–300, 09 2000. URL http://www.ncbi.nlm.nih.gov/pubmed/10899204.

Atsushi Nambu, Hironobu Tokuno, and Masahiko Takada. Functional significance of the cortico-subthalamo-pallidal 'hyperdirect' pathway. *Neuroscience research*, 43: 111–7, 06 2002. URL http://www.ncbi.nlm.nih.gov/pubmed/12067746.

D.J. Daniel J. Navarro and I.G. Ian G. Fuss. Fast and accurate calculations for first-passage times in Wiener diffusion models. *Journal of Mathematical Psychology*, 53 (4):222–230, August 2009. ISSN 00222496. doi: 10.1016/j.jmp.2009.02.003. URL http://linkinghub.elsevier.com/retrieve/pii/S0022249609000200.

Franz-Xaver F.-X. Neubert, Rogier B. Mars, Ethan R. Buch, Etienne Olivier, and Matthew F. S. Rushworth. Cortical and subcortical interactions during action reprogramming and their related white matter pathways. *Proceedings of the National Academy of Sciences of the United States of America*, 107(30): 13240–5, July 2010. ISSN 1091-6490. doi: 10.1073/pnas.1000674107. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2922153&tool=pmcentrez&rendertype=abstracthttp://www.pnas.org/cgi/doi/10.1073/pnas.1000674107.

Sander Nieuwenhuis, Annelies Broerse, Marjan M a Nielen, and Ritske de Jong. A goal activation approach to the study of executive function: an application to antisaccade tasks. *Brain and cognition*, 56(2):198–214, November 2004. ISSN 0278-2626. doi: 10.1016/j.bandc.2003.12.002. URL http://www.ncbi.nlm.nih.gov/pubmed/15518936.

J. T. Nigg. Is adhd a disinhibitory disorder? *Psychological bulletin*, 127(5):571–598, September 2001. ISSN 0033-2909. URL http://view.ncbi.nlm.nih.gov/pubmed/11548968.

J. T. Nigg, M. M. Wong, M. M. Martel, J. M. Jester, L. I. Puttler, J. M. Glass, K. M. Adams, H. E. Fitzgerald, and R. A. Zucker. Poor response inhibition as a predictor of problem drinking and illicit drug use in adolescents at risk for alcoholism and other substance use disorders. *Journal of the American Academy of Child and Adolescent Psychiatry*, 45(4):468–475, April 2006. ISSN 0890-8567. doi: 10.1097/01.chi.0000199028.76452.a9. URL http://dx.doi.org/10.1097/01.chi.0000199028.76452.a9.

Hå kan Nilsson, Jörg Rieskamp, and Eric-Jan Wagenmakers. Hierarchical Bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology*, 55(1):84–93, February 2011a. ISSN 00222496. doi: 10.1016/j.jmp.2010.08.006. URL http://linkinghub.elsevier.com/retrieve/pii/S0022249610001070.

Hå kan Nilsson, Jörg Rieskamp, and Eric-Jan Wagenmakers. Hierarchical Bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology*, 55(1):84–93, February 2011b. ISSN 00222496. doi: 10.1016/j.jmp.2010.08.006. URL http://linkinghub.elsevier.com/retrieve/pii/S0022249610001070.

S Nolen-Hoeksema. Responses to depression and their effects on the duration of depressive episodes. *Journal of abnormal psychology*, 100:569–582, 1991. URL http://psycnet.apa.org/journals/abn/100/4/569/.

I Noorani and R H S Carpenter. Antisaccades as decisions: LATER model predicts latency distributions and error responses. *The European journal of neuroscience*, 37 (September):1–9, November 2012. ISSN 1460-9568. doi: 10.1111/ejn.12025. URL http://www.ncbi.nlm.nih.gov/pubmed/23121177.

David Nutt and Guy Goodwin. ECNP Summit on the future of CNS drug research in Europe 2011: report prepared for ECNP by David Nutt and Guy Goodwin. *European neuropsychopharmacology : the journal of the European College of Neuropsychopharmacology*, 21(7):495–9, July 2011. ISSN 1873-7862. doi: 10.1016/j.euroneuro.2011.05.004. URL http://www.ncbi.nlm.nih.gov/pubmed/21684455.

Ignacio Obeso, Leonora Wilkinson, Enrique Casabona, Maria Luisa Bringas, Mario Alvarez, Lázaro Alvarez, Nancy Pavón, Maria-Cruz Rodríguez-Oroz, Raúl Macías, Jose a Obeso, and Marjan Jahanshahi. Deficits in inhibitory control and conflict resolution on cognitive and motor tasks in Parkinson's disease. *Experimental brain research. Experimentelle Hirnforschung. Expérimentation cérébrale*, 212(3):371–84, July 2011a. ISSN 1432-1106. doi: 10.1007/s00221-011-2736-6. URL http://www.ncbi.nlm.nih.gov/pubmed/21643718.

Ignacio Obeso, Leonora Wilkinson, and Marjan Jahanshahi. Levodopa medication does not influence motor inhibition or conflict resolution in a conditional stop-signal task in Parkinson's disease. *Experimental brain research. Experimentelle Hirnforschung. Experimentation cerebrale*, July 2011b. ISSN 1432-1106. doi: 10.1007/s00221-011-2793-x. URL http://www.ncbi.nlm.nih.gov/pubmed/21796541.

Jaap Oosterlaan, Gordon D. Logan, and Joseph A. Sergeant. Response inhibition in ad/hd, cd, comorbid ad/hd+cd, anxious, and control children: A meta-analysis of studies with the stop task. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, 39(03):411–425, 1998. URL http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=10427.

Randall C. O'Reilly and Michael J. Frank. Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, 18(2):283–328, 2006. URL http://www.ncbi.nlm.nih.gov/pubmed/16378516.

Randall C. O'Reilly and Yuko Munakata. *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain.* The MIT Press, Cambridge, MA, January 2000.

Randall C. O'Reilly, Yuko Munakata, Michael J. Frank, Thomas E. Hazy, and Contributors. *Computational Cognitive Neuroscience*. Wiki Book, 1st Edition, URL: http://ccnbook.colorado.edu, 2012. URL http://ccnbook.colorado.edu.

