DISS. ETH NO 25090

TRANSLATIONAL METHODS AND MODELS FOR COMPUTATIONAL PSYCHIATRY

A thesis submitted to attain the degree of

DOCTOR OF SCIENCE of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

EDUARDO ALBERTO APONTE PEREZ

Msc. University of Osnabrück

Born on 27.10.1987

Citizen of Colombia

accepted on the recommendation of

Prof. Dr. Klaas E. Stephan (examiner)

Prof. Dr. Peter König (co-examiner)

Dr. Jakob Heinzle (co-examiner)

2018

Table of Contents

Abstract

Computational psychiatry is a novel field devoted to improving the understanding and treatment of psychiatric disorders through quantitative methods. A significant part of this endeavor is the formulation and evaluation of models that describe relevant physiological and cognitive processes. In this dissertation, we aim to contribute to this field in two manners. First, we present and evaluate methods for model selection with an emphasis on models of brain connectivity. Our main contribution is to implement and extend thermodynamic integration, a method that has not been used in computational neuroscience in the past. Second, we develop statistical models of eye movements in two paradigms that are relevant for psychiatric research: the double step and the antisaccade task. The models developed here are probabilistic in nature and therefore we use the methods presented in the first section to solve several empirical questions related to learning, inhibitory control, and rule guided behavior in the oculomotor system.

Zusammenfassung

Komputationale Psychiatrie ist ein neues Forschungsgebiet, das sich dem Verständnis und der Behandlung von psychiatrischen Erkrankungen durch quantitative Methoden widmet. Ein signifikanter Teil dieses Bestrebens besteht in der Formulierung und Evaluierung von Modellen, welche relevante psychische und kognitive Prozesse beschreiben. In dieser Dissertation bestreben wir, in diesem Feld auf zwei Arten beizutragen. Zum einen präsentieren und evaluieren wir Methoden des Modellvergleichs, mit einem Fokus auf Modellen für Hirnkonnektivität. Unser primärer Beitrag besteht in der Implementierung und Erweiterung der thermodynamischen Integration, einer Methode welche in der Vergangenheit noch keine Anwendung in der Komputationalen Neurowissenschaft gefunden hat. Zum anderen entwickeln wir statistische Modelle von Augenbewegungen in zwei Paradigmen, die beide relevant für die psychiatrische Forschung sind: die Doppelschritt- und Antisakkadenaufgabe. Unsere Modelle sind in ihrer Natur probabilistisch, weshalb wir die Methoden des ersten Teils zur Beantwortung mehrerer empirischer Fragen bezüglich Lernens, inhibitorischer Kontrolle und regelbasierten Verhaltens im okulomotorischen System, anwenden.

Introduction

Computational psychiatry is a field devoted to the development and application of mathematical and quantitative methods to psychiatric research. Part of this enterprise is the formulation of models, or abstract representations of physiological and cognitive processes that are relevant to understanding mental health. Because models constitute scientific hypotheses, a central task for computational psychiatry is evaluating and comparing competing models of the same phenomena.

This dissertation aims to contribute to this young field in two manners. First, we consider the problem of how to decide between formal models applied to the same data using probabilistic methods. We undertake this task within the conceptual framework of Bayesian statistics, which offers one general metric of the adequacy of a model: its posterior probability. As we will clarify further in this thesis, despite being an intuitive concept, the probability of a model, known as the model evidence, is not a quantity that can be easily computed. The first part of this dissertation explores one option to estimate the model evidence that has received little attention in computational neuroscience: thermodynamic integration (TI). Despite its large computational costs, we hope to demonstrate the advantages of TI over other more conventional methods.

The second contribution of this dissertation is applying TI to empirical questions in computational psychiatry. Initially, we consider TI in the context of dynamical causal modeling (DCM; introduced by (Friston et al., 2003)). In contrast to other models in neuroscience, DCMs are not only used to simulate physiological processes, but also to infer the parameters and models that best explain experimental observations. We restrict our attention to DCM for functional magnetic resonance imaging (fMRI), and show in great detail the theory behind the methods proposed here. We then explore a practical implementation of this methodology, and compare it to other state-of-the-art alternatives.

In addition to DCM, the second application domain explored in this dissertation is models of saccadic eye movements. Research in this type of behavior has a long history in neuroscience and in psychiatry, as eye movements offer a surprisingly rich window into human cognition without the practical hurdles associated with other physiological measurements like electroencephalography (EEG) and fMRI. Moreover, some deficits in eye-movement behavior are likely to constitute endophenotypes or biomarkers of psychiatric and neurological illnesses (Radant et al., 2015; Myles et al., 2017).

The topics explored here are somewhat different but share an underlying agenda that unifies them: to devise and test models of human behavior and physiology that one day could be translated into applications in clinical psychiatry. Because of the differences between these topics, this dissertation is divided into two parts. The first section covers the methodological problem of computing the probability of a model, and the second section focuses on the development and evaluation of models of the oculomotor system.

The rest of the Introduction explains the structure of this dissertation. Each chapter is self-contained and has been redacted as an independent scientific article. Three of them have been published as clearly indicated in the corresponding chapter and the others are either in the final stage of preparation or have been submitted for peer review evaluation. To avoid duplication of efforts but to nevertheless offer a brief contextualization of each chapter, a short introduction precedes each of them. The chapters do not follow the same style in terms of citations and editing, as these have been prepared for different journals. We apologized to the reader in advance for these differences, which were kept to the minimum possible.

Part I: Methodological developments

In this dissertation, we address the problem of model selection from a statistical and probabilistic perspective. Concretely, we restrict our attention to generative models, characterized by formally stating how experimental data is believed to have been generated (Frassle et al., 2018).

In the following, we will refer to experimental data as y, which, depending on the context, might represent data such as reaction times

(RT) or blood-oxygen-dependent (BOLD) time series, for example. From the perspective of generative modelling, y is assumed to be generated by a mapping f from a set of parameters θ. For example, this mapping can represent the physiological process that connects brain activity to the BOLD signal. In addition, experimental data is also assumed to be corrupted by stochastic noise ϵ through a second mapping g. Thus, we informally write

$$g(f(\theta), \epsilon) \rightarrow y$$

meaning that observations y are the result of a generative process characterized by the mappings f and g, parameters θ, and noise ϵ. We further assume that the noise ϵ is generated by the process ω and write

$$\omega \rightarrow \epsilon.$$

This abstract representation encapsulates, among many others, linear regression models, in which the mapping from a set of coefficients $\theta = \beta$ is linear on a set of predictors X, and the noise is assumed to be additive and Gaussian distributed:

$$X\beta + \epsilon \rightarrow y.$$

Beyond the important case of linear regression, some common non-linear models can also be specified as a generative model. For example, in logistic regression, the mapping f is a sigmoid function, such as

$$f(\beta) = \frac{1}{1 + e^{-X\beta}}.$$

In this case, observations $y \in \{0,1\}$ are assumed to be generated from a Bernoulli distribution with mean $f(\beta)$.

The process that is thought to generate parameters θ is also an integral part of a generative model. We refer to this process as π and write:

$$\pi \rightarrow \theta.$$

Thus, it is in principle possible to *simulate* experimental observations by first generating parameters θ and noise ϵ from π and ω. From these, one can generate observations y through the mapping $g(f(\theta), \epsilon)$.

The components of a generative model are usually divided into its structural elements -the mappings f and g, and the description of the process from which parameters and noise are generated, π and ω- and the particular instantiations of the parameters θ and noise ϵ. All these elements together -the model, its parameters, and noise- represent the three forms of uncertainty that arise when studying experimental data:

not only are experimental observations corrupted by measurement noise; we are also uncertain about the adequacy of our models, and the parameters that best describe the data.

The approach used in this dissertation to formalize and quantify this uncertainty is probability theory. Without going into details, probability theory formalizes the notion that certain events are more likely than others, and thereby it can be used to quantify the certainty associated with a particular belief.

Informally, this is done through three simple rules (assuming a set of possible events or outcomes $y_1, \dots, y_N$):

- (*positivity of probabilities*) the probability of an outcome y_i, designated as $p(y_i)$, is a positive number,
- (*normalization*) the sum of the probability of all possible outcomes $\sum_{i=1}^{N} p\,(y_i)$ is 1, and
- (*additivity*) the probability that any of two disjoint outcomes occurs is equal to the sum of the probability of each outcome.

Based on these axioms, probability theory can be used to represent the uncertainty associated with a belief in a single positive number. For example, if the outcome space is understood as the set of models in consideration, we can express our initial confidence about the model m_i as the *prior* probability

$$p(m_i) \in [0,1].$$

Prior probabilities quantify the uncertainty associated with a model before taking into account any novel data or information. Absolute certainty about the correctness of model m_i implies

$$p(m_i) = 1.$$

By contrast, absolute certainty of the inadequacy of the model implies

$$p(m_i) = 0,$$

and all intermediate values represent varying degrees of certainty about a model.

Besides our prior beliefs about the model, we can quantify our prior uncertainty about the parameters θ as the probability

$$p(\theta|m),$$

to be read as 'the conditional probability of θ given m', or more succinctly, 'the probability of θ given m'. This corresponds to our beliefs

about how the parameters θ have been originally generated and are a structural part of the model.

Probability theory prescribes how the uncertainty about the model and its parameters should evolve in the light of experimental observations y. This is determined by the conditional probability of the data given the parameters θ, i.e., the likelihood

$$p(y|\theta, m).$$

This term depends on the model and its parameters, and therefore it is conditional on both of them. For instance, in a linear regression model, the likelihood is Gaussian because the noise ϵ (the difference between the predictions $X\beta$ and observations y), is assumed to be Gaussian.

The rule that dictates optimal belief updates is called Bayes' theorem, in honor of the English priest Thomas Bayes (1701-1761). Bayes' theorem states that our *posterior* belief about parameters θ after taking into account data y should be proportional to the product of the likelihood and prior

$$\underbrace{p(\theta|y,m)}_{posterior} = \frac{1}{Z}\underbrace{p(y|\theta,m)}_{likelihood}\underbrace{p(\theta|m)}_{prior}$$

where Z is a constant with respect to θ. This constant is necessary to warrant the normalization property, i.e.,

$$\sum_j p\left(\theta_j|y,m\right) = \sum_j \frac{p\left(y|\theta_j,m\right)p\left(\theta_j|m\right)}{Z} = 1.$$

It follows that

$$\sum_j p\left(y|\theta_j,m\right)p\left(\theta_j|m\right) = Z.$$

Similarly, the belief update about the model m is also governed by Bayes' rule, such that

$$p(m|y) = \frac{1}{C}p(y|m)p(m),$$

where C is again constant with respect to m and depends on the model space, i.e., the set of models under evaluation.

Fundamentally, probability theory indicates that the likelihood of the data conditioned on the model $p(y|m)$, also called *marginal likelihood*, is given by integrating out the uncertainty about the parameter θ. Hence, the marginal likelihood is precisely:

$$p(y|m) = \sum_j p\,(y|\theta_j, m)p(\theta_j|m) = Z.$$

This term is also known as the *model evidence*, and as we mentioned earlier in this introduction, this term is computationally laborious to calculate exactly. More specifically, to integrate out the uncertainty about θ is a difficult problem in the absence of analytical solutions because models can include thousands or tens of thousands of parameters, and, in the extreme case of non-parametric models, infinitely many. Thus, the ability to update beliefs according to Bayes' rule is often hampered by the challenge of computing the marginal likelihood of a model. Precisely for this reason, a wide variety of methods have been developed to compute or approximate the marginal likelihood of a model, including variational techniques, Markov Chain Monte Carlo (MCMC) methods, and approximations based on asymptotic theory (as explained in detail in Chapters 1 and 3).

The main goal of the first part of this dissertation is to present and to develop methods to assess models based on their marginal likelihood. To do this, we build upon TI, a method originally introduced in 1935 by John Kirkwood (Kirkwood, 1935) to compute the difference in thermodynamic potentials between two systems. Historically, Kirkwood and others were not interested in statistical inference, but in physical chemistry, where the analogous problem of computing the difference in free energy between two systems is paramount. The relationship between statistics and thermodynamics is explained in detail in Chapter 3.

Despite its long history, TI has rarely been used outside the field of physical chemistry, from which it emerged. It is not difficult to identify the reason for this: in spite of the evidence of highly accurate estimates, TI requires large computational resources that have not been available in the past. However, this is becoming a secondary issue as TI can be easily implemented in parallel architectures, fully leveraging multicore CPUs, GPUs, and computer clusters. Hence, part of the problem solved here is the engineering challenge of devising software able to exploit the computational gains offered by modern parallel architectures.

The first chapter of this dissertation is devoted to introducing TI in the context of model comparison, with an emphasis on clinical applications in computational psychiatry. A detailed explanation of model

comparison is presented. In addition, this chapter introduces a novel algorithm to compute predictive likelihoods -a quantity tightly related to the marginal likelihood of a model- for future clinical applications.

In the second chapter, we introduce an implementation of DCM for fMRI in GPUs. This is a tool that leverages parallel architectures either with the goal of model selection or to deal with large data sets. DCMs are an interesting application domain for TI for three key reasons. Firstly, from a translational perspective, DCMs offer mechanistic descriptions of brain connectivity that might be of clinical importance in the future (Stephan et al., 2015; 2017). Moreover, these models are a promising tool to help disentangle patient subpopulations based on their functional connectivity (Brodersen et al., 2011). Finally, as demonstrated in Chapter 3, some of the challenges presented by the differential equations used in DCM might require the type of parallelization gained through GPUs.

The main topic of Chapter 3 is the evaluation of TI in the context of DCM. In this chapter, TI is compared to other sampling-based methods, as well as the standard algorithm for DCM, variational Bayes under the Laplace approximation (VBL) (Friston et al., 2007). Our results demonstrate that there are important advantages in using TI when compared to more common approaches, in terms of less variance in the estimates of the model evidence. This is of great importance for clinical applications that require highly accurate and reliable results.

Part II: Applications to eye movement research

In Part II, we consider an application domain for the technology developed in the first section of this thesis: eye movement research in computational psychiatry. Although a small field, computational modelling in eye movement research with translational applications is promising for several reasons. First, there is strong evidence that behavioral deficits in some oculomotor tasks are endophenotypes of certain psychiatric diseases (Jaafari et al., 2011; Bittencourt et al., 2013; Terao et al., 2013). The most prominent examples are deficits in the antisaccade task in schizophrenia, with at least two large cohort studies providing striking evidence for this in the past 5 years (Reilly et al., 2014; Radant et al., 2015). In addition, the oculomotor system is well understood in the primate brain (Munoz and Everling, 2004), and the

effects of several psychoactive compounds on eye movements have been widely investigated (Reilly et al., 2008). However, only in recent years models of these deficits have appeared (as discussed in Chapter 6), and to date, our understanding of behavioral findings in computational terms is still limited.

The goal of the second part of this dissertation is to develop and evaluate formal computational models of eye movement behavior based on the methods proposed in Part I. The models developed here have been devised with the aim of applying them to translational research in computational psychiatry.

In Chapter 4, we briefly introduce the research agenda followed in Part II and identify some of the most promising questions and methods that could be used to address clinical questions in psychiatry using a computational approach to eye movement research.

In Chapter 5, we investigate the neurocorrelates of saccadic adaptation (SA) with the help of a computational model (McLaughlin, 1967). SA is a form of oculomotor plasticity triggered when the target of a saccade is stepped from its initial location just after the saccade has been initiated. SA is an interesting topic for computational psychiatry because it likely depends on corollary discharge and prediction errors, two types of computation hypothesized to be at the center of the deficits observed in schizophrenia (Adams et al., 2013) and autism (Pellicano and Burr, 2012; Friston et al., 2013; Haker et al., 2016). This chapter is an initial basic science study, part of a larger research program in oculomotor plasticity and prediction-error computation in schizophrenia, as detailed in Chapter 4.

In Chapter 6, we evaluate a series of models of voluntary control of eye movements in the antisaccade task (Hallett, 1978). In this paradigm, subjects are required to saccade in the opposite direction of a suddenly displayed stimulus. As mentioned above, deficits in this paradigm constitute a likely endophenotype of schizophrenia that might shed light on the computational fingerprint of this disease. The main result of this chapter is a novel model, *the Stochastic Early Reaction, Inhibition, and late Action* (SERIA) model, that captures both reaction time distributions and error rates in the antisaccade task, and also has predictive validity for the latency of compensatory saccades that follow an error.

SERIA is further developed in Chapter 6, where we demonstrate that it can be applied to the most common version of the antisaccade task. This chapter showcases the application of the methods developed in Chapter 1 to compute the evidence of hierarchical models. Thus, we can formally evaluate a series of hypothesis about the effect of task design on inhibitory and rule-guided behavior.

In Chapter 8, we study task switching costs in the context of the antisaccade task. Task-switching costs are defined as the degradation in performance that can occur when subjects are required to alternate between tasks that impose different demands (Allport et al., 1994), and have been shown to play a role in a range of neurological conditions. For example, Parkinson's disease patients exhibit difficulties in task switching in a number of paradigms (e.g., Woodward et al., 2002). To study this phenomenon, we use the mixed antisaccade task in Chapters 5 and 6 in consort with the SERIA model. Our results demonstrate robust switching costs in voluntary action generation, as well as in inhibitory control in the antisaccade task.