A Osman, S Kornblum, and D E Meyer. The point of no return in choice reaction time: controlled and ballistic stages of response preparation. *Journal of experimental psychology. Human perception and performance*, 12(3):243–258, 09 1986. URL http://www.ncbi.nlm.nih.gov/pubmed/2943853.

S. Palminteri, M. Lebreton, Y. Worbe, D. Grabli, A. Hartmann, and M. Pessiglione. Pharmacological modulation of subliminal learning in parkinson's and tourette's syndromes. *Proceedings of the National Academy of Sciences*, 2009.

Martin Pare and Doug P. Hanes. Controlled movement processing: Superior colliculus activity associated with countermanded saccades. *J. Neurosci.*, 23(16):6480–6489, July 2003. URL http://www.jneurosci.org/cgi/content/abstract/23/16/6480.

A. Parent and L. N. Hazrati. Functional anatomy of the basal ganglia. ii. the place of subthalamic nucleus and external pallidum in basal ganglia circuitry. *Brain Research. Brain Research Reviews*, 20(1):128–154, 05 1995. URL http://www.ncbi.nlm.nih.gov/pubmed/7711765.

Fabrice B R Parmentier, Gregory Elford, Carles Escera, Pilar Andrés, and Iria San Miguel. The cognitive locus of distraction by acoustic novelty in the cross-modal oddball task. *Cognition*, 106(1):408–32, January 2008. ISSN 0010-0277. doi: 10.1016/j.cognition.2007.03.008. URL http://dx.doi.org/10.1016/j.cognition.2007.03.008.

Anand Patil, David Huard, and Christopher J Fonnesbeck. PyMC: Bayesian Stochastic Modelling in Python. *Journal of statistical software*, 35(4):1–81, July 2010. ISSN 1548-7660. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3097064&tool=pmcentrez&rendertype=abstract.

Jane S Paulsen, Peggy C Nopoulos, Elizabeth Aylward, Christopher A Ross, Hans Johnson, Vincent A Magnotta, Andrew Juhl, Ronald K Pierson, James Mills, Douglas Langbehn, and Martha Nance. Striatal and white matter predictors of estimated diagnosis for Huntington disease. *Brain research bulletin*, 82(3-4):201–7, May 2010. ISSN 1873-2747. doi: 10.1016/j.brainresbull.2010.04.003. URL http://www.sciencedirect.com/science/article/pii/S0361923010000705.

Madeline Lee Pe, Joachim Vandekerckhove, and Peter Kuppens. A Diffusion Model Account of the Relationship Between the Emotional Flanker Task and Rumination and Depression. *Emotion (Washington, D.C.)*, 13(4):739–47, March 2013a. ISSN 1931-1516. doi: 10.1037/a0031628. URL http://www.ncbi.nlm.nih.gov/pubmed/23527499.

ML Pe, J Vandekerckhove, and P Kuppens. A diffusion model account of the relationship between the emotional flanker task and rumination and depression. *Emotion*, 13(739), 2013b. URL http://psycnet.apa.org/journals/emo/13/4/739/.

a Peltsch, a Hoffman, I Armstrong, G Pari, and D P Munoz. Saccadic impairments in Huntington's disease. *Experimental brain research*, 186(3):457–69, April 2008. ISSN 1432-1106. doi: 10.1007/s00221-007-1248-x. URL http://www.ncbi.nlm.nih.gov/pubmed/18185924.

R. Penadés, R. Catalán, K. Rubia, S. Andrés, M. Salamero, and C. Gastó. Impaired response inhibition in obsessive compulsive disorder. *European psychiatry : the journal of the Association of European Psychiatrists*, 22(6):404–410, September 2007. ISSN 0924-9338. doi: 10.1016/j.eurpsy.2006.05.001. URL http://dx.doi.org/10.1016/j.eurpsy.2006.05.001.

Fernando Pérez and Brian E. Granger. IPython: a System for Interactive Scientific Computing. *Computing in Science & Engineering*, 9(3):21–29, May 2007. URL http://ipython.org.

Martyn Plummer. Penalized loss functions for bayesian model comparison. *Biostatistics*, 9(3):523–539, 2008.

Dale J. Poirier. The growth of Bayesian methods in statistics and economics since 1970. *Bayesian Analysis*, 1(4):969–979, December 2006a. ISSN 1931-6690. URL http://projecteuclid.org/euclid.ba/1340370949.

Dale J. Poirier. The growth of Bayesian methods in statistics and economics since 1970. *Bayesian Analysis*, 1(4):969–979, December 2006b. ISSN 1931-6690. URL http://projecteuclid.org/euclid.ba/1340370949.

J Poland, B Von Eckardt, and W Spaulding. Problems with the DSM approach to classifying psychopathology. 1994. URL http://psycnet.apa.org/psycinfo/1995-97231-011.

Pierre Pouget, Gordon D. Logan, Thomas J. Palmeri, Leanne Boucher, Martin Pare, and Jeffrey D. Schall. Neural Basis of Adaptive Response Time Adjustment during Saccade Countermanding. *Journal of Neuroscience*, 31(35):12604–12612, August 2011. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.1868-11.2011. URL http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.1868-11.2011.

M. J. D. Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, 7(2):155–162, February 1964. ISSN 0010-4620. doi: 10.1093/comjnl/7.2.155. URL http://comjnl.oxfordjournals.org/content/7/2/155.abstract.

Andrews PW and Thomson JA. The bright side of being blue: Depression as an adaptation for analyzing complex problems. *Psychological Review*, pages 620–654, 2009.

B. Ramos and A. Arnsten. Adrenergic pharmacology and cognition: Focus on the prefrontal cortex. *Pharmacology & Therapeutics*, 113(3):523–536, March 2007. ISSN

01637258. doi: 10.1016/j.pharmthera.2006.11.006. URL http://dx.doi.org/10.1016/j.pharmthera.2006.11.006.