In Chapter 9 of this dissertation, SERIA is used to answer the following simple empirical question: what are the effects of a pro-dopaminergic (levodopa) and a pro-cholinergic (galantamine) drug in the voluntary control of eye movements in humans? This chapter describes the result of two double-blind, placebo controlled, within-subject experiments. Using the model developed in the previous chapters, we aim to disentangle the effects of the two different drugs. Our findings suggest that these two compounds have opposite effects in this task, with levodopa increasing and galantamine decreasing the RT of voluntary actions. This last chapter brings this dissertation closer to the goal of using formal mathematical models to address truly translational questions in psychiatry.

We conclude with an outlook on future applications of the methods, models, and empirical results obtained in this dissertation.

Because scientific research is not the enterprise of single individuals, all the work presented here has been done collaboratively and thus it is the result of the contributions of different individuals, acknowledged in each of the sections. In particular, Jakob Heinzle, Dario Schoebi, Dominic G. Tschan, Saee Paliwal, and Klaas E. Stephan have been instrumental to the present work.

Part I

Chapter 1

In the Introduction, we presented computational psychiatry (Montague et al., 2012; Stephan and Mathys, 2014; Wang and Krystal, 2014) as the discipline devoted to formulate mathematical models of mental illness, either in terms of behavior, cognition, physiology, or a combination of these dimensions. This goal invites an important epistemological question: How to choose between different, competing models of the same data? While according to falsificationism, scientific theories should be tested by the correctness of their predictions, the view explored here differs in asserting that model selection consists of comparing a set of plausible models and then opting for the one that better accounts for experimental data. What it means for a model to better account for data does not have a unique answer, and in this chapter, we discuss two different options: in-sample explanatory power, and out of sample predictive power. In the context of Bayesian statistics, the first option can be equated to using the marginal likelihood or evidence of a model as selection criterion, whereas the latter corresponds to using the predictive marginal likelihood.

The problem of model selection takes a special twist when the question is not which model accounts best for the totality of the data in consideration, but rather, one asks which model explains best a single observation. This is relevant for future clinical applications in computational psychiatry, in which different computational models might stand for different causal explanations of observable symptoms. For example, it is conceivable that disorders like schizophrenia, which historically emerged as the aggregate of different psychiatric constructs (Jablensky, 2010), might have more than a single biological origin. Thus, in some patients, symptoms might be caused by hypertonic striatal dopamine (Deserno et al., 2016), while in others, similar symptoms might be due to altered NMDA receptor activity in the prefrontal cortex (Krystal et al., 2003). These two pathways can potentially be formalized

as different generative models that relate biological pathomechanisms with physiological and behavioral data. Hence, in a clinical setting the question is not necessarily which is the model that best accounts for the whole population, but rather, which is the best model for a single patient.

Here, we argue that because the evidence of a model evaluated on single observations strongly depends on subjective priors, this method is not satisfactory in clinical applications. However, the evidence of a hierarchical model is far less susceptible to this problem. A definition of this type of models is provided in the chapter. Here, it is enough to characterize them as models in which observations are assumed to come from a common, although unknown population.

Our main conclusion is that while the model evidence is a robust tool when comparing hierarchical models, it is less so for non-hierarchical models. However, the model evidence is a global score that does not directly relate to a single observation and cannot be directly used to evaluate models on a subject-by-subject basis within a hierarchical model. This restriction does not hold for the Bayesian predictive likelihood, which can be derived from the model evidence, and is a more satisfactory score to evaluate models on single observations.

This chapter includes a technical contribution regarding the computation of the predictive likelihood using TI: We device a novel estimator of this quantity based on estimates of the model evidence. This estimator is compared to similar approaches and is shown to offer an accurate estimate at a little *additional* computational cost.

The posterior predictive marginal likelihood for single subject model selection

Eduardo A. Aponte[1,*], *Klaas E. Stephan*[1,2,3], *and Jakob Heinzle*[1]

[1]Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich. Wilfriedstrasse 6, 8004, Zurich, Switzerland.
[2]Wellcome Trust Centre for Neuroimaging, University College London. 12 Queen Square London WC1N 3BG,.
[3]Max Planck Institute for Metabolism Research. Gleueler Strasse 50, 50931, Cologne, Germany.
[*]Corresponding author: Eduardo A. Aponte, aponte@biomed.ee.ethz.ch

ABSTRACT

The marginal likelihood or evidence of a model is often used as criterion to evaluate models in psychology, cognitive neuroscience, and, more recently, in computational psychiatry. Based on this framework, we have recently proposed that the model evidence can be used not only for model selection, but that it can also be used for differential diagnosis in translational applications in computational psychiatry. Thus, instead of aiming for the model that best accounts for a population of subjects, in model-based differential diagnosis, one aims to find the diagnosis -represented by a model- that best explains each observation. Unfortunately, the model evidence strongly depends on the specification of the prior, a situation that is undesirable when the effects of hyperparameters are hard to predict or understand. One solution to this limitation is to rely on hierarchical models, in which the marginal likelihood depends only weakly on the specification of the hyperparameters. Because in this case the model evidence is only a global statistic, the predictive model evidence can be used to evaluate models on a subject-by-subject basis. To make this feasible in a practical setup, we present a novel estimator of the predictive likelihood based on thermodynamic integration. This estimator can be used to compute simultaneously the evidence of a hierarchical model as well as subject-by-subject predictive likelihoods at a minimal additional cost.

Keywords: *model evidence, Bayesian model selection, predictive likelihood, thermodynamic integration, differential diagnosis*

INTRODUCTION

The development of complex computational models in psychology, cognitive science and computational psychiatry has lead to a broadened interest in principled methods to evaluate and compare them. One approach that has become popular in these fields (e.g. Mulder and Wagenmakers 2016; Dienes 2016; Penny 2012) is to use the Bayesian log model evidence (LME). This is defined as the logarithm of the marginal likelihood of data y conditioned on a model m, with parameters θ, and hyperparameters w. Assuming that all the distributions discussed here admit a density with respect to a reference measure (which we assume to be the Lebesgue measure), the model evidence is defined as:

$$p(y|w,m) = \int p(y|\theta,w,m)p(\theta|w,m)d\theta, \tag{1}$$

where $p(y|\theta,w,m)$ is the likelihood function of data y, and $p(\theta|w,m)$ is the prior probability of the parameters θ. In a common setting, the data set y is assumed to be composed of N identically and independently distributed (i.i.d.) observations $\{y_1,\ldots,y_N\} = y$, that represent, e.g., experimental observations from N subjects. Under this condition, the marginal log likelihood is given by

$$\begin{aligned}
\ln p(y|w,m) &= \ln p(y_1,\ldots,y_N|w,m), &(2)\\
&= \sum_{i=1}^{N} \ln p(y_i|w,m), &(3)\\
&= \sum_{i=1}^{N} \ln \int p(y_i|\theta_i,w,m)p(\theta_i|w,m)d\theta_i. &(4)
\end{aligned}$$

In order to compare two models m_0 and m_1, the model evidence can be used, for example, in a likelihood ratio test

$$\frac{p(y|m_1)}{p(y|m_0)} \tag{5}$$

or, given a prior probability on the model space $p(m_i)$, to compute the posterior probability:

$$p(m_1|y) = \frac{p(y|m_1)p(m_1)}{p(y|m_0)p(m_0) + p(y|m_1)p(m_1)}. \tag{6}$$

The likelihood ratio of two models, or Bayes factor, is a widely used method for model selection (Kass and Raftery, 1995).

Despite its conceptual simplicity, the LME requires the marginalization of any free parameter θ, with the caveat that in most but the simplest models no analytical expression is available. For this reason, a large numbe of method to approximate or estimate the LME either through sampling (for example, Meng and Wong 1996; Gelman and Meng 1998; Raftery et al. 2007), approximate Bayesian inference (Friston et al., 2007; Minka, 2001), or asymptotic theory (Schwarz, 1978; Watanabe, 2013) has been developed.

In the case of hierarchical models, in which observations are assumed to be conditionally independent given a set of unknown parameters, the LME is not usually used to compare models. Instead, a more common method to evaluate hierarchical models is cross-validation, in which the goal is to estimate the predictive power of a model. This is done by fitting the model to a set of observations $y = (y_1, \ldots, y_N)$ and then by evaluating it on a test set y_{test}. Here, we will limit our discussion to leave-one-out cross validation, in which the test data set consists of a single observation y_0. Excellent presentations of this topic from Bayesian and non-Bayesian perspectives can be found in Hastie et al. (2001) Chap. 7, Robert (2007) Chap. 7, Vehtari and Ojanen (2012), Burnham and Anderson (2003).

The predictive error of a model, *Err*, is defined as the expected negative utility or loss $-u(y_0, y, m)$, with respect to the data generation process g from which the training and test data sets are sampled

$$Err = -\int g(y) \left[\int g(y_0) u(y_0, y, m) dy_0 \right] dy, \tag{7}$$

$$= -\,\mathrm{E}\left[\mathrm{E}\left[u(y_0, y, m) \right]_{g(y_0)} \right]_{g(y)}. \tag{8}$$

For example, in a typical machine learning application, a classifier is trained on labeled data $y_i = (x_i, l_i)$, where x_i are predictors or independent variables, and l_i are categorical labels. The classifier is evaluated by examining the predicted label $l_{predict}$ based on an unseen predictor x_0. In this context, it is common to use the 0–1 loss function, which assigns 1 if the correct label is predicted, and 0 otherwise.

A second common utility function is the log posterior predictive likelihood (LPL)

$$u(y_0, y, m) = \ln p(y_0 | y, w, m), \tag{9}$$

$$= \ln \int p(y_0 | \theta, w, m) p(\theta | y, w, m) d\theta. \tag{10}$$

This represents the marginal log likelihood of the test observation y_0, after the initial prior $p(\theta|w, m)$ is replaced by the posterior distribution of the parameters $p(\theta|y, w, m)$ conditioned on the training data y. This term is sometimes called Bayes generalization error (Watanabe, 2009), and it is commonly used as a Bayesian model selection criterion (Vehtari and Ojanen, 2012).

Recently, we proposed that the LME can be used as a method for hypothesis testing on single subjects in the context of clinical applications (Heinzle et al., 2016; Stephan et al., 2017). This differs epistemologically from the problem of model selection and development (Gelman and Shalizi, 2013), in which the goal is to device models that offer general and parsimonious explanations of current data, and that predict future observations. Rather, in Heinzle et al. (2016); Stephan et al. (2017), we proposed to use model comparison on a subject-by-subject basis to evaluate plausible hypotheses of the causes or pathomechanisms behind behavioral and physiological symptoms. These hypotheses are represented by generative models that describe the formation of observable symptoms from a set of unobservable causes. Thus, in this scenario, one is not interested in developing a single model that generalizes to as many future observations as possible, but rather to test a series of hypotheses that could better account for individual symptoms and, eventually, guide treatment.

Here, we argue that the predictive likelihood in the context of hierarchical models is a more appropriate method for model selection when one is interested in clinical applications. In particular, we demonstrate in an example from our own work that using the model evidence in non-hierarchical models can lead to 'prior driven' conclusions, a well documented phenomenon sometimes referred to as the Jeffreys-Lindey paradox (Gelman and Shalizi, 2013). This problem is obiquitous in the complex nonlinear models used in computational psychiatry, because it is often hard to predict the role of an individual parameter in a model. Thus, in many situations there is no subjective method to select a satisfactory prior parametrization for a model. Importantly, we show here that this phenomenon is less severe in hierarchical models, but, because the LME provides only a single global score, we suggest to use the LPL when comparing models on a subject-by-subject basis. As the sample size increases, this corresponds

to the model evidence conditioned on the population distribution and thus, it is conceptually similar to the model evidence as a comparison criterion.

To provide estimates of the predictive likelihood, we derive three simple estimators based on thermodynamic integration (TI), a sampling method. Although TI has to our knowledge rarely been used in neuroscience or psychology (but see Chapter 5 to 8), it is a particularly promising approach in biological science (Ballnus et al., 2017) for reason that we have discussed elsewhere (see Chap. 3). One of the estimators proposed here has, to our knowledge, not been described previously in the literature. Because the theory underlying TI is conceptually rich, we discuss it in some detail, emphasizing its connections with other more commonly used techniques.

This paper is organized as follows. First, we describe in more detail both the LME and LPL as methods for model selection. We do not aim to review the very large body of literature on this topic (see Vehtari and Ojanen 2012; Gelman et al. 2014; Han and Carlin 2001). Rather, our goal is to motivate both methods and to explain why the evidence of a non-hierarchical model depends strongly on the specification of its prior. We then proceed to introduce TI and derive an estimator of the LME using this approach, from which we derive three estimators of the leave-one-out predictive likelihood. The first two estimators are based on samples obtained when computing the LME. The third estimator is based on a more computationally intensive approach, designed to generate highly accurate estimates. Finally, we briefly present the cognitive model evaluated here and apply TI to a toy example and to an empirical data set published before.

Our main result shows that if TI is used to compute the LME, accurate estimates of the predictive likelihood can be obtained at a minimal marginal cost. Moreover, the LME and the predictive likelihood yield similar conclusion for sample sizes in the range of common psychological experiments. However, the predictive likelihood offers a formally sound method to evaluate models on a subject-by-subject basis, which is required in clinical settings. Please note that we did not aim to evaluate the computational efficiency of the algorithms developed here, as this falls outside of the scope of this paper.

THE LOG MODEL EVIDENCE

As explained above, one motivation to use differences in LME as model selection criterion is these have an intuitive but formal interpretation in terms of a likelihood ratio test. A second motivation for using the model evidence for model selection is the observation that it includes a penalization term that accounts for the complexity of the model (MacKay, 2002; Stephan et al., 2009; Vandekerckhove et al., 2015). In the following we will refer to the LME with the letter F, in analogy to the concept of negative free energy, as it is the counterpart of this term in thermal physics. In the variational literature, F is often expressed as the difference of an accuracy term A and a penalization or complexity term S. To avoid clutter in the notation, we assume the conditioning on hyperparameters w and model m and make it implicit whenever possible. The LME can be expressed as:

$$
\begin{aligned}
F &= \ln p(y), && (11)\\
&= \int p(\theta|y) \ln \frac{p(y|\theta)p(\theta)}{p(\theta|y)} d\theta, && (12)\\
&= \underbrace{\int p(\theta|y) \ln p(y|\theta) d\theta}_{\text{Accuracy}} - \underbrace{\int p(\theta|y) \ln \frac{p(\theta|y)}{p(\theta)} d\theta}_{\text{Complexity}}, && (13)\\
&= A - S && (14)
\end{aligned}
$$

where $S \geq 0$.

The term A, which in statistical physics is denominated the internal energy U, corresponds to the expected log likelihood of the data under the posterior distribution. The complexity term S is called the Kullback-Leibler (KL) divergence between the posterior and prior distributions. Because S can be shown to be always positive or equal to zero, the model evidence includes a term that penalizes the expected fit of a model according to the divergence between the posterior and prior distributions of the parameters θ. It is commonly argued that this term prevents overfitting, by penalizing overparametrized models (for example, see MacKay 2002; Penny 2012).

When observations are assumed to be independent, the average LME is

equal to

$$\frac{F_N}{N} = \frac{1}{N}\sum_{i=1}^{N} \ln p(y_i), \tag{15}$$

$$= \frac{1}{N}\sum_{i=1}^{N} A_i - S_i. \tag{16}$$

As the number of independent observations N goes to infinity, F_N/N converges to the expected accuracy minus complexity of the model (as long as these exist), where the expectation is taken over the data generating process g:

$$\frac{1}{N}\sum_{i=1}^{N} A_i - S_i \to \int g(y_0)\,(A_0 - S_0)\,dy_0 \tag{17}$$

Because the observations are assumed to be independent, the complexity term scales with the number of observations and thereby, regardless of the sample size, the prior distribution can have a large impact on the LME.

If observations are assumed to come from an unknown but common distribution, as in a hierarchical model, and assuming that the posterior density of θ converges in distribution to $\hat{p}(\theta)$ as the number of observations increases (for which conditions are provided in, for example, Ghosh et al. 1994), the averaged model evidence behaves differently. In this case, the log model evidence can be written as

$$\ln p(y) = \sum_{i=1}^{N}\left[\int p(\theta|y)\ln p(y_i|\theta)d\theta - \int p(\theta|y)\ln\frac{p(\theta|y)}{p(\theta)}d\theta\right]. \tag{18}$$

Thus, we can write

$$\frac{F_N}{N} = \frac{1}{N}\sum_{i=1}^{N} A_i + \frac{S_N}{N}. \tag{19}$$

If the complexity S_N is bounded when $N \to \infty$, the average model evidence F_N/N converges to

$$\mathrm{E}\left[A\right]_g = \mathrm{E}\left[\mathrm{E}\left[\ln p(y_0|\theta)\right]_{\hat{p}(\theta)}\right]_{g(y_0)} \tag{20}$$

where g is the true data generating process. Thus, in the case of non-hierarchical models, the complexity term scales with the number of observations, while in the case of hierarchical models, the complexity term does not grow with the number of observations (if the posterior density converges to a density $\hat{p}(\theta)$). Thus, as the sample size increases, the complexity term tends to be diluted.