Julia A. Rao, Deborah L. Harrington, Sally Durgerian, Christine Reece, Lyla Mourany, Katherine Koenig, Mark J. Lowe, Vincent A. Magnotta, Jeffrey D. Long, Hans J. Johnson, Jane S. Paulsen, and Stephen M. Rao. Disruption of response inhibition circuits in prodromal Huntington disease. *Cortex*, 58:72–85, September 2014. ISSN 00109452. doi: 10.1016/j.cortex.2014.04.018. URL http://www.ncbi.nlm.nih.gov/pubmed/24959703.

R Ratcliff, MG Philiastides, and P Sajda. Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. *Proceedings of the National Academy of Sciences*, 106(16):6539–6544, 2009. URL http://www.pnas.org/content/106/16/6539.short.

Roger Ratcliff and Michael J. Frank. Reinforcement-based decision making in corticostriatal circuits: mutual constraints by neurocomputational and diffusion models. *Neural Computation*, 24(5):1186–1229, May 2012. URL http://www.ncbi.nlm.nih.gov/pubmed/22295983.

Roger Ratcliff and Gail McKoon. The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, 20:873–922, Apr 2008. URL http://www.ncbi.nlm.nih.gov/pubmed/18085991.

Roger Ratcliff and Jeffrey N. Rouder. Modeling response times for two-choice decisions. *Psychological Science*, 9:347, January 1998.

Roger Ratcliff and Francis Tuerlinckx. Estimating parameters of the diffusion model: approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic bulletin & review*, 9(3):438–81, September 2002. ISSN 1069-9384. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2474747&tool=pmcentrez&rendertype=abstract.

Roger Ratcliff, Anjali Thapar, and Gail McKoon. Individual differences, aging, and IQ in two-choice tasks. *Cognitive psychology*, 60(3):127–57, May 2010. ISSN 1095-5623. doi: 10.1016/j.cogpsych.2009.09.001. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2835850&tool=pmcentrez&rendertype=abstract.

N. J. Ray, N. Jenkinson, J. Brittain, P. Holland, C. Joint, D. Nandi, P. G. Bain, N. Yousif, A. Green, J. S. Stein, and T. Z. Aziz. The role of the subthalamic nucleus in response inhibition: evidence from deep brain stimulation for parkinson's disease. *Neuropsychologia*, 47(13):2828–2834, Nov 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19540864.

J. L. Reilly, M. S. Harris, M. S. Keshavan, and J. A. Sweeney. Adverse effects of risperidone on spatial working memory in first-episode schizophrenia. *Archives of general psychiatry*, 63(11):1189–1197, November 2006. ISSN 0003-990X. doi: 10.1001/archpsyc.63.11.1189. URL http://dx.doi.org/10.1001/archpsyc.63.11.1189.

J. L. Reilly, M. S. Harris, T. T. Khine, M. S. Keshavan, and J. A. Sweeney. Antipsychotic drugs exacerbate impairment on a working memory task in first-episode schizophrenia. *Biological psychiatry*, 62(7):818–821, October 2007. ISSN 0006-3223. doi: 10.1016/j.biopsych.2006.10.031. URL http://dx.doi.org/10.1016/j.biopsych.2006.10.031.

Benedikt Reuter and Norbert Kathmann. Using saccade tasks as a tool to analyze executive dysfunctions in schizophrenia. *Acta psychologica*, 115(2-3):255–69, 2004. ISSN 0001-6918. doi: 10.1016/j.actpsy.2003.12.009. URL http://www.ncbi.nlm.nih.gov/pubmed/14962403.

K. Richard Ridderinkhof, Birte U. Forstmann, Scott a. Wylie, Borís Burle, and Wery P. M. van den Wildenberg. Neurocognitive mechanisms of action control: resisting the call of the Sirens. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(2):

174–192, March 2011. ISSN 19395078. doi: 10.1002/wcs.99. URL http://doi.wiley.com/10.1002/wcs.99.

K Richard Ridderinkhof. Micro- and macro-adjustments of task set: activation and suppression in conflict tasks. *Psychological research*, 66(4):312–23, November 2002. ISSN 0340-0727. doi: 10.1007/s00426-002-0104-7. URL http://www.ncbi.nlm.nih.gov/pubmed/12466928http://academic.research.microsoft.com/Publication/2054125/activation-and-suppression-in-conflict-tasks-empirical-clarification-through-dis
//dare.uva.nl/record/122002.

K Richard Ridderinkhof, Wery P M van den Wildenberg, Sidney J Segalowitz, and Cameron S Carter. Neurocognitive mechanisms of cognitive control: the role of prefrontal cortex in action selection, response inhibition, performance monitoring, and reward-based learning. *Brain and cognition*, 56(2):129–40, November 2004. ISSN 0278-2626. doi: 10.1016/j.bandc.2004.09.016. URL http://www.ncbi.nlm.nih.gov/pubmed/15518930.

K Richard Ridderinkhof, Anouk Scheres, Jaap Oosterlaan, and Joseph a Sergeant. Delta plots in the study of individual differences: new tools reveal response inhibition deficits in AD/Hd that are eliminated by methylphenidate treatment. *Journal of abnormal psychology*, 114(2):197–215, May 2005. ISSN 0021-843X. doi: 10.1037/0021-843X.114.2.197. URL http://www.ncbi.nlm.nih.gov/pubmed/15869351.

K Richard Ridderinkhof, Birte U. Forstmann, Scott a. Wylie, Borís Burle, Wery P. M. van den Wildenberg, and K. Richard Ridderinkhof. Neurocognitive mechanisms of action control: resisting the call of the Sirens. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(2):174–192, March 2011. ISSN 19395078. doi: 10.1002/wcs.99. URL http://doi.wiley.com/10.1002/wcs.99.

Trevor W Robbins, Claire M Gillan, Dana G Smith, Sanne de Wit, and Karen D

Ersche. Neurocognitive endophenotypes of impulsivity and compulsivity: towards dimensional psychiatry. *Trends in cognitive sciences*, 16(1):81–91, January 2012. ISSN 1879-307X. doi: 10.1016/j.tics.2011.11.009. URL http://www.ncbi.nlm.nih.gov/pubmed/22155014.