CROSS-VALIDATION AND THE PREDICTIVE LIKELIHOOD

As mentioned above, the goal of cross-validation is to estimate the expected negative utility or loss of a model. The expected negative utility is defined as:

$$-\mathrm{E}\left[\mathrm{E}\left[u(y_0, y, m)\right]\right] = -\int \left[\int u(y_0, y, m) g(y_0) dy_0\right] g(y) dy. \tag{21}$$

When the utility function is the LPL, there are two main motivations for using cross validation. First, cross validation controls for overfitting, which occurs when a model displays high in-sample likelihood on the training dataset, but performs poorly on out-of-sample observations. In other words, overfitted models have poor predictive power, but perform well on the data used for training. Hence, by definition, the model with the lowest expected loss has the highest out-of-sample predictive power.

The second motivation for using the LPL is that the model that minimizes the conditional expected loss *Err* given training data y, also minimizes the KL divergence between the data generating process $g(y_0)$ and the predictive likelihood $p(y_0|y)$ (cf. Burnham and Anderson (2003), Chap. 7):

$$\mathrm{KL}\left[g||p\right] = \int g(y_0) \ln \frac{g(y_0)}{p(y_0|y)} dy_0. \tag{22}$$

This term can be expressed as the negative entropy of g, $-H_g$ and the expected loss of the predictive distribution

$$-\mathrm{E}\left[\ln p(y_0|y)\right]_{g(y_0)}. \tag{23}$$

Thus, the conditional error can be expressed as

$$Err = H_g + \mathrm{KL}\left[g||p\right]. \tag{24}$$

Since H_g is constant with respect to the model being evaluated, the model that minimizes the expected loss minimizes the KL divergence between the data generating processes g and predictive posterior distribution $p(y_0|y)$.

From a Bayesian perspective, the LPL corresponds to the log likelihood of a new observation y_0, conditioned on previous observations and the model. In the following we assume $0, \ldots, N$ observations and for

simplicity in notation, we denote the set $\{0, \ldots, N\} \setminus \{i\}$ as $\setminus i$. Also, we denote all the observations $y_0, \ldots, y_N$ as simply y and the set of all observation minus $\{y_i\}$ as $y_{\setminus i}$. The predictive likelihood of observation y_0 is then given by:

$$p(y_0|y_{\setminus 0}) = \int p(y_0|\theta)p(\theta|y_{\setminus 0})d\theta. \tag{25}$$

Thus, this is the marginal likelihood of the new observation under the distribution of θ conditioned on all other observations. The predictive posterior can be rewritten as

$$p(y_0|y_{\setminus 0}) = \int p(y_0|\theta)\frac{p(y_{\setminus 0}|\theta)p(\theta)}{Z_{\setminus 0}}d\theta = \frac{Z}{Z_{\setminus 0}}, \tag{26}$$

where

$$Z = \int \left[\prod_{i=0}^{N} p(y_i|\theta)\right] p(\theta)d\theta, \quad Z_{\setminus 0} = \int \left[\prod_{i=1}^{N} p(y_i|\theta)\right] p(\theta)d\theta. \tag{27}$$

In other words, one can formulate the predictive likelihood as the ratio of the normalization constants of the joint and leave-one-out densities.

The LPL contains a penalization term analogous to the LME:

$$\ln \frac{Z}{Z_{\setminus 0}} = \mathrm{E}\left[\ln p(y_0|\theta)\right]_{p(\theta|y)} - \mathrm{KL}\left[p(\theta|y)||p(\theta|y_{\setminus 0})\right]. \tag{28}$$

Hence, in the LPL the complexity term is the divergence between the leave-one-out and 'full' posterior densities. Because for a large sample size the influence of the prior on the posterior is negligible, its effect on the LPL is often minimal. We note that this is a heuristic argument, and a formal treatment is outside the scope of the present discussion (for a review see Ghosh and Ramamoorthi 2003).

In summary, both the LME and the LPL are commonly used for assessing models. The former includes a penalization term that corresponds to the KL divergence between prior and posterior. The predictive likelihood corresponds to a marginal likelihood in a model in which the prior has been replaced by a leave-one-out density. It includes also a penalization term, which is the divergence between the empirical leave-one-out density and the posterior conditioned on the totality of the data.

In the next section we proceed to show how both the LME and LPL can be estimated using TI.

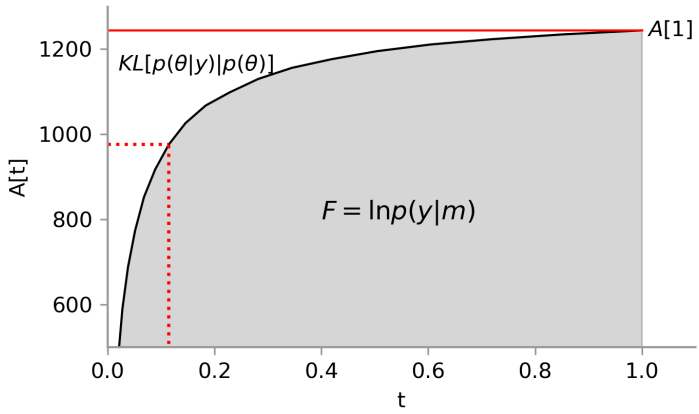


Figure 1: The model evidence corresponds to the area below the curve $A[t] = \mathrm{E}\left[\ln p(y|\theta)\right]_{p(\theta|y,t)}$, which is the accuracy term in the variational formulation of TI. It follows that the area between $A[t]$ and below the constant $A[1]$ is equal to the KL divergence between the prior and posterior up to an arbitrary additive constant. An analoguous relationship holds for every $t \in [0, 1]$.

THERMODYNAMIC INTEGRATION

TI is based on the idea of constructing a piecewise differentiable path γ between the prior and posterior densities. Imposing the conditions

$$F[\gamma(0)] = \ln \int p(\theta) d\theta = 0, \quad F[\gamma(1)] = \ln \int p(y|\theta) p(\theta) d\theta, \tag{29}$$

the main TI equality is:

$$F[\gamma(1)] - F[\gamma(0)] = \int_0^1 \left[\frac{\partial}{\partial t} F\left[\gamma(t)\right] \right] dt = \int_0^1 \left[\frac{dF}{d\gamma} \gamma'(t) \right] dt. \tag{30}$$

Note that the path integral in Eq. 30 is independent of the piecewise differentiable curve γ. Eq. 30 can be seen as the difference in the log normalization constant of two distributions, in this case the prior and unnormalized posterior. More generally, the TI equality can be used to compute changes in LME between distributions, a fact that we will exploit later on.

For non-hierarchical models, typically one assumes $\gamma(t) = t$ and

$$F[t] \stackrel{\text{def}}{=} \ln \int p(y|\theta)^t p(\theta) d\theta. \tag{31}$$

We use now the notation

$$p(\theta|y,t) = \frac{p(y|\theta)^t p(\theta)}{Z_t}, \quad Z_t = \int p(y|\theta)^t p(\theta) d\theta. \tag{32}$$

From this, one obtains that

$$\int_0^1 \frac{\partial}{\partial t} F \Big|_{t=\tau} d\tau = \int_0^1 \mathrm{E}\left[\ln p(y|\theta)\right]_{p(\theta|y,\tau)} d\tau, \tag{33}$$

provided that F and $\partial F/\partial t$ are almost everywhere continuous with respect to t and θ.

This formulation closely maps to the formulation of the LME in terms of accuracy and complexity as depicted in Fig 1. In particular, one can interpret the integral in Eq. 33 as the area below the curve defined by

$$\partial F/\partial t = \mathrm{E}\left[\ln p(y|\theta)\right]_{p(\theta|y,t)} \tag{34}$$

$$\stackrel{\text{def}}{=} A[t] \tag{35}$$

for $t \in [0, 1]$. Note that for $t = 1$, $A[t]$ corresponds to the accuracy of the model.

Eq. 33 can be used to compute the model evidence by estimating the accuracy terms $A[t]$ through sampling and computing the outer integral with the help of a trapezoidal rule or any other numerical integration method.

THERMODYNAMIC INTEGRATION FOR HIERARCHICAL MODELS

The TI method can be easily extended to a model with latent states. Consider a simple model with latent variables θ, and parameters β

$$p(y|\beta)p(\beta|\theta)p(\theta). \tag{36}$$

In this case, the accuracies $A[t]$ required to approximate the LME can be estimated by expressing the likelihood as a marginal. We write the joint probability of the model as

$$p(y|\theta)p(\theta) = \left[\int p(y|\beta)p(\beta|\theta)d\beta\right] p(\theta). \tag{37}$$

The following expression can be used to construct a path between prior and posterior

$$F[t] = \ln \int \left[\int p(y|\beta)^t p(\beta|\theta)d\beta\right] p(\theta)d\theta. \tag{38}$$

This leads to

$$\frac{\partial}{\partial t}F = \frac{1}{Z_t}\int\left[\int \frac{d}{dt}p(y|\beta)^t p(\beta|\theta)d\beta\right] p(\theta)d\theta, \tag{39}$$

$$= \frac{1}{Z_t}\int\left[\int [\ln p(y|\beta)] p(y|\beta)^t p(\beta|\theta)d\beta\right] p(\theta)d\theta, \tag{40}$$

$$= \frac{1}{Z_t}\int\left[\int [\ln p(y|\beta)] \frac{p(y|\beta)^t p(\beta|\theta)}{p(y|\theta,t)}d\beta\right] p(y|\theta,t)p(\theta)d\theta, \tag{41}$$

$$= \mathrm{E}\left[\mathrm{E}\left[\ln p(y|\beta)\right]_{p(\beta|\theta,y)}\right]_{p(\theta|y)}. \tag{42}$$

From the law of total expectation, this integral can be estimated using the expression

$$\frac{1}{K}\sum_{k=1}^{K}\ln p(y|\beta^{(k)}) \tag{43}$$

where samples $\beta^{(k)}$, $k = 1,\ldots,K$, have been obtained from the power posterior

$$\frac{1}{Z_t}p(y|\beta)^t p(\beta|\theta)p(\theta). \tag{44}$$

In the case of hierarchical models, the path between prior and posterior can be constructed similarly. We consider a set of experimental observations y_i, $i \in I = \{1,\ldots,N\}$, which are modeled using the subject specific parameters β_i and the population parameter θ.

$$\ln p(y|t) = \ln p(y_1,\ldots,y_N|t) \tag{45}$$

$$= \ln \int p(\theta)\prod_{i\in I}\left[\int p(y_i|\beta_i)^t p(\beta_i|\theta)d\beta_i\right] d\theta. \tag{46}$$

In Appendix A, it is shown that the estimator

$$\frac{1}{K}\sum_{k=1}^{K}\sum_{i\in I}\ln p(y_i|\beta_i^{(k)}), \tag{47}$$

where $\beta_i^{(k)}$ are samples from the joint power posterior distributions, can be used to compute the accuracies $A[t]$ required in TI.

From the log model evidence to the log predictive likelihood

To our knowledge, TI has been used to compute differences in the normalization constants of two distribution, but not to compute the

LPL in the context of model comparison. Our contribution here is to show how the samples from the power posteriors required to compute the model evidence of a hierarchical model in TI can be used to estimate the leave-one-out LPL using importance sampling. Although a similar method was presented in Vehtari and Lampinen (2002), the method proposed by these authors was not developed in the context of TI.

A BIASED ESTIMATOR FOR THE LOG PREDICTIVE LIKELIHOOD

We assume here that the LME of the totality of the data $y_0, \dots, y_N$ has been computed using the TI method proposed above. Thus, samples of the log likelihood $\ln p(y_i|\beta_i)$ for a grid of temperatures $t_0, \dots, t_M$ are available. Our goal is to compute the ratio of two normalization constants $Z/Z_{\setminus 0}$ using the samples obtained through TI.

We observe that when the number of observations is large enough (Vehtari and Ojanen, 2012):

$$p(\theta|y,t) \approx p(\theta|y_{\setminus 0},t). \tag{48}$$

Thus, one can obtain a biased estimator of the log ratio $\ln Z/Z_{\setminus 0}$ using the TI equality

$$\ln \frac{Z}{Z_{\setminus 0}} = \int_0^1 \mathrm{E}\left[\sum_{i=0}^{N} \ln p(y_i|\theta)\right]_{p(\theta|y,t)} dt - \int_0^1 \mathrm{E}\left[\sum_{i=1}^{N} \ln p(y_i|\theta)\right]_{p(\theta|y_{\setminus 0},t)} dt, \tag{49}$$

$$\approx \int_0^1 \mathrm{E}\left[\sum_{i=0}^{N} \ln p(y_i|\theta)\right]_{p(\theta|y,t)} dt - \int_0^1 \mathrm{E}\left[\sum_{i=1}^{N} \ln p(y_i|\theta)\right]_{p(\theta|y,t)} dt, \tag{50}$$

$$= \int_0^1 \mathrm{E}\left[\ln p(y_0|\theta)\right]_{p(\theta|y,t)} dt. \tag{51}$$

This estimator requires no additional computational cost when the LME is estimated using TI. Note that comparing Eq. 28 and 51, we can conclude that in this approximation the penalization term associated with the LPL is effectively set to zero, while still accounting for the uncertainty associated with the parameters β_i. The severity of this bias depends on the assumption in Eq. 48 (Vehtari and Ojanen, 2012). This estimator is interesting because it demonstrates the relationship between the LME of a hierarchical model and the LPL form the TI perspective.

AN UNBIASED ESTIMATOR

Although the estimator in Eq. 51 is consistent but weakly biased for a large sample size N, it is possible to correct the bias using the joint power posteriors as importance distributions, similarly as in Vehtari and Lampinen (2002). To our knowledge this estimator has not been proposed before. The goal is to compute $Z_{\backslash 0}$ using samples obtained from the posterior distribution conditioned on the totality of the data. This can be achieved using the same TI method as before and then adjusting the estimator with the correct importance weights.

The log marginal $Z_{\backslash 0}$ can be computed using the TI equality:

$$\ln Z_{\backslash 0} = \int_0^1 \mathrm{E}\left[\sum_{i=1}^{N} \ln p(y_i|t,\theta)\right]_{p(\theta|y_{\backslash 0},t)} dt \tag{52}$$

In order to use the TI identity, it is necessary to estimate the accuracies:

$$A_{\backslash 0}[t] \stackrel{\text{def}}{=} \mathrm{E}\left[\sum_{i=1}^{N} \ln p(y_i|\theta,t)\right]_{p(\theta|y_{\backslash 0},t)} \tag{53}$$

$$= \int p(\theta|t,y_{\backslash 0}) \left[\sum_{i=1}^{N} \int p(\beta_i|\theta,y_i,t) \ln p(y_i|\beta_i) d\beta_i\right] d\theta. \tag{54}$$

To simplify the notation, we define

$$\phi(\theta) \stackrel{\text{def}}{=} \sum_{i=1}^{N} \int p(\beta_i|\theta,y_i,t) \ln p(y_i|\beta_i) d\beta_i \tag{55}$$

such that

$$A_{\backslash 0}[t] = \int p(\theta|t,y_{\backslash 0}) \phi(\theta) d\theta. \tag{56}$$

At this point, one can use importance sampling, such that:

$$\int p(\theta|t,y_{\backslash 0}) \phi(\theta) d\theta = \int p(\theta|t,y_{\backslash 0}) \frac{p(y_0|t,\theta)}{p(y_0|t,\theta)} \phi(\theta) d\theta, \tag{57}$$

$$\propto \int p(\theta|t,y) \frac{\phi(\theta)}{p(y_0|t,\theta)} d\theta, \tag{58}$$

$$= \int p(\theta|t,y) \left[\int p(\beta_0|y_0,\theta,t) \frac{\phi(\theta)}{p(y_0|\beta_0)^t} d\beta_0\right] d\theta. \tag{59}$$

The last step corresponds to the same equality underlying the harmonic mean estimator. The importance density is proportional to the posterior

density of θ conditioned on y up to a constant. The likelihood ratio required by the importance distribution is

$$w(\theta, \beta_0) = 1/p(y_0|\beta_0)^t. \tag{60}$$

Using the law of total expectation and the method of self-normalizing importance sampling, the importance sampling estimator of the accuracies is

$$\frac{\sum_{k=1}^{K}\left[w(\theta^{(k)}, \beta_0^{(k)})\sum_{i=1}^{n}\ln p(y_i|\beta_i^{(k)})\right]}{\sum_{k=1}^{K} w(\theta^{(k)}, \beta_0^{(k)})}. \tag{61}$$

Thus, it is possible to use the posterior of the totality of the data as an importance distribution to estimate the leave-one-out normalization constant $Z_{\setminus 0}$. In that case, the TI integral is computed over the adjusted expected values in Eq. 61. Since we assumed that the LME of the totality of the data $\ln Z$ is known, the LPL is readily available. We refer to this estimator as the unbiased TI estimator.

The log ratio TI method

We conclude with a TI estimator that does not use the samples from the joint distribution. For this, one can exploit the more general fact that TI can be used to compute *differences in log marginal likelihood* to formulate another estimator of the LPL. Using Eq. 26 we note that this can be formulated as

$$\ln \frac{Z}{Z_{\setminus 0}} = \int_0^1 \frac{\partial}{\partial t} F[t] d\tau, \tag{62}$$

$$F[t] \stackrel{\text{def}}{=} \ln \int \left[\int p(y_0|\beta_0)^t p(\beta_0|\theta) d\beta_0\right] p(y_{\setminus 0}|\theta) p(\theta) d\theta. \tag{63}$$

In this case, it is not the joint likelihood of the whole model that is relaxed by the parameter t, but only the likelihood of the test set y_0. This method does not require to explicitly compute the LME of the totality of the data $\ln Z$. The disadvantage of this approach is that it requires samples from the power posteriors $Z_t^{-1}\left[\int p(y_0|\beta_0)^t p(\beta_0|\theta) d\theta_0\right] p(y_{\setminus 0}|\theta) p(\theta)$ for each observation y_i, increasing the computational costs by the number of observations. Moreover, when the posterior landscape is challenging it is desirable to sample from an array of temperatures to implement population Markov chain Monte Carlo (MCMC) sampling (Calderhead and

Girolami, 2009; Ballnus et al., 2017). However, one would expect that this is a highly accurate estimator and we will use it as gold standard when comparing the biased and corrected estimators proposed above. In the following we refer to this estimator as the *log ratio TI method*.

AN ILLUSTRATION: THE GAMMA-GAUSSIAN CASE

To illustrate the performance of the estimators presented here, we consider a simple model for which the LME and the LPL can be computed analytically: A Gaussian likelihood with unknown mean μ and precision λ. To keep the model tractable, the prior of λ is assumed to be the Gamma distribution and the prior distribution of μ is assumed to be Gaussian with precision λ and mean μ_0. Later on, we will extend this model into a simple hierarchical model, as discussed below.

The joint probability of $y = [y_1, \ldots, y_N]$ observations, and parameters μ and λ can be written as

$$p(y_1, \ldots, y_N, \mu, \lambda) = \left[\left(\frac{\lambda}{2\pi}\right)^{N/2} \exp -\frac{\lambda}{2}\sum_{i=1}^{N}(y_i - \mu)^2\right] \left[\left(\frac{\lambda}{2\pi}\right)^{1/2} \exp -\frac{\lambda}{2}(\mu_0 - \mu)^2\right]\left[\frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-\lambda b)\right]. \tag{64}$$

The hyperparameters a and b are the shape and scale parameters of the prior density of λ. To obtain a simple expression for the posterior for arbitrary $t \in [0, 1]$, the model can be reparametrized in terms of the summary statistics and 'tempered' sample size. Hence, the model can be fully expressed in terms of the coordinates $\eta(t) = [\eta_0(t), \eta_1(t), \eta_2(t), \eta_3(t)]$:

$$\eta_0(t) = tn + 1, \tag{65}$$

$$\eta_1(t) = \frac{tn\sum_{i=1}^{n}\frac{y_i}{n} + \mu_0}{tn+1}, \tag{66}$$

$$\eta_2(t) = \frac{2b}{tn+1} + \frac{tn\sum_{i=1}^{n}\frac{y_i^2}{n} + \mu_0^2}{tn+1}, \tag{67}$$

$$\eta_3(t) = a. \tag{68}$$

Under this parametrization, the unnormalized power posteriors can be

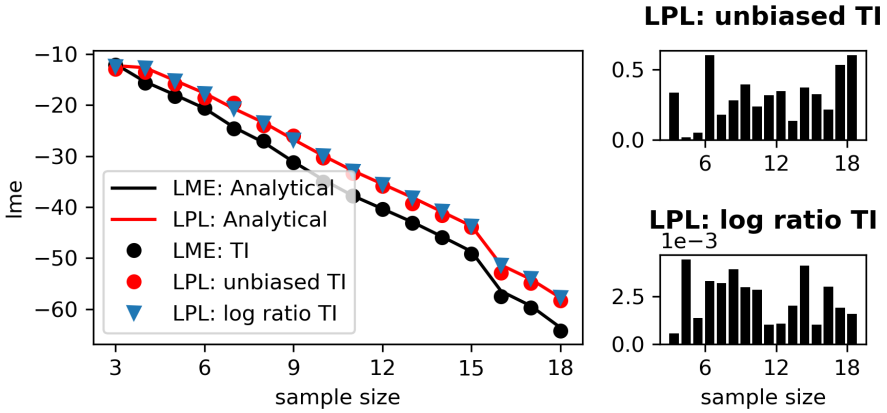


Figure 2: Analytical and estimated LME and LPL. Estimates were computed using 16 chains and 100 samples per chain. **Left:** Analytical and estimated mean LME and LPL for increasingly larger sample sizes with the Gamma-Gaussian model. **Right:** Variance of the two TI estimators of the LPL. As expected, the log ratio TI method displays lower variance than the unbiased TI estimator.

written as:

$$q(\lambda, \mu|\eta(t)) = \lambda^{\eta_3(t)+\eta_0(t)/2-1} \exp -\frac{\lambda\eta_0(t)}{2}(\mu - \eta_1(t))^2 \exp -\frac{\lambda\eta_0(t)}{2}(\eta_2(t) - \eta_1^2(t)). \tag{69}$$

The normalization constant is given by

$$\int q(\lambda, \mu|\eta(t)) d\mu d\lambda = \frac{\left[\frac{\eta_0(t)(\eta_2(t)-\eta_1^2(t))}{2}\right]^{\eta_3(t)+(\eta_0(t)-1)/2} \sqrt{\eta_0(t)}}{\sqrt{2\pi}\Gamma(\eta_3(t) + (\eta_0(t)-1)/2)} \tag{70}$$

Using this parametrization, one can sample from the unnormalized power posteriors by first obtaining samples of λ and then samples from μ. This is explained in detail in Appendix B.

In Fig. 2, we illustrate the LME and the sum of LPL (also called sometimes pseudo-log model evidence Geisser and Eddy 1979; Vehtari and Lampinen 2002), where the hyperparameters were $m_0 = 5, a = 2, b = 2$. The data was generated from a Gaussian distribution centered at -2 and std. 5. All simulations were performed with 16 chains, and 100 samples per chain. To compute the variance of the estimators, the simulations were repeated 10 times. Although both the unbiased TI estimator, as well as the log ratio TI method yielded highly accurate estimates, the latter method displayed lower variance.

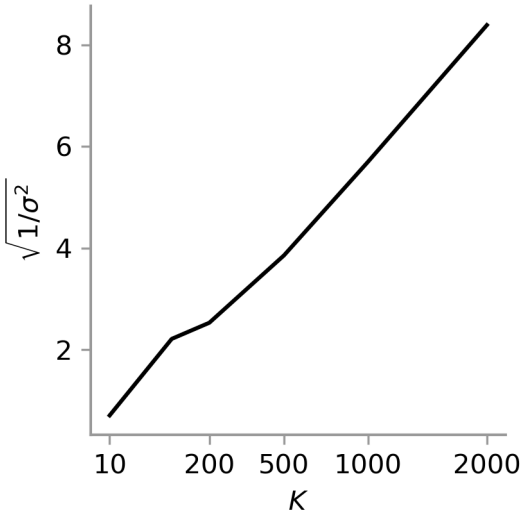


Figure 3: Convergence rate of the unbiased TI estimator. The inverse variance of the ubiased TI estimator as a function of the number of samples drawn for the Gamma-Gaussian model displayed in squared root scale. The sample size is 18 observations. The variance of the estimator is computed by estimating the LPL 50 times for $K = 10, 100, 200, 500, 1000, 2000$.

The key quantity to understand this behavior (Wolpert and Schmidler, 2012) is the tail behavior of the importance weights

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{p(y_0|\theta_0^{(k)})^t} \tag{71}$$

where

$$\theta_0^{(k)} \sim p(\theta_0^{(k)}|y,t) \tag{72}$$

As shown in Wolpert and Schmidler (2012), this estimator is only reliable if the variance of the importance weights exists. Interestingly, in contrast to the examples discussed by Wolpert and Schmidler (2012), in this particular case it can be easily shown that the variance of the importance weights always exists. Thus, the unbiased TI should display $\sqrt{K}$ convergence rate. This is exemplified in 3, which displays the inverse variance of the unbiased TI estimator as a function of the number of samples K.

SUMMARY

Here, we have shown the relationship of the model evidence with the leave-one-out Bayesian predictive likelihood and its relationship with cross validation. In particular, the Bayesian predictive likelihood can be expressed as the ratio of two normalization constants and therefore as a path integral. This path integral can be approximated with the samples used to compute the model evidence of a hierarchical model conditioned on the totality of the data. We proceed to describe the cognitive model and empirical data on which we applied the methods above.

METHODS

THE HIERARCHICAL BAYESIAN FILTER

We consider an example from our own work based on the Hierarchical Bayesian Filter (HGF), a model of learning under environmental uncertainty. The HGF (Mathys et al., 2011) is a probabilistic inference algorithm for an extension of the model proposed in Behrens et al. (2007). This model consists of stacked one dimensional $AR(1)$ processes $X_k(t)$, in which $0 < t < N$ is an index over discrete time, k with $0 < k \leq K$ is an index over the AR processes, and K is the order of the HGF. Usually K is restricted to either $K = 1$ or $K = 2$. Fundamentally, the variance of $X_k(t)$ is assumed to be a function of $X_{k+1}(t)$ such that

$$X_k(t) \sim N(x_k(t-1), \exp(x_{k+1}(t) + \omega_k)), \tag{73}$$

$$X_K(t) \sim N(x_K(t-1), \exp(\omega_K)). \tag{74}$$

For a binary output $U(t)$, the *emissions* of the process are distributed according to

$$p(U(t) = 1 | x_1(t)) = \frac{1}{1 + \exp(-x_1(t))}. \tag{75}$$

The HGF is an approximate single forward pass algorithm that exploits local conditional independencies, similarly as in variational Bayes. Recently, it has received great attention as a cognitive model of

inference under environmental volatility because its update equations can be interpreted as precision weighted prediction error updates. Given that prediction errors are considered a fundamental type of computation performed by the brain (Rao and Ballard, 1999; Friston and Kiebel, 2009), the model proposed by Mathys et al. (2011) has been used in several domains, such as reverse learning (Iglesias et al., 2013), social inference (Diaconescu et al., 2014), gambling (Paliwal et al., 2014), the Posner task (Vossel et al., 2015), etc.. In computational psychiatry, HGF has now been used in, for example, autism (Lawson et al., 2017) and psychosis (Powers et al., 2017) research among others.

For our purposes, we simply consider the HGF as a likelihood function of the form

$$p(y_1, \ldots, y_n | u_1, \ldots, u_n, \theta, m) \tag{76}$$

where y_i correspond to behavioral measurements, u_i to experimental manipulations, m to the specific model evaluated and θ a set of model parameters, which are postulated to reflect idiosyncrasies in the cognition of different individuals. Usually, the likelihood in Eq. 76 can be seen as composed of, first, the maximum a posteriori (MAP) estimates of the states $X_k(t)$ of the HGF

$$f(u_1, \ldots, u_n, \theta) = [x(1), \ldots, x(n)] \tag{77}$$

where $x(t)$ is the vector of MAP estimates $[x_1(t), \ldots, x_K(t)]$ at time t. Second, the MAP estimates are used as arguments of a link likelihood function called observation function

$$p(y_t | x(t), \theta) \tag{78}$$

which describes the probability of a behavioral response given the deterministic putative beliefs predicted by the HGF, given the experimental inputs and subject specific parameters.

In Appendix B, we present a simple hierarchical extension of the HGF based on the Gamma-Gaussian model discussed before. Note that other extensions are possible and the methods develop here depend only on the assumption that the subject specific parameters are conditionally independent given a set of unknown parameters representing the population distribution. In the next section we briefly describe the empirical data set used here, which was previously reported in Iglesias et al. (2013).

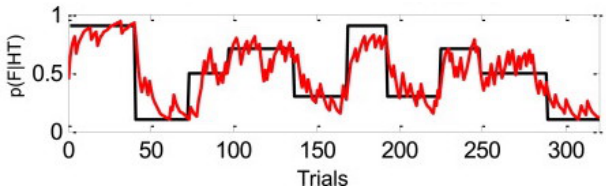


Figure 4: Probability of a face given high tone as a function of trials and a posterior estimate of it according to a second order HGF. Taken from Iglesias et al. (2013).

DATA

We investigated a fraction of the data reported in Iglesias et al. (2013). Briefly, 47 subjects participated in a reverse learning task. Participants were presented an auditory stimulus (AS; a high or low pitch tone) and were asked to guess a visual stimulus (VS; a face or a house) based on the AS. Participants were rewarded for correct guesses. Fundamentally, the conditional densities

$$p(VS|AS) \tag{79}$$

were not constant but changed across trials, as schematically shown in Fig. 4.

Four of the models reported in Iglesias et al. (2013) were evaluated as summarized in Tab. 1. All models included a form of learning in which beliefs about the densities $p(VS|AS)$ were updated across trials. Models m_2 and m_4 used a temporal difference algorithm, but model m_4 included a flexible learning rate according to Sutton (1992). Model m_1 and m_3 were HGFs of order 2 and 1, respectively. This implies that model m_1 included a term representing the volatility of the probability $p(VS|AS)$, whereas in m_4 the volatility was assumed to be constant. All models used a similar observation function according to which the probability of a decision was given by the rule

$$y = 0 \leftrightarrow \text{Response face} \qquad u = 0 \leftrightarrow \text{Low pitch tone} \tag{80}$$

$$y = 1 \leftrightarrow \text{Response house} \qquad u = 1 \leftrightarrow \text{High pitch tone} \tag{81}$$

$$\ln p(y|b,u) = (y \oplus u) \ln \sigma(b) + (1 - y \oplus u) \ln(1 - \sigma(b)), \tag{82}$$

$$\sigma(b) = \frac{1}{1 + \exp(-\beta b)}, \tag{83}$$

where $\oplus$ corresponds to the XOR operator. The parameter β is usually called inverse decision temperature and represents the slope of the

Models in Iglesias et al. (2013)	
m_1	HGF of second order, i.e., with state dependent volatility. Described in detail in Mathys et al. (2011).
m_2	Temporal difference algorithm in which $p(VS\|AS)$ is updated with a fixed learning rate.
m_3	HGF of first order, i.e., no volatility update. Described in detail in Mathys et al. (2011).
m_4	Temporal difference algorithm with variable learning rate as described by Sutton (1992).

Table 1: Summary of the models analyzed from Iglesias et al. (2013).

decision rule. The term b represents a belief about the VS and AS contingency and is computed in a different manner in each of the different models; thus, while the belief update was different in each model, the decision rule was equal across them.

The prior distribution of the parameters was the same as the one used in Iglesias et al. (2013) except that the prior mean of the inverse decision temperature parameter β was assumed to be 1 (as opposed to 48) if not otherwise stated. We were interested in the LME of hierarchical and non-hierarchical models, and the bias and variance of the estimators proposed here.

Implementation

The sampler was implemented in MATLAB and is available as part of the tapas toolbox (https://www.tnu.ethz.ch/en/software/tapas.html). In all the simulations, we used a temperature schedule following a 5th order power rule (Calderhead and Girolami, 2009). The Metropolis-Hastings (MH) step used a Gaussian proposal distribution. During the burn-in phase, the kernel was adapted following Shaby and Well (2011). In addition, we included a population MCMC step in which the parameters were swapped across chains (Calderhead and Girolami, 2009). The number of samples collected is detailed in Tab. 2.

Sampler parameters			
	# chains	# burn-in samples	# kept samples
Iglesias et al. (2013) non hierarchical	16	3000	3000
Iglesias et al. (2013) Hierarchical	32	3000	3000
Iglesias et al. (2013) log ratio TI	32	4000	2000

Table 2: Parameters of the sampler.

RESULTS

In Tab. 3 the LME of each of the models with and without a hierarchical prior is displayed. In the case of the non-hierarchical models, model m_3, which does not include a belief about volatility, was preferred over model m_1. While both models did not display a strong difference in LME ($\Delta LME < 3$) when using a hierarchical prior, hierarchical models obtained a higher LME when compared to non-hierarchical models. Fig. 5 depicts the accuracy as a function of temperature. From Fig. 5, it can be concluded that the higher evidence in favor of the hierarchical models is due to a small improvement in the fit as seen when temperature approaches 1. Moreover, at lower temperatures the accuracy had a higher slope, as the prior mean was quickly replaced by the sample population mean when t increases.