Ralph J. Roberts, Lisa D. Hager, and Christine Heron. Prefrontal cognitive processes: Working memory and inhibition in the antisaccade task. *Journal of Experimental Psychology: General*, 123:374, January 1994.

MD Robinson and BP Meier. Introversion, inhibition, and displayed anxiety: The role of error reactivity processes. *Journal of Research in . . .*, 41:558–578, 2007. URL http://www.sciencedirect.com/science/article/pii/S0092656606000845.

N. P. Rougier, D. Noelle, T. S. Braver, J. D. Cohen, and R. C. O'Reilly. Prefrontal cortex and the flexibility of cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences*, 102(20):7338–7343, January 2005.

J. Rowe, K. Friston, R. Frackowiak, and R. Passingham. Attention to action: specific modulation of corticocortical interactions in humans. *NeuroImage*, 17(2):988–998, October 2002. ISSN 1053-8119. URL http://view.ncbi.nlm.nih.gov/pubmed/12377172.

Leonid L Rubchinsky, Nancy Kopell, and Karen A Sigvardt. Modeling facilitation and inhibition of competing motor programs in basal ganglia subthalamic nucleus-pallidal circuits. *Proceedings of the National Academy of Sciences of the United States of America*, 100:14427–32, 12 2003. URL http://www.ncbi.nlm.nih.gov/pubmed/14612573.

AJ Rush, MH Trivedi, and HM Ibrahim. 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic. *Biological . . .*, 54:

573–583, 2003. URL http://www.sciencedirect.com/science/article/pii/S0006322302018668.

Barbara J Sahakian, Gavin Malloch, and Christopher Kennard. A UK strategy for mental health and wellbeing. *Lancet*, 375(9729):1854–5, May 2010. ISSN 1474-547X. doi: 10.1016/S0140-6736(10)60817-3. URL http://www.thelancet.com/journals/a/article/PIIS0140-6736(10)60817-3/fulltext.

M. Sakagami, Tsutsui K., J. Lauwereyns, M. Koizumi, S. Kobayashi, and O. Hikosaka. A code for behavioral inhibition on the basis of color, but not motion, in ventrolateral prefrontal cortex of macaque monkey. *The Journal of Neuroscience*, 21(13): 4801–4808, Jul 2001. URL http://www.ncbi.nlm.nih.gov/pubmed/11425907.

Russell Schachar and Gordon D. Logan. Impulsivity and inhibitory control in normal development and childhood psychopathology. *Developmental Psychology*, 26(5): 710–720, 1990. ISSN 0012-1649. doi: 10.1037/0012-1649.26.5.710. URL http://dx.doi.org/10.1037/0012-1649.26.5.710.

J. Schlag and M. Schlag-Rey. Evidence for a supplementary eye field. *J Neurophysiol*, 57(1):179–200, January 1987. URL http://jn.physiology.org/cgi/content/abstract/57/1/179.

M. Schlag-Rey and J. Schlag. Visuomotor functions of central thalamus in monkey. I. Unit activity related to spontaneous eye movements. *J Neurophysiol*, 51(6):1149–1174, June 1984. URL http://jn.physiology.org/cgi/content/abstract/51/6/1149.

M Schlag-Rey, N Amador, H Sanchez, and J Schlag. Antisaccade performance predicted by neuronal activity in the supplementary eye field. *Nature*, 390:398, 12 1997. URL http://www.ncbi.nlm.nih.gov/pubmed/9389478.

Robert Schmidt, Daniel Leventhal, Jeff Pettibone, Alaina Case, and Joshua Berke.

Suppressing Actions in the Basal Ganglia. In *The 9th annual Computational and Systems Neuroscience meeting*, page 139, 2012.

W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 03 1997. URL http://www.ncbi.nlm.nih.gov/pubmed/9054347.

Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, March 1978. ISSN 2168-8966. URL http://projecteuclid.org/euclid.aos/1176344136.

W Schwarz. The ex-Wald distribution as a descriptive model of response times. *Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc*, 33(4):457–69, November 2001. ISSN 0743-3808. URL http://www.ncbi.nlm.nih.gov/pubmed/11816448.

J Sethuraman. A constructive definition of Dirichlet priors. 1991. URL http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA238689.

D J Sharp, V Bonnelle, X De Boissezon, C F Beckmann, S G James, M C Patel, and M A Mehta. Distinct frontal systems for response inhibition, attentional capture, and error processing. *Proceedings of the National Academy of Sciences of the United States of America*, Mar 2010. URL http://www.ncbi.nlm.nih.gov/pubmed/20220100.

Weixing Shen, Marc Flajolet, Paul Greengard, and D James Surmeier. Dichotomous dopaminergic control of striatal synaptic plasticity. *Science (New York, N.Y.)*, 321(5890):848–851, 08 2008. URL http://www.ncbi.nlm.nih.gov/pubmed/18687967.

RM Shiffrin, MD Lee, and W Kim. A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive ...*, 2008a. URL http://onlinelibrary.wiley.com/doi/10.1080/03640210802414826/full.

RM Shiffrin, MD Lee, and W Kim. A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive ...*, 2008b. URL http://onlinelibrary.wiley.com/doi/10.1080/03640210802414826/full.

D. Simmonds, J. Pekar, and S. Mostofsky. Meta-analysis of go/no-go tasks demonstrating that fmri activation associated with response inhibition is task-dependent. *Neuropsychologia*, 46(1):224–232, 2008. ISSN 00283932. doi: 10.1016/j.neuropsychologia.2007.07.015. URL http://dx.doi.org/10.1016/j.neuropsychologia.2007.07.015.

J R Simon. Reactions toward the source of stimulation. *Journal of experimental psychology*, 81:174–176, 10 1969. URL http://www.ncbi.nlm.nih.gov/pubmed/5812172.