In Tab.4, we report the mean difference and standard deviation across subjects between the biased and unbiased estimator with respect to log

LME			
m_1	m_2	m_3	m_4
Hierarchical			
-6879.87	-7007.46	-6878.47	-7233.31
Non Hierarchical			
-6947.49	-7064.16	-6989.55	-7251.94

Table 3: LME of the models reported in Iglesias et al. (2013).

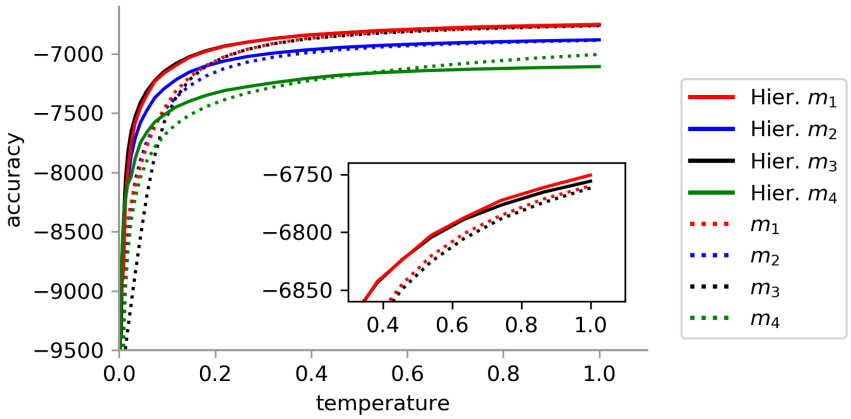


Figure 5: Expected value of the log likelihood as a function of the temperature for the models in Iglesias et al. (2013). Non hierarchical models are represented as broken lines.

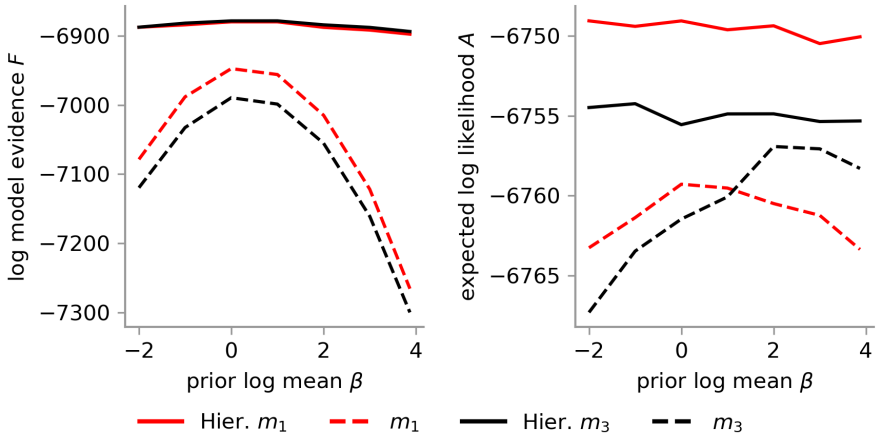


Figure 6: LME and accuracy at $t = 1$. The LME and accuracy at $t = 1$ was computed for different prior means of the log decision temperature $\ln \beta$. **Left:** The LME is highly sensitive to the prior mean in non hierarchical models. This is not the case in hierchical models, in which the LME varies slowly as a function of this hyperparameter. Note that Bayes factors would lead to different conclusions. **Right:** The accuracy as a function of decision temperature β.

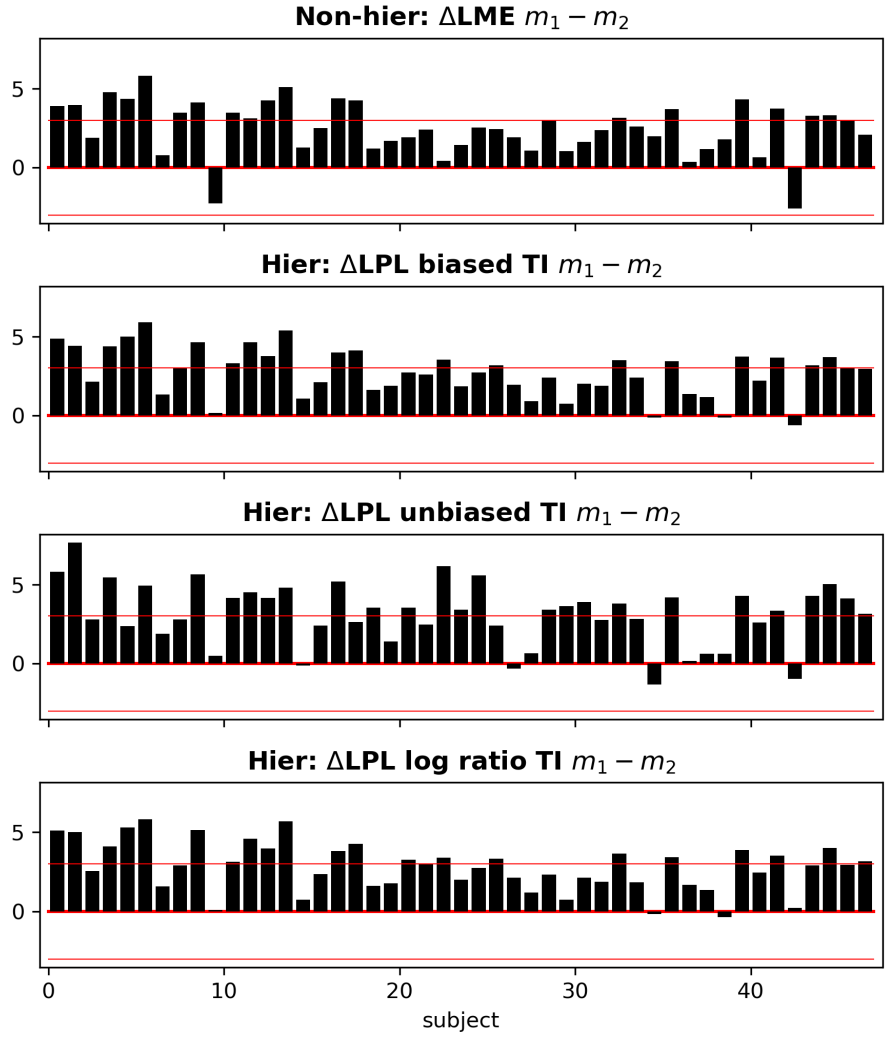


Figure 7: Subject-by-subject model comparison between m_1 and m_2. Red line depics a difference of ± 3, usually considered strong evidence in favor of one of the two models.

ratio TI estimator. Both estimators performed well. While the biased estimator was on average a pessimistic estimator, its variance was low in comparison to the corrected estimator, which showed for all models low bias but high variance. When considered on a subject-by-subject basis (Fig. 7 and 8), while there were no large differences between models m_1 and m_3 except for one subject, there was clearly strong evidence in favor of the HGF for several subjects when compared to a delta learning rule model without a flexible learning rule (model m_3).

Finally, we examined the pseudo LME (Geisser and Eddy, 1979; Vehtari

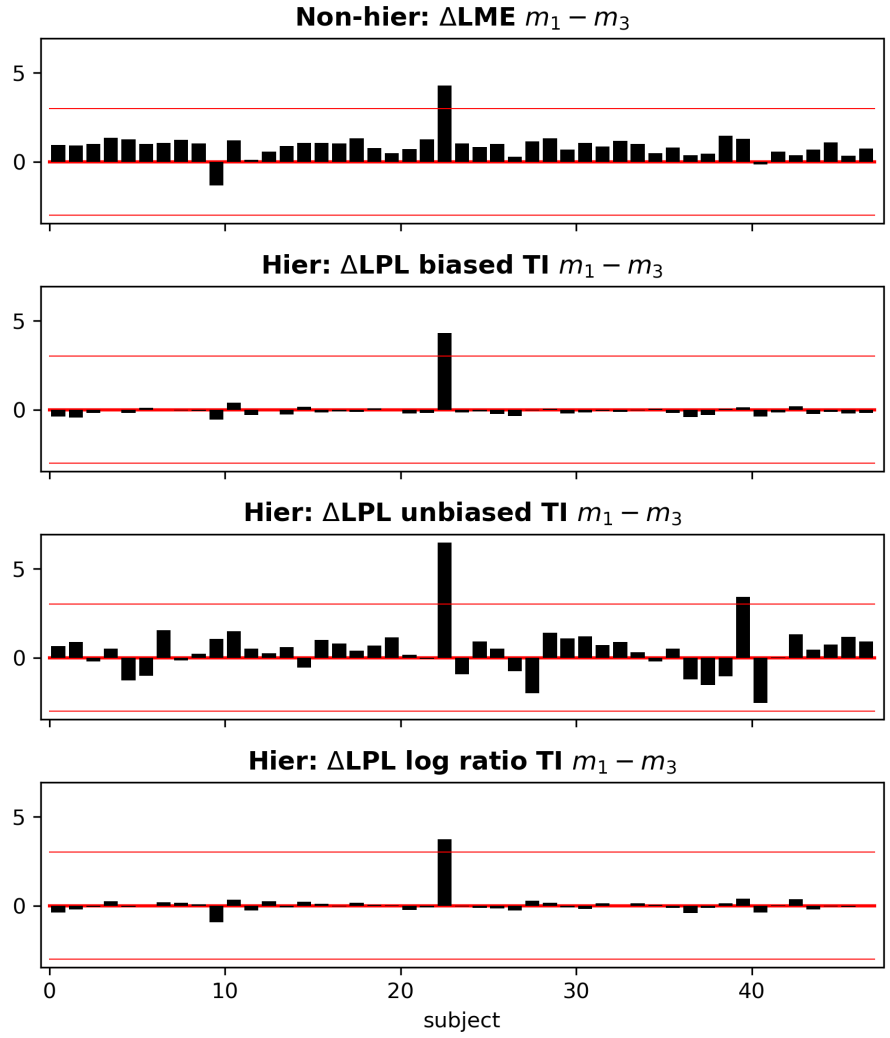


Figure 8: Subject-by-subject model comparison between m_1 and m_3. Although all estimators yield similar conclusions, the unbiased TI estimator displays much higher variance.

Mean diff. estimators							
m_1		m_2		m_3		m_4	
Biased estimator							
-0.410	(0.344)	-0.308	(0.366)	-0.329	(0.320)	-0.188	(0.367)
Corrected estimator							
0.268	(0.875)	-0.031	(0.858)	-0.114	(0.784)	-0.098	(1.000)

Table 4: Mean difference and standard deviation between the biased and unbiased estimators when compared to the log ratio TI estimator for the data reported in Iglesias et al. (2013).

and Lampinen, 2002)

$$\sum_i \ln p(y_i|y_{\setminus i}, m) \tag{84}$$

which has been suggested as a method for model scoring. The rationale of this score is that if for each i, $p(\theta|y_{\setminus i})$ is a good approximation of the population density, then this score corresponds to LME of a model where the prior is the population density. The pseudo LME score is reported for the three estimators. Note that for the biased estimator, this corresponds to the LME estimated through TI. The corrected unbiased and log ratio TI estimators yielded similar values and both favored model m_3 over all other models. However, the unbiased estimator suggested strong evidence in favor of model m_3.

Predictive pseudo LME			
m_1	m_2	m_3	m_4
Biased TI estimator			
-6879.87	-7007.46	-6878.47	-7233.31
Unbiased TI estimator			
-6848.04	-6994.47	-6868.34	-7229.09
log ratios TI estimator			
-6860.61	-6992.99	-6863.00	-7224.49

Table 5: Pseudo LME (Vehtari and Lampinen, 2002) according to the three estimators proposed here for the data reported in Iglesias et al. (2013). In the case of the biased estimator, the pseudo Bayes LME corresponds to the actual LME.

DISCUSSION

Our results reflect the long standing concern that the LME is highly sensitive to the selection of a particular prior density for the model parameters (Kass and Raftery, 1995). This represents a serious issue, as it conflicts with the postulate of scientific objectivity. This concern is aggravated by the difficulty of assessing the effect of prior densities on nonlinear models, in which parameters can interact in subtle manners. Scoring models using the LME of their hierarchical versions partially addresses this concern because prior distributions are diluted in a population dependent prior. In terms of the TI formulation of the LME, this is represented by a higher slope of the accuracy $A[t]$ at lower temperatures. Also, hierarchical models can lead to better fits in terms of the expected log likelihood under the posterior distribution. We note, however, that although the LME can be influenced by the particular priors used, this concern can be addressed with several measures such as sensitivity analyses (see for example the discussion in Kass and Raftery 1995), and thus it does not constitute a general reason to not use this criterion for model selection. However, it is important to highlight that Bayesian model comparison should be accompanied by other quantitative and qualitative tests (Gelman et al., 2003; Gelman and Shalizi, 2013).

An alternative score besides the model evidence is the Bayesian predictive likelihood. This method quantifies the likelihood of a new observation, conditioned on previous observations. Thus, it provides information about the generalizability of a model, as opposed to the likelihood of already observed data. Because this quantity can be understood as the ratio of two normalization constants, it can be easily approximated using samples obtained from computing the LME of a hierarchical model with TI.

Here, we proposed three estimators. The first is a biased approach that simply uses the joint power posteriors as an approximation of the leave-one-out densities. The second estimator corrects this bias by using importance sampling. To account for the presence of a marginal likelihood in the correction factor, we proposed to use the equality underlying the harmonic mean estimator. One of the advantages of this approach is that it addresses the main concern about importance sampling in this context: although the weights

$$\frac{1}{p(y|\theta)^t} \tag{85}$$

can have infinite variance in many applications of the harmonic mean estimator (Raftery et al., 2007), the temperature parameter effectively counteracts the high variance of this term at low temperatures. Moreover, as t approaches 1, the variance of this term decreases. Therefore, in practice the method that we proposed here is less susceptible to the problems usually attributed to the harmonic mean estimator in terms of unacceptable convergence rates (Wolpert and Schmidler, 2012).

Both the biased and unbiased estimator provided fairly accurate approximations of the predictive density at minimal *added* computational costs compared to the direct computation of the log ratio using the TI inequality. Thus, although the latter method should be preferred when the predictive likelihood is paramount, the biased and unbiased estimators computed from samples used to estimate the LME are acceptable approximations.

TRANSLATIONAL APPLICATIONS

In Stephan et al. (2017), we suggested to use differences in LME for diagnosis and prediction of treatment response in computational psychiatry. This requires subject specific scores for each model tested. The methods developed here can translate group level results, in terms of the LME of a hierarchical model, to subject specific scores. In particular, one can use the Bayesian predictive likelihood as a subject specific decision rule for model comparison, by conditioning on group level results. Thus, one can use predictive pseudo Bayes factors of the form

$$\frac{p(y_i|y_{\setminus i}, m_1)}{p(y_i|y_{\setminus i}, m_2)}. \tag{86}$$

This approach has the advantage of integrating previous knowledge when comparing hypothesis regarding a new observation and it should be preferred since the effect of *subjective* priors can yield inadequate conclusions as shown here.

A further advantage of producing subject specific scores is that these can be used in the context of hierarchical models that consider model assignment as a random variable, such as Stephan et al. (2009) and more recently Rigoux et al. (2014). This depends on the assumption that the leave-one-out distribution is a good approximation for the population distribution.

SUMMARY

This paper contains two novel ideas. First we extended the classical TI approach to develop an estimator of hierarchical models. Second, we developed three estimators of the Bayesian predictive likelihood based on the TI equality. We applied it to a multilevel extension of the HGF, a cognitive model of belief updates under environmental volatility. We showed that considering the LME of group models reduces the dependency of the results with respect to hyperparameters. Moreover, we showed that the biased and unbiased estimators proposed here performed comparatively well to subject-wise estimator of the predictive likelihood using the TI equality. Given the reduced computational costs of estimating only once the LME evidence of a hierarchical model, we suggest to use the unbiased estimator.

APPENDIX

A. THERMODYNAMIC INTEGRATION FOR HIERARCHICAL MODELS

Let $y_i, i \in I = \{1, \dots, N\}$ be a set of N observations, and let the power posterior density of the model m at temperature $t \in [0, 1]$ be

$$\frac{1}{Z_t} p(\theta) \prod_{i \in I} p(y_i|\theta, t) \tag{87}$$

where

$$p(y_i|\theta, t) = \int p(y|\beta_i)^t p(\beta_i|\theta) d\beta_i. \tag{88}$$

The model evidence is given by

$$F = \int_0^1 \frac{\partial F}{\partial t} dt \tag{89}$$

$$= \int_0^1 A[t] dt \tag{90}$$

We show that the estimator

$$\frac{1}{K} \sum_{k=1}^{K} \sum_i \ln p(y_i|\beta_i^{(k)}) \to A[t] \tag{91}$$

where $\beta_i^{(j)}$ are samples from the power posterior

$$\left[\prod_{i \in I} p(\beta_i|y_i, \theta, t) \right] p(\theta|y_1, \dots, y_N, t) \tag{92}$$

and $K \to \infty$.

The proof follows from the TI fundamental equation

$$\int_0^1 \frac{\partial}{\partial t} F[t] \Big|_\tau d\tau = \int_0^1 \frac{\partial}{\partial t} \ln \int \left(\prod_{i \in I} p(y_i|t, \theta) \right) p(\theta) d\theta \Big|_\tau d\tau \tag{93}$$

$$= \int_0^1 \frac{1}{Z_\tau} \int \left(\sum_{i \in I} \frac{\partial}{\partial t} p(y_i|t, \theta)|_\tau \prod_{j \in I \setminus \{i\}} p(y_j|\theta, \tau) \right) p(\theta) d\theta d\tau \tag{94}$$

$$= \int_0^1 \int \frac{p(\theta) \left(\prod_{i \in I} p(y_i|\tau, \theta) \right)}{Z_\tau} \sum_{i \in I} \frac{\int [\ln p(y_i|\beta_i)] p(y_i|\beta_i)^\tau p(\beta_i|\theta) d\beta_i}{p(y_i|\tau, \theta)} d\theta d\tau \tag{95}$$

We note that all the densities are well normalized, such that we can write Eq. 95 as a expected value of the form

$$\int_0^1 \frac{\partial}{\partial t} F[t]\Big|_\tau \, d\tau = \int_0^1 \mathrm{E}\left[\sum_{i\in I} \mathrm{E}\left[\ln p(y_i|\beta_i)\right]_{p(\beta_i|y_i,\theta,\tau)}\right]_{p(\theta|y_1,\ldots,y_N,\tau)} d\tau. \tag{96}$$

The statement follows from the law of total expectation and the law of large numbers.

B. A HIERARCHICAL HGF

In the following, we lump all experimental manipulations used with subject s_i, in the variable u_i and the corresponding responses in y_i. The goal is to compute the posterior density of the subject dependent parameters θ_i, conditioned on a model m, manipulations u_i, and responses y_i and the corresponding normalization constant.

When applying the HGF to data it is customary to assume a Gaussian prior density on the subject dependent parameters. Thus, this model can be extended by defining a hierarchical prior that models the sample mean μ. According to the model (Fig.9), observations are generated from this mean with variance λ following:

$$p(y_1,\ldots,y_n,u_1,\ldots,u_n,\mu,\lambda) = \left[\Pi_{i=1}^n p(y_i|u_i,\theta_i)p(\theta_i|\mu,\lambda)\right] p(\mu,\lambda), \tag{97}$$

$$p(\theta_i|\mu,\lambda) = \frac{\lambda^{1/2}}{\sqrt{2\pi}} \exp\left(-\frac{\lambda}{2}(\theta_i-\mu)^2\right), \tag{98}$$

$$p(\mu,\lambda) = \frac{\lambda^{1/2}}{\sqrt{2\pi}} \exp\left(-\frac{\lambda}{2}(\theta_0-\mu)^2\right) \frac{\beta^a \lambda^{a-1}}{\Gamma(a)} \exp(-\beta\lambda). \tag{99}$$

Here Γ denotes the Gamma function. Later, we use the same symbol to denote the Gamma p.d.f.. Which is meant should be clear from the context. The prior density $p(\mu,\lambda)$ corresponds to the product of a normal and a Gamma density. For simplicity in notation, θ_i is assumed here to be a scalar quantity, but the extension to a higher dimensional space is structurally equivalent. The term $p(y_i|u_i,\theta_i)$ constitutes the likelihood defined by the HGF as explained before. This model has the desirable property that the parameters λ and μ are conditionally independent of y when θ_i has been observed. This could be leveraged

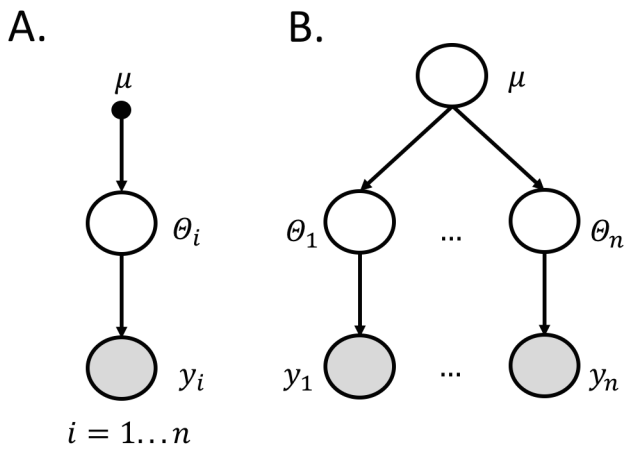


Figure 9: Graphical representation of hierarchical and non-hierarchical models. Observed nodes are depicted in gray. A. In a non-hierarchical model all observations y_i are assumed to be generated from a known population mean μ. Observations are independent of each other. B. In a hierarchical model, observations are assumed to be generated from an unknown population mean μ. Importantly, observations are not any more independent. For simplicity we have not specified a variance term in the generating process $p(\theta_i|\mu)$.

to design an efficient variational algorithm using the Laplace approximation on the conditional probability

$$p(\theta_i|y_i, u_i, \mu, \lambda) \tag{100}$$

as in Mathys et al. (2011); Friston et al. (2007). Instead, the E step typical of variational algorithms can be replaced by a Gibbs sampling step. For this, we consider the conditional posterior density of μ, λ

$$p(\mu, \lambda|\theta_0, \ldots, \theta_n) = N\left(\mu; \sum_{i=0}^{n} \frac{\theta_i}{n+1}, \frac{1}{\lambda(n+1)}\right) \Gamma\left(\lambda; a + n/2, \beta + \frac{(n+1)}{2}\text{Var}[\theta_i]\right) \tag{101}$$

where $\text{Var}[\theta_i]$ is the sample variance of θ_i. Note that the prior θ_0 is simply treated as a further observation with the same weighted contribution as the parameters $\theta_1, \ldots, \theta_n$.

To obtain a sample from the conditional posterior, one first draws a sample of λ. The expected value of the precision is

$$E[\lambda|\theta_0, \ldots, \theta_n] = \frac{a + n/2}{\beta + \text{Var}[\theta_i](n+1)/2}. \tag{102}$$

Its variance is

$$\mathrm{Var}[\lambda|\theta_0,\ldots,\theta_n] = \frac{\alpha + n/2}{(\beta + \mathrm{Var}[\theta_i](n+1)/2)^2}. \tag{103}$$

As n grows, the expected value of λ approaches $1/\mathrm{Var}[\theta_i]$ and its variance approaches zero.

In the second part of the Gibbs step, a sample of μ is drawn from a Gaussian distribution with mean $\sum_{i=0}^{n} \theta_i/(n+1)$ and variance $1/\lambda(n+1)$.

Finally the posterior densities of the parameters of each subject are conditionally independent of each other given samples μ_k and λ_k:

$$p(\theta_i|y_i,u_i,\mu_k,\lambda_k) \propto p(y_i|u_i,\theta_i)p(\theta_i|\mu_k,\lambda_k) = p(y_i|u_i,\theta_i)N(\theta_i;\mu_k,1/\lambda_k). \tag{104}$$

This can be exploited in a local Metropolis Hastings step using the acceptance-rejection probability

$$\min\left\{1, \frac{p(y_i|\theta_i^*)p(\theta_i^*|\mu_k,\lambda_k)}{p(y_i|\theta_i)p(\theta_i|\mu_k,\lambda_k)}\right\}, \tag{105}$$

where θ_i^* corresponds to a sample obtained from a Gaussian kernel centered on θ_i. Note that while the variance $1/(n+1)\lambda$ is used when drawing a sample of the population average, the conditional $p(\theta_i|\mu,\lambda)$ is a Gaussian density with variance $1/\lambda$.

REFERENCES

Ballnus, B., Hug, S., Hatz, K., Gorlitz, L., Hasenauer, J., and Theis, F. J. Comprehensive benchmarking of Markov chain Monte Carlo methods for dynamical systems. *BMC Syst Biol*, 11(1):63, Jun 2017.

Behrens, T. E., Woolrich, M. W., Walton, M. E., and Rushworth, M. F. Learning the value of information in an uncertain world. *Nat. Neurosci.*, 10(9):1214–1221, Sep 2007.

Burnham, K. P. and Anderson, D. R. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2003.

Calderhead, B. and Girolami, M. Estimating bayes factors via thermodynamic integration and population MCMC. *Computational Statistics & Data Analysis*, 53(12): 4028–4045, 2009. doi: 10.1016/j.csda.2009.07.025. URL `http://dx.doi.org/10.1016/j.csda.2009.07.025`.

Diaconescu, A. O., Mathys, C., Weber, L. A., Daunizeau, J., Kasper, L., Lomakina, E. I., Fehr, E., and Stephan, K. E. Inferring on the intentions of others by hierarchical Bayesian learning. *PLoS Comput. Biol.*, 10(9): e1003810, Sep 2014.

Dienes, Z. How bayes factors change scientific practice. *Journal of Mathematical Psychology*, 72:78 – 89, 2016. ISSN 0022-2496. doi: http://dx.doi.org/10.1016/j.jmp.2015.10.003. URL `http://www.sciencedirect.com/science/article/pii/S0022249615000607`. Bayes Factors for Testing Hypotheses in Psychological Research: Practical Relevance and New Developments.

Friston, K. and Kiebel, S. Predictive coding under the free-energy principle. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 364(1521):1211–1221, May 2009.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., and Penny, W. Variational free energy and the Laplace approximation. *Neuroimage*, 34(1):220–234, Jan 2007.

Geisser, S. and Eddy, W. A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160, 1979.

Gelman, A. and Meng, X.-L. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185, 1998.

Gelman, A. and Shalizi, C. R. Philosophy and the practice of Bayesian statistics. *Br J Math Stat Psychol*, 66(1):8–38, Feb 2013.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2003.

Gelman, A., Hwang, J., and Vehtari, A. Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24 (6):997–1016, 2014. ISSN 1573-1375. doi: 10.1007/s11222-013-9416-2. URL `http://dx.doi.org/10.1007/s11222-013-9416-2`.

Ghosh, J. K., Ghosal, S., and Samanta, T. Stability and convergence of the posterior in non-regular problems. *Statistical Decision Theory and Related Topics V*, pages 183–199, 1994.

Ghosh, J. and Ramamoorthi, R. *Bayesian Nonparametrics*. Springer Series in Statistics. Springer, 2003. ISBN 9780387955377. URL `https://books.google.ch/books?id=4jhE7N23hRcC`.

Han, C. and Carlin, B. Markov chain monte carlo methods for computing bayes factors: A comparative review. *Journal of the American Statistical Association*, 96(455):1122–1132, 2001.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Heinzle, J., Aponte, E. A., and Stephan, K. E. Computational models of eye movements and their application to schizophrenia. *Current Opinion in Behavioral Sciences*, 11:21–29, 2016.

Iglesias, S., Mathys, C., Brodersen, K. H., Kasper, L., Piccirelli, M., den Ouden, H. E., and Stephan, K. E. Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. *Neuron*, 80(2): 519–530, Oct 2013.

Kass, R. E. and Raftery, A. E. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, June 1995. ISSN 01621459. doi: 10.2307/2291091. URL `http://dx.doi.org/10.2307/2291091`.

Lawson, R. P., Mathys, C., and Rees, G. Adults with autism overestimate the volatility of the sensory environment. *Nat. Neurosci.*, 20(9):1293–1299, Sep 2017.

MacKay, D. J. C. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, New York, NY, USA, 2002. ISBN 0521642981.

Mathys, C., Daunizeau, J., Friston, K. J., and Stephan, K. E. A bayesian foundation for individual learning under uncertainty. *Front Hum Neurosci*, 5:39, 2011.

Meng, X. and Wong, W. H. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, pages 831–860, 1996.

Minka, T. P. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pages 362–369, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-800-1. URL `http://dl.acm.org/citation.cfm?id=2074022.2074067`.

Mulder, J. and Wagenmakers, E. Editors' introduction to the special issue 'bayes factors for testing hypotheses in psychological research: Practical relevance and new developments'. *Journal of Mathematical Psychology*, 72:1 – 5, 2016. ISSN 0022-2496. doi: http://dx.doi.org/10.1016/j.jmp.2016.01.002.

Paliwal, S., Petzschner, F. H., Schmitz, A. K., Tittgemeyer, M., and Stephan, K. E. A model-based analysis of impulsivity using a slot-machine gambling paradigm. *Front Hum Neurosci*, 8:428, 2014.

Penny, W. D. Comparing dynamic causal models using AIC, BIC and free energy. *Neuroimage*, 59(1):319–330, Jan 2012.

Powers, A. R., Mathys, C., and Corlett, P. R. Pavlovian conditioning-induced hallucinations result from overweighting of perceptual priors. *Science*, 357(6351):596–600, 08 2017.

Raftery, A., Newton, M., Satagopan, J., and Krivitsky, P. Estimating the Integrated Likelihood via Posterior Simulation Using the Harmonic Mean Identity. In *Bayesian Statistics 8*, pages 1–45, 2007.

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

Rigoux, L., Stephan, K. E., Friston, K. J., and Daunizeau, J. Bayesian model selection for group studies - revisited. *Neuroimage*, 84:971–985, Jan 2014.

Robert, C. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

Schwarz, G. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

Shaby, E. and Well, M. Exploring and adaptative Metropolis algorithm. Technical report, Dep. of Statistical Sciences Duke University, 2011.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., and Friston, K. J. Bayesian model selection for group studies. *Neuroimage*, 46(4): 1004–1017, Jul 2009.

Stephan, K. E., Schlagenhauf, F., Huys, Q. J., Raman, S., Aponte, E. A., Brodersen, K. H., Rigoux, L., Moran, R. J., Daunizeau, J., Dolan, R. J., Friston, K. J., and Heinz, A. Computational neuroimaging strategies for single patient predictions. *Neuroimage*, 145(Pt B):180–199, Jan 2017.

Sutton, R. S. Gain adaptation beats least squares. In *Proceedings of the 7th Yale workshop on adaptive and learning systems*, pages 161–166, 1992.

Vandekerckhove, J., Matzke, D., and Wagenmakers, E. Model comparison and the principle of parsimony. In Busemeyer, J., Wang, Z., Townsend, J. T., and Eidels, A., editors, *Oxford library of psychology. The Oxford handbook of computational and mathematical psychology*, pages pp. 300–319. University Press, New York, NY, US, 2015.

Vehtari, A. and Lampinen, J. Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Comput*, 14(10):2439–2468, Oct 2002.

Vehtari, A. and Ojanen, J. A survey of bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6: 142–228, 2012.

Vossel, S., Mathys, C., Stephan, K. E., and Friston, K. J. Cortical Coupling Reflects Bayesian Belief Updating in the Deployment of Spatial Attention. *J. Neurosci.*, 35(33):11532–11542, Aug 2015.

Watanabe, S. *Algebraic geometry and statistical learning theory*, volume 25. Cambridge University Press, 2009.

Watanabe, S. A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, 14(Mar):867–897, 2013.

Wolpert, R. L. and Schmidler, S. C. α-stable limit laws for harmonic mean estimators of marginal likelihoods. *Statistica Sinica*, pages 1233–1251, 2012.

Chapter 2

A prominent approach to model brain activity that has promising applications in computational psychiatry is DCM. The goal of this framework is to device physiological models of experimental measurements (such as eye movements (Adams et al., 2016), galvanic skin conductance (Bach et al., 2010), EEG (Kiebel et al., 2009), or fMRI (Friston et al., 2003)) that can be probabilistically fitted to experimental data. In contrast to most physiological models in neuroscience, DCMs are used both as forward models that generate predictions and simulations, as well as the backbone of statistical inference from experimental data to their putative biological causes. This is done, vaguely speaking, by minimizing the difference between simulations of the forward model and experimental data.

The ambitious program of using physiologically plausible models both as forward and backward models comes at the price of high computational and mathematical complexity. This chapter presents an engineering solution to these two challenges that relies on specialized hardware to increase the number of concurrent simulations that can be generated from a model. This is interesting in the scenario presented in the Chapter 1: the computation of the model evidence using TI in both hierarchical and non-hierarchical models. Thereby, an efficient implementation of DCM that can generate large numbers of simulations in parallel renders the utilization of the methods presented in Chapter 2 feasible. We compare TI to other inference methods in Chapter 3.

This chapter has been published as *Aponte, E. A., Raman, S., Sengupta, B., Penny, W. D., Stephan, K. E., & Heinzle, J. (2016). mpdcm: A toolbox for massively parallel dynamic causal modeling. Journal of neuroscience methods, 257, 7-16.* It is a verbatim copy of the document:

https://doi.org/10.1016/j.jneumeth.2015.09.009.