S. A. Sloman. The empirical case for two systems of reasoning. *Pscyhological Bulletin*, 119:3–22, January 1996.

Philip L Smith and Roger Ratcliff. Psychology and neurobiology of simple decisions. *Trends in neurosciences*, 27(3):161–8, March 2004. ISSN 0166-2236. doi: 10.1016/j.tins.2004.01.006. URL http://www.ncbi.nlm.nih.gov/pubmed/15036882.

Pl Smith. Stochastic Dynamic Models of Response Time and Accuracy: A Foundational Primer. *Journal of mathematical psychology*, 44(3):408–463, September 2000. ISSN 0022-2496. doi: 10.1006/jmps.1999.1260. URL http://www.ncbi.nlm.nih.gov/pubmed/10973778.

RP Snaith, M Hamilton, and S Morley. A scale for the assessment of hedonic tone the Snaith-Hamilton Pleasure Scale. *The British Journal of ...*, 167:99–103, 1995. URL http://bjp.rcpsych.org/content/167/1/99.short.

HR Snyder. Major depressive disorder is associated with broad impairments on neuropsychological measures of executive function: a meta-analysis and review. *Psychological bulletin*, 139:81–132, 2013. URL http://psycnet.apa.org/psycarticles/2012-13786-001.

HR Snyder and RH Kaiser. Opposite effects of anxiety and depressive symptoms on executive function: The case of selecting among competing options. *Cognition & ...*, 28:893–902, 2014. URL http://www.tandfonline.com/doi/abs/10.1080/02699931.2013.859568.

Marc A Sommer and Robert H Wurtz. A pathway in primate brain for internal monitoring of movements. *Science (New York, N.Y.)*, 296:1480–1482, 05 2002. URL http://www.ncbi.nlm.nih.gov/pubmed/12029137.

Marc A Sommer and Robert H Wurtz. What the brain stem tells the frontal cortex. i. oculomotor signals sent from superior colliculus to frontal eye field via mediodorsal thalamus. *Journal of neurophysiology*, 91(3):1381–1402, Mar 2004a. URL http://www.ncbi.nlm.nih.gov/pubmed/14573558.

Marc A Sommer and Robert H Wurtz. What the brain stem tells the frontal cortex. I. Oculomotor signals sent from superior colliculus to frontal eye field via mediodorsal thalamus. *Journal of neurophysiology*, 91(3):1381–402, March 2004b. ISSN 0022-3077. URL http://www.ncbi.nlm.nih.gov/pubmed/14573558.

Marc A Sommer and Robert H Wurtz. Influence of the thalamus on spatial visual processing in frontal cortex. *Nature*, 444(7117):374–7, November 2006. ISSN 1476-4687. URL http://dx.doi.org/10.1038/nature05279.

D. L. Sparks. The brainstem control of saccadic eye movements. *Nature Reviews Neuroscience*, 3:952–964, January 2002.

David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and Angelika van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Sta-*

tistical Society: Series B (Statistical Methodology), 64(4):583–639, October 2002a. ISSN 1369-7412. doi: 10.1111/1467-9868.00353. URL http://doi.wiley.com/10.1111/1467-9868.00353.

DJ Spiegelhalter, NG Best, and Bradley P. Carlin. Bayesian measures of model complexity and fit. *Journal of the Royal . . .*, 2002b. URL http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00353/full.

Klaas Enno Stephan, Will D Penny, Jean Daunizeau, Rosalyn J Moran, and Karl J Friston. Bayesian model selection for group studies. *NeuroImage*, 46 (4):1004–17, July 2009. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2009.03.025. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2703732&tool=pmcentrez&rendertype=abstract.

Matthew Stephens and David J Balding. Bayesian statistical methods for genetic association studies. *Nature reviews. Genetics*, 10(10):681–90, October 2009a. ISSN 1471-0064. doi: 10.1038/nrg2615. URL http://dx.doi.org/10.1038/nrg2615.

Matthew Stephens and David J Balding. Bayesian statistical methods for genetic association studies. *Nature reviews. Genetics*, 10(10):681–90, October 2009b. ISSN 1471-0064. doi: 10.1038/nrg2615. URL http://dx.doi.org/10.1038/nrg2615.

A Stevens. Event-related fMRI of auditory and visual oddball tasks. *Magnetic Resonance Imaging*, 18(5):495–502, June 2000. ISSN 0730725X. doi: 10.1016/S0730-725X(00)00128-4. URL http://dx.doi.org/10.1016/S0730-725X(00)00128-4http://linkinghub.elsevier.com/retrieve/pii/S0730725X00001284.

Gregory P. Strauss, Michael J. Frank, James a. Waltz, Zuzana Kasanova, Ellen S. Herbener, and James M. Gold. Deficits in Positive Reinforcement Learning and Uncertainty-Driven Exploration Are Associated with Distinct Aspects of Negative Symptoms in Schizophrenia. *Biological Psychiatry*, 69(5):424–431, March 2011a. ISSN 00063223. doi: 10.1016/j.biopsych.2010.10.015. URL http://dx.doi.org/10.1016/j.biopsych.2010.10.015http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3039035&tool=pmcentrez&rendertype=abstracthttp://linkinghub.elsevier.com/retrieve/pii/S0006322310011066.

Gregory P. Strauss, Michael J. Frank, James a. Waltz, Zuzana Kasanova, Ellen S. Herbener, and James M. Gold. Deficits in Positive Reinforcement Learning and Uncertainty-Driven Exploration Are Associated with Distinct Aspects of Negative Symptoms in Schizophrenia. *Biological Psychiatry*, 69(5):424–431, March 2011b. ISSN 00063223. doi: 10.1016/j.biopsych.2010.10.015. URL http://linkinghub.elsevier.com/retrieve/pii/S0006322310011066.

V Stuphorn, T L Taylor, and J D Schall. Performance monitoring by the supplementary eye field. *Nature*, 408:857, 12 2000. URL http://www.ncbi.nlm.nih.gov/pubmed/11130724.