Journal of Neuroscience Methods 257 (2016) 7–16

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Computational neuroscience

mpdcm: A toolbox for massively parallel dynamic causal modeling

Eduardo A. Aponte[a,*], Sudhir Raman[a], Biswa Sengupta[b], Will D. Penny[b], Klaas E. Stephan[a,b,c], Jakob Heinzle[a]

[a] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and Swiss Federal Institute of Technology (ETH), 8032 Zurich, Switzerland
[b] Wellcome Trust Centre for Neuroimaging, University College London, 12 Queen Square, London WC1N 3BG, UK
[c] Max Planck Institute for Metabolism Research, 50931 Cologne, Germany

HIGHLIGHTS

- "mpdcm" is a toolbox for fast simulation of Dynamic causal models for fMRI on GPUs.
- Parallelization reduces computation time by up to two orders of magnitude.
- This enables the use of sampling algorithms for Bayesian inference.
- The mpdcm toolbox is openly available under the GPLv3 license.

ARTICLE INFO

Article history:
Received 13 May 2015
Received in revised form 15 August 2015
Accepted 8 September 2015
Available online 16 September 2015

Keywords:
Dynamic causal modeling
GPU
Markov chain Monte Carlo
Thermodynamic integration
Parallel tempering
Model inversion
Model evidence
Bayesian model comparison

ABSTRACT

Background: Dynamic causal modeling (DCM) for fMRI is an established method for Bayesian system identification and inference on effective brain connectivity. DCM relies on a biophysical model that links hidden neuronal activity to measurable BOLD signals. Currently, biophysical simulations from DCM constitute a serious computational hindrance. Here, we present *Massively Parallel Dynamic Causal Modeling* (**mpdcm**), a toolbox designed to address this bottleneck.
New method: **mpdcm** delegates the generation of simulations from DCM's biophysical model to graphical processing units (GPUs). Simulations are generated in parallel by implementing a low storage explicit Runge–Kutta's scheme on a GPU architecture. **mpdcm** is publicly available under the GPLv3 license.
Results: We found that **mpdcm** efficiently generates large number of simulations without compromising their accuracy. As applications of **mpdcm**, we suggest two computationally expensive sampling algorithms: thermodynamic integration and parallel tempering.
Comparison with existing method(s): **mpdcm** is up to two orders of magnitude more efficient than the standard implementation in the software package SPM. Parallel tempering increases the mixing properties of the traditional Metropolis–Hastings algorithm at low computational cost given efficient, parallel simulations of a model.
Conclusions: Future applications of DCM will likely require increasingly large computational resources, for example, when the likelihood landscape of a model is multimodal, or when implementing sampling methods for multi-subject analysis. Due to the wide availability of GPUs, algorithmic advances can be readily available in the absence of access to large computer grids, or when there is a lack of expertise to implement algorithms in such grids.

1. Introduction

Dynamic causal modeling (DCM) (Friston et al., 2003) is a widely used Bayesian framework for inference on effective connectivity from neurophysiological data. When applied to fMRI, it is a hierarchical generative model that integrates an explicit model of neuronal population interactions with a biophysical model of

* Corresponding author. Tel.: +41 44 634 91 12; fax: +41 44 634 9131.
E-mail address: aponte@biomed.ee.ethz.ch (E.A. Aponte).

http://dx.doi.org/10.1016/j.jneumeth.2015.09.009
0165-0270/

the regional blood oxygen level dependent (BOLD) signals caused by those interactions. Currently, parameter estimation in DCM rests on a variational Bayesian scheme which maximises negative free energy (Friston et al., 2003, 2007). Alternative inference techniques have been proposed, for example, Markov chain Monte Carlo (MCMC) sampling (Chumbley et al., 2007; Raman et al. unpublished results, see also Sengupta et al., 2015 for an MCMC approach to DCM for electrophysiological data). However, variational Bayes has not yet been replaced as the method of choice for inference in DCM. One reason for this is computational efficiency: sampling methods are limited by the high computational cost of evaluating DCM's likelihood function, because this requires integrating the neuronal as well as the hemodynamic model equations in DCM. (For simplicity, in the following we use the term "simulations" to refer to integration of the model's state equations.)

In this paper, we introduce *Massively Parallel Dynamic Causal Modeling*, **mpdcm**, a toolbox designed to overcome some of the computational limitations of DCM. **mpdcm** delegates the simulations of DCM's biophysical model to graphical processing units (GPUs). The goal of our toolbox is to facilitate the implementation of algorithms for statistical inference in DCM that require large numbers of simulations, as in the case of multi-subject hierarchical models (Raman et al. unpublished results) or sampling methods. **mpdcm** is mostly written in the C programming language and is accompanied by a MATLAB interface compatible with Statistical Parametric Mapping (SPM). It is available under the GPL license as part of the open source TAPAS software at www.translationalneuromodeling.org/software.

There is a growing interest in using GPUs in the context of imaging (Shi et al., 2012; Eklund et al., 2013) and particularly in the field of MRI. For example, Hernandez et al. (2013) used a massively parallel algorithm to accelerate a sampling method for diffusion weighted imaging. In the case of fMRI, mature toolboxes covering several aspects of the analysis pipeline, including coregistration and statistical inference, are available (Eklund et al., 2014). Ferreira da Silva (2011) used GPUs to apply Bayesian inference to BOLD time series modeling using sampling methods. Recently, Jing et al. (2015) proposed to accelerate group ICA by using GPUs, a method integrating a toolbox originally developed for EEG (Raimondo et al., 2012). In the context of DCM, Wang et al. (2013) presented a massively parallel implementation of DCM for event-related potentials. This implementation aimed mostly at speeding up the variational EM algorithm for inference in DCM for EEG. Here, we propose a more general approach in the context of DCM for fMRI that can not only increase the efficiency of the variational EM algorithm, but also the efficiency of, for example, sampling methods. An introductory account of parallel programming can be found in Suchard et al. (2010).

In the following, we first briefly introduce DCM and present our toolbox. We then compare the accuracy and performance of **mpdcm** to the standard implementation in SPM. We also showcase the implementation of path sampling and parallel tempering, two computationally expensive sampling algorithms that can be easily optimized with **mpdcm**. Finally, we discuss future applications, extensions, and limitations.

2. Methods

2.1. Dynamic causal modeling

Dynamic causal models describe continuous time interactions between nodes (neuronal populations) in a pre-specified neural network. The time course of neuronal activity is modeled bythe bilinear differential equation

$$\dot{x} = Ax + \sum_{n=1}^{N} u_i, B_i x + C_u, \tag{1}$$

where x is a vector of neuronal states, u is an N dimensional vector of inputs to the network, and A, B, and C are matrices that represent the connection strengths between nodes and the influence of external inputs on nodes and connections. The evolution of x therefore depends on both the interactions between nodes and experimental inputs. In order to predict measurable signals, DCM for fMRI also invokes a biophysical model, which consists of weakly nonlinear differential equations that link neuronal activity to BOLD signals, including an extended Balloon model (Buxton et al., 1998; Friston et al., 2003; Stephan et al., 2007). Briefly, the BOLD signal is modeled as a nonlinear function of the venous compartment volume and the deoxyhemoglobin content (Stephan et al., 2007). The venous compartment is assumed to behave as a balloon, where the total blood volume is the difference of the blood inflow and outflow. The deoxyhemoglobin content is simply the difference of (1) the product of the blood inflow and oxygen extraction rate and (2) the outflow rate times the deoxyhemoglobin concentration. Importantly, blood inflow into the venous compartment is triggered by putative brain activity x. For a detailed treatment see Stephan et al. (2007). Because the Balloon model comprises nonlinear differential equations that cannot be solved analytically, numerical integration is necessary to generate predictions from the model. To simplify notation from here on, we refer to all states (neuronal or hemodynamic) as x and the state equations as f, such that:

$$\dot{x} = f\left(x, u, \theta\right), \tag{2}$$

where u are inputs, and θ are subject specific parameters, including matrices A, B and C, as well as several hemodynamic parameters.

States x are linked to observable quantities via a forward model that we denote by the function g:

$$\hat{y} = g\left(x, u, \theta\right), \tag{3}$$

where $\hat{y}$ is the predicted BOLD signal.

DCM can be used in two different ways: identification of system structure and parameter estimation. The former requires model comparison, i.e., determining which model, from a predefined set of alternatives (model space), best explains a set of empirical observations. Typical differences between models include the absence or presence of certain connections or modulations thereof (entries in matrices A or B). The latter consists in determining the posterior distribution of the parameters θ given experimental observations and a model (Penny et al., 2010). Summary statistics of these distributions can be used, for example, as dependent variables in statistical tests, which can indicate the effect of a treatment or condition on effective connectivity.

Model comparison (Penny et al., 2004) and parameter estimation (Friston et al., 2003) rest on the generative nature of DCM. A generative model requires two things: first, a likelihood function that defines the probability of observed data given a prediction or simulation of the model. Model predictions are fully specified by the parameters θ, experimental input u, the neuronal state equation $f(x, u, \theta)$, and the forward model $g(x, u, \theta)$. Second, a generative model requires a prior distribution of the parameters θ, i.e., the probability of θ before any observation. The relationship between these elements is determined by Bayes' formula:

$$p\left(\theta\,|y, m\right) = \frac{p\left(y\,\middle|\,\theta, m\right) p\left(\theta\,|m\right)}{p\left(y\,|m\right)}, \tag{4}$$

where y is a set of experimental observations, θ are the parameters, and m is the model being evaluated. In Bayesian statistics,

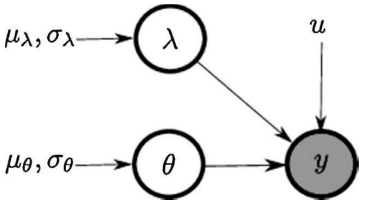


Fig. 1. Graphical representation of the generative model underlying DCM. All distributions are Gaussian. Experimental data y is modeled as being conditionally dependent on the model parameters θ and a region-wise noise scaling parameter λ.

parameter estimation (model inversion) is equivalent to computing the posterior probability $p\left(\theta \left| y, m\right.\right)$. Identification of system structure can be achieved by comparing different models on the basis of their marginal likelihood or model evidence $p(y|m)$ (Kass and Raftery, 1995; Penny et al., 2004). This quantity is the probability of a set of observations y given a model m after marginalizing over parameters θ. Comparing models via their marginal likelihood is an alternative or complementary approach to cross-validation and other methods routinely used in frequentist statistics for model comparison (Kass and Raftery, 1995; Gelman et al., 2003; MacKay, 2003)

DCM is defined by a Gaussian noise model and Gaussian priors (Friston et al., 2003). Assuming that there are $r = 1, \ldots, R$ nodes or brain regions, the model is specified by the joint probability

$$p\left(y_1, \ldots, y_R \left| \theta, \lambda_1, \ldots, \lambda_R, m\right.\right) = p\left(\theta|m\right) \prod_{r=1}^{R} p\left(y_r \left| \theta, \lambda_r, m\right.\right) p\left(\lambda_r|m\right), \quad (5)$$

$$p\left(y_r \left| \theta, \lambda_r, m\right.\right) = N\left(y_r | g\left(x_r, u, \theta\right), \lambda_r^{-1} I\right), \quad (6)$$

$$p\left(\theta|m\right) = N\left(\theta \left| \mu_\theta, \sigma_\theta^2 I\right.\right), \quad (7)$$

$$p\left(\ln \lambda_r|m\right) = N\left(\ln \lambda_r \left| \mu_\lambda, \sigma_\lambda^2\right.\right), \quad (8)$$

where N denotes the Gaussian probability density function, y_r corresponds to the time series extracted from brain region r, λ_r corresponds to a region-specific scaling factor of the covariance of the noise of the observed BOLD signal, and I is an identity matrix (assuming temporal auto-correlation in the BOLD signals has been removed by whitening). A graphical representation of the model is presented in Fig. 1.

In summary, classical DCM for fMRI assumes a bilinear model of the interactions between brain regions and a nonlinear model of the ensuing BOLD signal. Simulations of the model are used to construct a likelihood function by assuming that the observed BOLD signal is Gaussian distributed around the predicted signal. Given the nonlinear model of the hemodynamics that cause the BOLD signal, predicting data points requires numerical integration of the differential Eq. (2).

The main computational hurdle for DCM is that it requires the simulation of predicted BOLD signals. The goal of **mpdcm** is to accelerate the numerical integration of the biophysical model of DCM by exploiting the computational power of GPUs. Here, we optimized only the generation of biophysical simulations, because the computational costs of DCM are mostly incurred by the simulation step. This implies that **mpdcm** can be used to optimize any inference algorithm that relies on a model with the same likelihood function, independently of any other aspects of the model. In the next section, we briefly present the implementation details of our toolbox.

2.2. Massively parallel simulation of DCMs

The temporal integration of DCM's biophysical model on massively parallel architectures is accompanied by three main challenges: First, the problem is iterative in nature. This property forces a serial implementation of the time updates and imposes a ceiling to the performance gains achievable. Thus, part of our goal was to exploit parallelizations on each update step to minimize the serial component of the algorithm. Second, the integration of DCM's dynamic system cannot be rendered into the framework of single instruction/multiple data paradigm, where GPU architectures excel, because each state of the model evolves according to different dynamics. Third, connectivity between areas and the dynamics of the states require high levels of communication between threads and thus increase the effect of memory latencies.

We implemented two integration schemes in **mpdcm**. First, we implemented an integrator based on Euler's method (Butcher, 2008), which approximates x at time $t+h$ through the equation:

$$x_{t+h} = x_t + hf\left(x_t, u_t, \theta\right). \quad (9)$$

Euler's integration scheme does not require the storage of any intermediate result between iterations and hence has low memory requirements. Although memory is nowadays often not a major concern in CPU-based computations, memory access and size are usually the main bottlenecks for GPU-based computations. The advantages of Euler's integration come at the price of reduced accuracy and, more importantly, possible accumulation of errors. In order to reduce these inaccuracies, we also implemented a fixed step sized Runge–Kutta scheme of fourth order (Butcher, 2008). This standard technique uses a linear combination of different, iterative evaluations of f to approximate the dynamics of x. Runga–Kutta's method requires four times more floating point operations that cannot be parallelized and, more importantly, twice as much memory for the storage of intermediate results. In order to reduce the memory needed to store intermediate steps, we implemented a modified version of Runge–Kutta's method (Blum, 1962). This modification takes advantage of linear dependencies between the coefficients of f to reduce the number of required registers for each state variable from four to three registers. Both integration methods are compared in Section 3.

In our GPU implementation, we exploited three sources of parallelism. First, simulations for systems with different sets of parameters θ or inputs u can be computed in parallel. As it will become clear in the next section, several applications require simulations from the same model with different parametrizations. Second, the update of each node of the network can be performed in parallel. Finally, because f is a multivariate function, its evaluation in each dimension can be parallelized. Similarly, we parallelized the evaluation of the forward model g for each region.

An important consideration in our implementation is memory access. In contrast to CPUs, GPUs have only a very limited amount of fast memory that can be shared between threads in a block. Currently, a single multiprocessor has by default access to 48 kb of high latency shared memory, and access to a high latency global memory of several gigabytes. In **mpdcm**, the experimental inputs u are loaded to global memory and kept there, while model parameters are stored in shared memory. States are stored in shared memory, and predicted output is stored in global memory and transferred to the CPU memory after execution. This implies that each time step requires the access to slow global memory.

In Algorithm 1 we present pseudo code of our implementation of Euler's integration method. Several details have been

left out in order to clarify the main aspect of the implementation. Euler's method can be easily extended to accommodate the Runge Kutta's fourth order method. The implementation of the Runge Kutta's algorithm is described in the supplementary materials.

Algorithm 1:
Requires: *itheta*—array of pointers to model parameters
Requires: *ntheta*—number of elements of *itheta*.
Requires: *iu*—array of experimental inputs.
Requires: *iy*—arrays of pointers to predicted signals.
Requires: *nt*—total number of integration steps.
Requires: *dt*—1 over sampling rate (integer)
Requires: *h*—time constant.
Requires: *nregions*—number of regions.
Requires: *NUM_SM_PROCESSORS*—number of streaming multiprocessors.
Requires: *MAX_THREADS*—maximal number of threads supported by each multiprocessor.
Requires: *MAX_NUM_NODES*—maximal network size supported.
1: Allocate shared float array *buf*0 of size *MAX_NUM_NODES* * 5.
2: Allocate shared float array *buf*1 of size *MAX_NUM_NODES* * 5.
3: Allocate local pointer to model parameters *theta*.
4: Allocate local pointer to float array *y*.
5: Allocate local pointer to float array u.
6: *nblocks* = *NUM_SM_PROCESSORS* * (*MAX_THREADS*/(*MAX_NUM_NODES* * 5))
7: Set thread block size to (*MAX_THREADS*, 5).
8: Set grid size to (*nblocks*, 1).
9: *l* = *blockId.x* * (*MAX_NUM_NODES*/*nregions*) + (*threadId.x*/*nregions*)
10: **while** *l* < *ntheta* in parallel **do**
11: *inrange* = *threadId.y* < *nregions* * (*MAX_NUM_THREADS*/*nregions*)
12: *nlow* = *nregions* ∗ (*threadId.y*/*nregions*)
13: *nhigh* = *nlow* + *nregions*
14: *theta* = *itheta*[*l*]
15: *u* = *iu*[*l*]
16: *y* = *iy*[*l*]
17: **for** *t* = 0 **to** *nt* − 1 **do**
18: **for** *i* = 0 **to** *MAX_NUM_NODES* − 1 in parallel **do**
19: **for** *j* = 0 **to** 4 in parallel **do**
20: **if** *inrange* **then**
21: *buf*1[*i*, *j*] = *buf*0[*i*, *j*] + *h* ∗ *f*_*i*(*buf*0[*nlow* : *nhigh*, 0 : 4], *theta*, *u*[*nt*], *i*)
22: **end if**
23: **end for**
24: synchronize threads
25: **if** *nt* % *dt* == 0 **and** *inrange* **and** *threadId.y* == 0 **then**
26: *y*[*nt*/*nd*, *threadId.x* % *nr*] = *g*(*theta*, *buf*0[*i*, 0 : 4])
27: **end if**
28: **end for**
29: Swap *buf*0 and *buf*1
30: **end for**
31: *l* = *l* + *nblocks* * (*MAX_NUM_NODES*/*nregions*)
32: **end while**

We exploited three levels of parallelization. First, several simulations can be performed in parallel (line 10). Second, the state equations of each node are computed in parallel (line 18–19). Crucially, by setting *MAX_NUM_NODES* to a multiple of the minimal warp size, different thread warps are assigned to the evaluation of different state equations (line 19). Importantly, to maximize occupancy, the number of blocks is based on the maximal number of possible threads that can be hold by each streaming multiprocessor (line 6). Further implementational details are omitted here for clarity.