Veit Stuphorn and Jeffrey D. Schall. Executive control of countermanding saccades by the supplementary eye field. *Nat Neurosci*, 9(7):925–931, July 2006. ISSN 1097-6256. doi: 10.1038/nn1714. URL http://dx.doi.org/10.1038/nn1714.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, January 1998. URL http://www.cs.ualberta.ca/~sutton/book/ebook/the-book.html.

Nicole Swann, Howard Poizner, Melissa Houser, Sherrie Gould, Ian Greenhouse, Weidong Cai, Jon Strunk, Jobi George, and Adam R Aron. Deep brain stimulation of the subthalamic nucleus alters the cortical profile of response inhibition in the beta frequency band: a scalp EEG study in Parkinson's disease. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 31

(15):5721–9, April 2011a. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.6135-10.2011. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3086079&tool=pmcentrez&rendertype=abstract.

Nicole C Swann, Weidong Cai, Christopher R Conner, Thomas A Pieters, Michael P Claffey, Jobi S George, Adam R Aron, and Nitin Tandon. Roles for the pre-supplementary motor area and the right inferior frontal gyrus in stopping action : electrophysiological responses and functional and structural connectivity. *NeuroImage*, September 2011b. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2011.09.049. URL http://linkinghub.elsevier.com/retrieve/pii/S1053811911011141.

Sarah J Tabrizi, Douglas R Langbehn, Blair R Leavitt, Raymund Ac Roos, Alexandra Durr, David Craufurd, Christopher Kennard, Stephen L Hicks, Nick C Fox, Rachael I Scahill, Beth Borowsky, Allan J Tobin, H Diana Rosas, Hans Johnson, Ralf Reilmann, Bernhard Landwehrmeyer, and Julie C Stout. Biological and clinical manifestations of Huntington's disease in the longitudinal TRACK-HD study: cross-sectional analysis of baseline data. *Lancet neurology*, 8(9):791–801, September 2009. ISSN 1474-4422. doi: 10.1016/S1474-4422(09)70170-X. URL http://www.sciencedirect.com/science/article/pii/S147444220970170X.

Sarah J Tabrizi, Rachael I Scahill, Gail Owen, Alexandra Durr, Blair R Leavitt, Raymund a Roos, Beth Borowsky, Bernhard Landwehrmeyer, Chris Frost, Hans Johnson, David Craufurd, Ralf Reilmann, Julie C Stout, and Douglas R Langbehn. Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington's disease in the TRACK-HD study: analysis of 36-month observational data. *Lancet neurology*, 12(7):637–49, July 2013. ISSN 1474-4465. doi: 10.1016/S1474-4422(13)70088-7. URL http://www.ncbi.nlm.nih.gov/pubmed/23664844.

Stefano Taverna, Ema Ilijic, and D. James Surmeier. Recurrent collateral connections of striatal medium spiny neurons are disrupted in models of Parkinson's disease. *The Journal of Neuroscience*, 28(21):5504–5512, 5 2008. URL http://www.jneurosci.org/cgi/content/abstract/28/21/5504.

James T. Townsend and F. Gregory Ashby. *The stochastic modeling of elementary psychological processes*. Cambridge University Press, 1983a. ISBN 0521241812. doi: 102-212-801.

James T. Townsend and F. Gregory Ashby. *The stochastic modeling of elementary psychological processes*. Cambridge University Press, 1983b. ISBN 0521241812. doi: 102-212-801.

Brandon M Turner and Per B Sederberg. A generalized, likelihood-free method for posterior estimation. *Psychonomic bulletin & review*, November 2013. ISSN 1531-5320. doi: 10.3758/s13423-013-0530-0. URL http://www.ncbi.nlm.nih.gov/pubmed/24258272.

Brandon M. Turner and Trisha Van Zandt. A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology*, 56(2):69–85, April 2012. ISSN 00222496. doi: 10.1016/j.jmp.2012.02.005. URL http://linkinghub.elsevier.com/retrieve/pii/S0022249612000272.

Maia Tiago V. Two-factor theory the actor-critic model and conditioned avoidance. *Learning & behavior*, 38(1):50–67, Feb 2010. URL http://www.ncbi.nlm.nih.gov/pubmed/20065349.

Maia Tiago V. and Frank Michael J. From reinforcement learning models to psychiatric and neurological disorders. *Nature neuroscience*, 14(2):154–162, Feb 2011. URL http://www.ncbi.nlm.nih.gov/pubmed/21270784.

Martijn G. van Koningsbruggen, Tom Pender, Liana Machado, and Robert D. Rafal. Impaired control of the oculomotor reflexes in parkinson's disease. *Neuropsychologia*, June 2009. ISSN 1873-3514. doi: 10.1016/j.neuropsychologia.2009.06.018. URL http://dx.doi.org/10.1016/j.neuropsychologia.2009.06.018.

Leendert van Maanen, Scott D Brown, Tom Eichele, Eric-Jan Wagenmakers, Tiffany Ho, John Serences, and Birte U Forstmann. Neural Correlates of Trial-to-Trial Fluctuations in Response Caution. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 31(48):17488–17495, November 2011. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.2924-11.2011. URL http://www.ncbi.nlm.nih.gov/pubmed/22131410.

Joachim Vandekerckhove and Francis Tuerlinckx. Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods*, 40(1):61–72, February 2008. ISSN 1554-351X. doi: 10.3758/BRM.40.1.61. URL http://brm.psychonomic-journals.org/cgi/doi/10.3758/BRM.40.1.61.

Joachim Vandekerckhove, Francis Tuerlinckx, and Michael D Lee. Hierarchical diffusion models for two-choice response times. *Psychological methods*, 16(1):44–62, March 2011. ISSN 1939-1463. doi: 10.1037/a0021765. URL http://www.ncbi.nlm.nih.gov/pubmed/21299302.

Elena M Vazey and Gary Aston-Jones. The emerging role of norepinephrine in cognitive dysfunctions of Parkinson's disease. *Frontiers in behavioral neuroscience*, 6:48, January 2012. ISSN 1662-5153. doi: 10.3389/fnbeh.2012.00048. URL http://www.ncbi.nlm.nih.gov/pubmed/22848194.

Frederick Verbruggen and Gordon D. Logan. Response inhibition in the stop-signal paradigm. *Trends in Cognitive Sciences*, 12:418–424, Nov 2008. URL http://www.ncbi.nlm.nih.gov/pubmed/18799345.

Frederick Verbruggen and Gordon D. Logan. Models of response inhibition in the stop-signal and stop-change paradigms. *Neuroscience & Biobehavioral Reviews*, 33 (5):647–661, May 2009a. ISSN 01497634. doi: 10.1016/j.neubiorev.2008.08.014. URL http://dx.doi.org/10.1016/j.neubiorev.2008.08.014.

Frederick Verbruggen and Gordon D Logan. Models of response inhibition in the stop-signal and stop-change paradigms. *Neuroscience and biobehavioral reviews*, 33:647–661, May 2009b. URL http://www.ncbi.nlm.nih.gov/pubmed/18822313.

Frederick Verbruggen, Adam R Aron, Michal A Stevens, and Christopher D Chambers. Theta burst stimulation dissociates attention and action updating in human inferior frontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 107, Aug 2010. URL http://www.ncbi.nlm.nih.gov/pubmed/20631303.

D Vickers. Evidence for an accumulator model of psychophysical discrimination. *Ergonomics*, 13(1):37–58, January 1970. ISSN 0014-0139. doi: 10.1080/00140137008931117. URL http://dx.doi.org/10.1080/00140137008931117.

V. Voon, M. Pessiglione, C. Brezing, C. Gallea, H. H. Fernandez, R. J. Dolan, and M. Hallett. Mechanisms underlying dopamine-mediated reward bias in compulsive behaviors. *Neuron*, 65:135–142, 2010.

a Voss and J Voss. A fast numerical algorithm for the estimation of diffusion model parameters. *Journal of Mathematical Psychology*, 52:1–9, November 2007. ISSN 00222496. doi: 10.1016/j.jmp.2007.09.005. URL http://linkinghub.elsevier.com/retrieve/pii/S0022249607000685.

Eric-Jan Wagenmakers, Han L. J. Maas, and Raoul P. P. P. Grasman. An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, 14(1):3–22, February 2007. ISSN 1069-9384. doi: 10.3758/BF03194023. URL http://www.springerlink.com/index/10.3758/BF03194023.

Eric-Jan Wagenmakers, Tom Lodewyckx, Himanshu Kuriyal, and Raoul Grasman. Bayesian hypothesis testing for psychologists: a tutorial on the savage-dickey method. *Cognitive psychology*, 60, May 2010. URL http://www.ncbi.nlm.nih.gov/pubmed/20064637.

G Wagner, E Sinsel, T Sobanski, and S Köhler. Cortical inefficiency in patients with unipolar depression: an event-related FMRI study with the Stroop task. *Biological ...*, 59:958–965, 2006. URL http://www.sciencedirect.com/science/article/pii/S0006322305014344.

A. Wald. *Sequential Analysis.* Wiley, 1947.

David J. Wales and Jonathan P. K. Doye. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *The Journal of Physical Chemistry A*, 101(28):5111–5116, July 1997. ISSN 1089-5639. doi: 10.1021/jp970984n. URL http://dx.doi.org/10.1021/jp970984n.

Jonathan D Wallis and Earl K Miller. Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. *The European journal of neuroscience*, 18:2069–81, 11 2003. URL http://www.ncbi.nlm.nih.gov/pubmed/14622240.

James A Waltz, Michael J Frank, Benjamin M Robinson, and James M Gold. Selective reinforcement learning deficits in schizophrenia support predictions from computational models of striatal-cortical dysfunction. *Biological psychiatry*, 62: 756–764, 09 2007. URL http://www.ncbi.nlm.nih.gov/pubmed/17300757.

James A. Waltz, Michael J. Frank, Thomas V. Wiecki, and James M. Gold. Altered probabilistic learning and response biases in schizophrenia: behavioral evidence and neurocomputational modeling. *Neuropsychology*, 25(1):86–97, Jan 2011. URL http://www.ncbi.nlm.nih.gov/pubmed/21090899.

Yan Wang, Masaki Isoda, Yoshiya Matsuzaka, Keisetsu Shima, and Jun Tanji. Prefrontal cortical cells projecting to the supplementary eye field and presupplementary motor area in the monkey. *Neuroscience research*, 53(1):1–7, September 2005. ISSN 0168-0102. doi: 10.1016/j.neures.2005.05.005. URL http://www.ncbi.nlm.nih.gov/pubmed/15992955.

Masayuki Watanabe and Douglas P. Munoz. Neural correlates of conflict resolution between automatic and volitional actions by basal ganglia. *European Journal of Neuroscience*, 30(11):2165–2176, 2009. URL http://dx.doi.org/10.1111/j.1460-9568.2009.06998.x.

Masayuki Watanabe and Douglas P. Munoz. Presetting basal ganglia for volitional actions. *The Journal of Neuroscience*, 30(2), Jul 2010. URL http://www.ncbi.nlm.nih.gov/pubmed/20668198.

Masayuki Watanabe and Douglas P. Munoz. Probing basal ganglia functions by saccade eye movements. *The European Journal of Neuroscience*, 33(11):2070–2090, Jun 2011. URL http://www.ncbi.nlm.nih.gov/pubmed/21645102.

Stephen P Wegener, Kevin Johnston, and Stefan Everling. Microstimulation of monkey dorsolateral prefrontal cortex impairs antisaccade performance. *Experimental brain research. Experimentelle Hirnforschung. Expérimentation cérébrale*, 190(4):463–73, October 2008. ISSN 1432-1106. doi: 10.1007/s00221-008-1488-4. URL http://www.ncbi.nlm.nih.gov/pubmed/18641976.

CN Corey N White, Roger Ratcliff, MW Michael W Vasey, and Gail McKoon. Using diffusion models to understand clinical disorders. *Journal of mathematical psychology*, 54(1):39–52, February 2010a. ISSN 0022-2496. doi: 10.1016/j.jmp.2010.01.004. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2859713&tool=pmcentrez&rendertype=abstracthttp://www.sciencedirect.com/science/article/pii/S0022249610000155.

Corey N White, Roger Ratcliff, Michael W Vasey, and Gail McKoon. Anxiety enhances threat processing without competition among multiple inputs: a diffusion model analysis. *Emotion (Washington, D.C.)*, 10(5):662–77, October 2010b.

ISSN 1931-1516. doi: 10.1037/a0019474. URL http://www.ncbi.nlm.nih.gov/pubmed/21038949.

Corey N. Whitea, Roger Ratcliff, Jeffrey J S. Starns, and Corey N White. Diffusion models of the flanker task: Discrete versus gradual attentional selection. *Cognitive Psychology*, 63(1989):210–238, September 2010a. ISSN 1095-5623. doi: 10.1016/j.cogpsych.2011.08.001. URL http://dx.doi.org/10.1016/j.cogpsych.2011.08.001.

Corey N. Whitea, Roger Ratcliff, and Jeffrey S. Starns. Diffusion models of the flanker task: Discrete versus gradual attentional selection. *Cognitive Psychology*, (1989), 2010b.

Thomas V. Wiecki and Michael J. Frank. A computational model of inhibitory control in frontal cortex and basal ganglia. *Psychological Review*, 120(2):329–355, Apr 2013. URL http://www.ncbi.nlm.nih.gov/pubmed/23586447.

Thomas V. Wiecki, Jeffrey Poland, and Michael J. Frank. Sequential sampling models in computational psychiatry: Bayesian parameter estimation and classification. a.

Thomas V. Wiecki, Jeffrey Poland, and Michael J. Frank. Model-based cognitive neuroscience approaches to computational psychiatry: clustering and classification. b.

Thomas V Wiecki, Imri Sofer, and Michael J Frank. HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in python. c.

Thomas V Wiecki, Imri Sofer, and Michael J Frank. HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in neuroinformatics*, 7:14, January 2013a. ISSN 1662-5196. doi: 10.3389/fninf.2013.00014. URL http://www.frontiersin.org/Neuroinformatics/10.3389/fninf.2013.00014/abstracthttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3731670&tool=pmcentrez&rendertype=abstract.

Thomas V Wiecki, Imri Sofer, and Michael J Frank. Hddm: Hierarchical bayesian estimation of the drift-diffusion model in python. *Frontiers in neuroinformatics*, 7, 2013b. URL http://www.ncbi.nlm.nih.gov/pubmed/23935581.

T.V. Wiecki and M.J. Frank. Neurocomputational models of motor and cognitive deficits in Parkinson's disease. In Anders Bjorklund and M. Angela Cenci, editors, *Progress in Brain Research: Recent Advances in Parkinson's Disease - Part I: Basic Research*, volume 183, chapter 14, pages 275–297. Elsevier, 2010. URL http://www.ncbi.nlm.nih.gov/pubmed/20696325.

S Windmann and T Krüger. Subconscious detection of threat as reflected by an enhanced response bias. *Consciousness and Cognition*, 1998. URL https://www.sciencedirect.com/science/article/pii/S1053810098903373http://www.sciencedirect.com/science/article/pii/S1053810098903373.

Simon N Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–4, August 2010. ISSN 1476-4687. doi: 10.1038/nature09319. URL http://www.ncbi.nlm.nih.gov/pubmed/20703226.

Scott A. Wylie, K. Richard Ridderinkhof, William J. Elias, Robert C. Frysinger, Theodore R. Bashore, Kara E. Downs, Nelleke C. van Wouwe, and Wery P. M. van den Wildenberg. Subthalamic nucleus stimulation influences expression and suppression of impulsive behaviour in parkinson's disease. *Brain*, Sep 2010. URL http://www.ncbi.nlm.nih.gov/pubmed/20861152.

Gui Xue, Adam R. Aron, and Russell A. Poldrack. Common neural substrates for inhibition of spoken and manual responses. *Cereb. Cortex*, 18(8):1923–1932, August 2008. doi: 10.1093/cercor/bhm220. URL http://dx.doi.org/10.1093/cercor/bhm220.

N. Yeung and J. D. Cohen. The impact of cognitive deficits on conflict monitoring. predictable dissociations between the error-related negativity and n2. *Psychological* *science : a journal of the American Psychological Society / APS*, 17(2):164–171, February 2006. ISSN 0956-7976. doi: 10.1111/j.1467-9280.2006.01680.x. URL http://dx.doi.org/10.1111/j.1467-9280.2006.01680.x.

Nick Yeung, Matthew M Botvinick, and Jonathan D Cohen. The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychological review*, 111(4):931–959, 10 2004a. URL http://www.ncbi.nlm.nih.gov/pubmed/15482068.

Nick Yeung, Matthew M Botvinick, and Jonathan D Cohen. The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychological review*, 111(4):931–959, 10 2004b. URL http://www.ncbi.nlm.nih.gov/pubmed/15482068.

Kareem A Zaghloul, Christoph T Weidemann, Bradley C Lega, Jurg L Jaggi, Gordon H Baltuch, and Michael J Kahana. Neuronal activity in the human subthalamic nucleus encodes decision conflict during action selection. *The Journal of neuroscience*, 32, Feb 2012. URL http://www.ncbi.nlm.nih.gov/pubmed/22396419.

Bram B Zandbelt and Matthijs Vink. On the role of the striatum in response inhibition. *PloS one*, 5(11):e13848, January 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0013848. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2973972&tool=pmcentrez&rendertype=abstract.

Baltazar Zavala, John-Stuart Brittain, Ned Jenkinson, Keyoumars Ashkan, Thomas Foltynie, Patricia Limousin, Ludvic Zrinzo, Alexander L Green, Tipu Aziz, Kareem Zaghloul, and Peter Brown. Subthalamic nucleus local field potential activity during the Eriksen flanker task reveals a novel role for theta phase during conflict monitoring. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 33(37):14758–66, September 2013. ISSN 1529-2401. doi:

10.1523/JNEUROSCI.1036-13.2013. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3771028&tool=pmcentrez&rendertype=abstract.