2.3. Comparing simulation accuracy

We verified the accuracy of Euler's and Runge–Kutta's method by comparing simulations from **mpdcm** against simulations from SPM's function *spm_int_J.m*, which was introduced for nonlinear DCMs (Stephan et al., 2008). All results are based on SPM8, revision 5236. The implementation in *spm_int_J.m* uses a local linearization of the dynamics of *x* (Ozaki, 1992). The update of *x* is given by

$$x_{t+h} = x_t + J^{-1}\left(e^{hJ} - I\right) f\left(x, u, \theta\right), \tag{10}$$

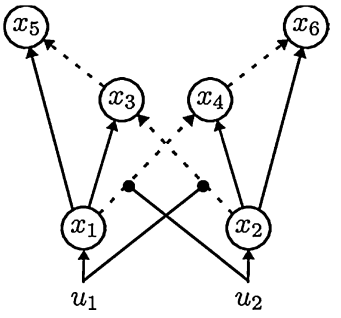


Fig. 2. Exemplary DCM network. Solid arrows represent positively signed (excitatory) connections. Broken arrows correspond to negatively signed (inhibitory) connections. Self-inhibitory connections are not displayed. Inputs u_1 and u_2 activate directly only nodes 1 and 2, and modulate the connections between nodes 1 and 4 and between nodes 2 and 3.

where J is the Jacobian of f with respect to x, and I is the identity matrix. Although this scheme is a highly accurate method, it requires the computation of the Jacobian of f, one evaluation of f, and one matrix exponentiation per update.

To evaluate the three integration schemes, five sets of model parameters of a six node network were drawn randomly. Fig. 2 displays the structure of the network. We assumed a TR of 2.0 s. The inputs u had a sampling rate of 8.0 Hz and consisted of regular box car functions with a width of 20.0 s. Box car functions are commonly used to model DCM inputs as they represent categorical experimental conditions. The step size h used for all simulations was 0.125 s, therefore matching with the sampling rate of u. A total of 512 scans were simulated.

2.4. Comparing simulation performance

We compared the performance of **mpdcm** against the default integrator in SPM used for inference. This method is implemented in *spm_int.m*. It relies on a linear approximation of f

$$\dot{x} = f\left(x, u, \theta\right) \approx J(u)x. \tag{11}$$

Using the same numerical integration scheme as in Eq. (10), which effectively corresponds to the analytical solution assuming linearity of f with respect to x, the iterative updates are given by

$$x_{t+h} = e^{hJ(u_t)}x_t. \tag{12}$$

Because $J(u)$ needs to be recomputed only when u changes, increasing the sampling rate of u when u is composed of only step functions does not change the results nor the number of computations necessary to generate simulations (Friston, 2002). This method is a compromise between computational time and accuracy, as it approximates f with a linear system, but preserves the nonlinearities in g.

To further evaluate the performance gains provided by **mpdcm**, we also implemented a highly optimized, multithreaded *C* version of Euler's method and compared it against **mpdcm**. This function is also available as part of our toolbox. We use *openmp* to multithread our implementation at the level of simultaneous simulations. Although a more fine-grained parallelization is possible, for a large number of simulations this simple scheme warrants constant high occupancy, as no synchronization is required. Moreover, a more fine-grained parallelization would also come at the cost of increased latency in the initialization of parallel threads, a problem not found in GPU units (Suchard et al., 2010).

We compared the differences in timing of a single simulation and then proceeded to test the gains in performance when simulating several models in parallel. Because SPM's method lacks

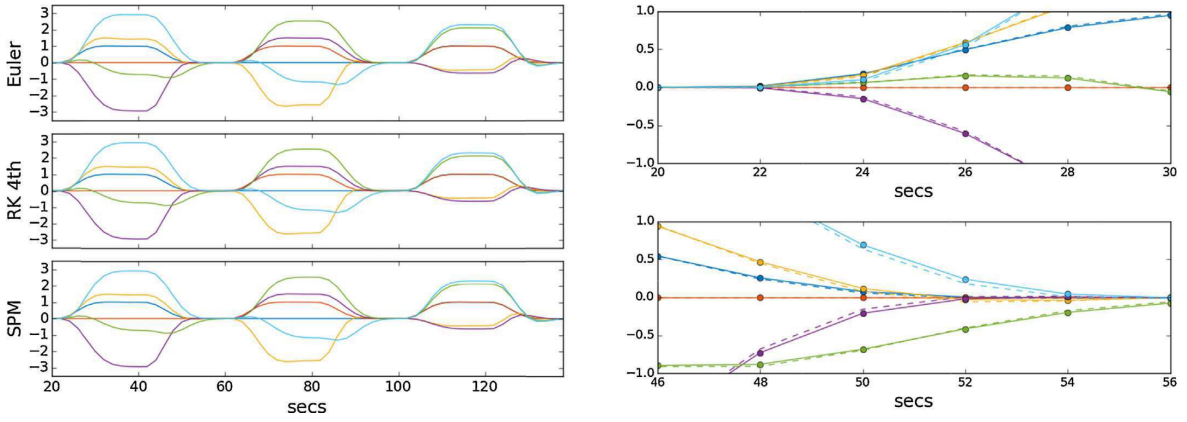


Fig. 3. Left: simulated BOLD signal using SPM, Euler and Runge–Kutta fourth order integration methods. Right: The same data zoomed-in in two different intervals. Solid lines correspond to SPM, broken lines to Euler's method, and circles to Runge–Kutta's method. Differences are only apparent during transients.

any optimization for parallel simulations, its computational time of multiple simulations is simply additive. Thus, we did not apply SPM's method to more than one simulation. The experimental input u used for these simulations had the same parameters as in the previous section.

2.5. Environment

Currently, **mpdcm** targets *Nvidia* graphical processing units and is written mostly in the *C* programming language using the Compute Unified Device Architecture (CUDA) API. *Nvidia* GPUS are off-the-shelf chip sets that make massively parallel architectures affordable for standard desktop computers. **mpdcm** has no dependencies on specialized libraries provided by the manufacturer except for the standard *C* library and the *math* library.

Simulations were performed on a dedicated machine with a twelve core AMD Opteron 6174 processor, a professional *Nvidia* Tesla M2050, Linux CentOS 6.5, and MATLAB 8.1. **mpdcm** was compiled with gcc 4.4.7. The Tesla M2050 has a single precision arithmetic peak performance of 1030.6 giga floating point operations per seconds (GFLOPS), and 512.2 GFLOPS for double precision arithmetic. All GPU and CPU simulations were performed on this machine if not indicated otherwise. In order to investigate the differences in performance that single and double floating point representation could produce, some of the simulations were performed on a machine equipped with an Intel i7 4770k, a *Nvidia* Geforce GTX 760, Linux Ubuntu 13.10, gcc 4.8.1 and MATLAB 8.1. The peak performance of single precision arithmetic of the GTX 760 is 2257.0 GFLOPS and only 94 GFLOPS of double precision arithmetic. It is important to notice that our goal was not to benchmark the two GPUs, but to evaluate the performance change produced by different ratios of single to double precision peak FLOPS. All other simulations were performed in single precision, unless otherwise stated. In the remainder of this paper, we first compare accuracy and performance gains in the simulation of BOLD data in DCM and then explore the application of our toolbox for model comparison.

3. Results

3.1. Accuracy

Fig. 3 shows a fragment of one simulation of the network using **mpdcm** and SPM. In this case, the range of the BOLD signals was $[-5.0, 3.4]$. The absolute difference between the signal simulated using Euler and Runge–Kutta integration schemes was always lower than 0.11 (max. mean absolute difference 0.013). Euler's method diverged always less than 0.11 from SPM (max. mean absolute difference 0.013), whereas the Runge–Kutta's method diverged less than 4×10^{-4} from SPM (mean absolute difference 9×10^{-6}). To compare all five sets of random parametrizations, we computed the variance of the difference between the predicted signal using Euler's and SPM's method, and Runge–Kutta's and SPM's method as a percentage of the total variance of the predicted signals. For simplicity, the total variance of the signal was computed using the simulations from SPM. We found that the variance of the difference from the predicted signals was never larger than 0.1% (Euler's method) and $10^{-5} \times 0.1\%$ (Runge–Kutta's method) of the total variance of the signal indicating that all three integration schemes resulted in highly similar simulations, with Runge–Kutta and SPM yielding virtually identical simulations. Finally, we considered the effect of floating point number representation on the simulations. In order to verify that single precision representation did not cause any systematic rounding errors, we repeated the simulations both under single and double precision and compared the relative error of the single precision implementation with respect to the double precision. The relative error was defined as

$$\frac{abs(double(svalue) - dvalue)}{abs(dvalue)} \tag{13}$$

where *svalue* and *dvalue* stand for the values obtained using single and double precision, respectively. We found that the maximal relative error did not exceed 2×10^{-5} for both SPM and **mpdcm**.

3.2. Performance gains

One of the main reasons for introducing **mpdcm** was to increase performance for costly simulations needed for inversion of DCM models. Here, performance on a three and six node network was assessed by running 20 consecutive simulations. The median values were used as summary statistics. Only minimal variance in the execution times was observed. Fig. 4 summarizes the results.

mpdcm can reduce the computation time required for generating simulations from DCM with respect to SPM by up to two orders of magnitude. It is important to note that although increasing the sampling rate of the input u has a linear effect on the performance of **mpdcm**, it has no effect on SPM's method. Thus, even though in practice the sampling rate of the input u was 16 times larger in the simulations generated with **mpdcm**, our toolbox performed better while preserving the nonlinearities of f. The difference between

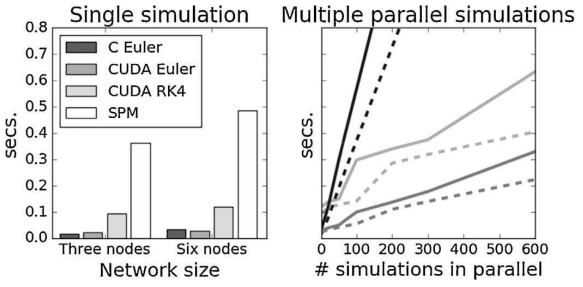


Fig. 4. Performance measurements. Left panel: Single simulation. Right panel: Multiple parallel simulations. The broken and solid lines correspond to DCMs with 3 and 6 nodes, respectively. Colors are matched to the left panel. Vertical axis is shared between panels. We found that mpdcm was faster integrating a single model (left panel) than the standard SPM integrator. This advantage scales with the number of regions, as the computation of f is performed in parallel. In the right panel, the time required for performing simulations in parallel is displayed as a function of the number of simultaneous parallel simulations. **mpdcm**'s Euler scheme was approximately 10 times faster (9.8 times for a 3 nodes network, and 9.9 times for a 6 node network) than the C implementation when 600 systems where integrated in parallel.

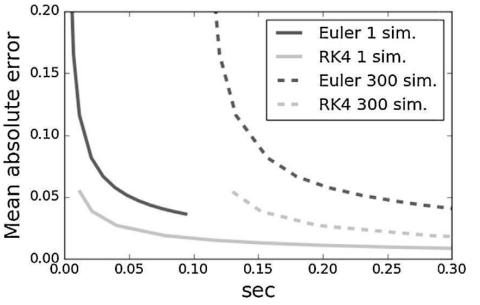


Fig. 6. Mean absolute error as a function of computation time in seconds. As expected, the error of the Euler method decreased slowly compared to Runge–Kutta's method.

the GPU implementation and the multithreaded CPU implementation is lower, with the GPU Euler's method being an order of magnitude faster than a CPU implementation when 600 simulations are performed in parallel. **mpdcm** is only marginally faster than a CPU implementation when a single system is evaluated.

We proceeded to investigate whether the performance of **mpdcm** was affected by the floating point number representation used. In Fig. 5, we compared the performance of **mpdcm** when using single and double precision floating point arithmetic in two different GPUs. In the *Nvidia* Geforce GTX 760, in the case of the Euler integrator, double precision was only marginally slower than single precision. However, double precision representation had a larger effect on Runge–Kutta's method. In particular, we found that for 600 parallel simulations, a double precision implementation was almost between two and three times slower than a single float implementation. This effect disappeared on the *Nvidia* 2050 M, where only minor differences between single and double precision arithmetic were found.

Finally, we considered the increases in accuracy obtained as a function of computation time. We compared Euler's and Runge–Kutta's method by simulating a six node network with the same parameters as those used in the previous section. However, input u was sampled at only 0.5 Hz while the time step h was varied from 1.0 s to 0.03514 s. A single simulation and 300 parallel and simultaneous simulations were compared. The mean absolute error was obtained against a simulation using Runge–Kutta's method with a time step of $h = 2^{-8}$. Results are displayed in Fig. 6.

Please note that these results should be considered only a rough indication of performance gains. Differences in performance depend on several parameters, such as the number of nodes, TR, sampling rates, input and output size and structure, and hardware. Also, in the case of CPU implementation the linear increase in time can of course be compensated by the use of a larger number of CPUs.

4. Applications

4.1. Thermodynamic integration

mpdcm was developed to increase the performance of DCM in the context of computationally demanding inference problems. One particularly interesting challenge is the evaluation of the model evidence or marginal likelihood of a DCM.

Comparison of alternative DCMs is usually done by comparing the marginal likelihood between models given a data set y. This corresponds to computing the Bayes factors between models:

$$\frac{p(y|m_1)}{p(y|m_2)} = \frac{\int p\left(y\middle|\theta, m_1\right) p\left(\theta|m_1\right) \mathrm{d}\theta}{\int p\left(y\middle|\theta, m_2\right) p\left(\theta|m_2\right) \mathrm{d}\theta}. \tag{14}$$

Although a powerful approach, exact computation of marginal likelihoods is usually not possible because almost all but the simplest models lead to intractable integrals in Eq. (14). For this reason, approximations such as the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) or the negative free energy are typically used (Penny et al., 2004). In addition, several sampling-based methods have been proposed in the literature to compute

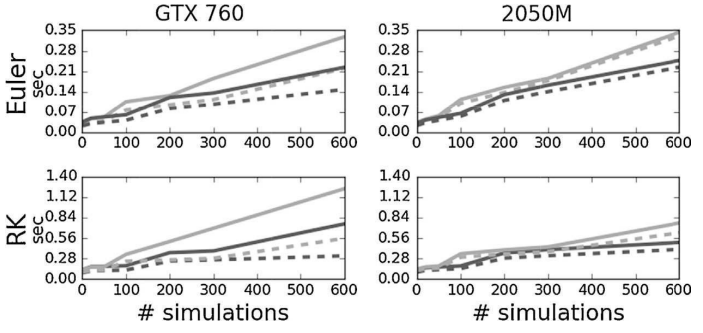


Fig. 5. Computation time in seconds as a function of the number of parallel simulations in two different Nvidia cards. Black and gray lines display DCMs of 3 and 6 regions. Solid and broken lines display results with double and single precision arithmetic, respectively. The effect of floating point representation was more pronounced in the GTX 760.

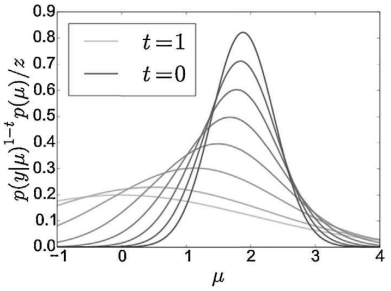


Fig. 7. Normalized power posterior densities $p\left(y\left|\theta,m\right.\right)^{1-t}p\left(\theta\left|m\right.\right)/z$. The likelihood and prior distributions are Gaussian. The prior distribution of μ is $N(0,\sqrt{2})$. y is a single observation with $y=2$. Eight different values of t are displayed. As the temperature decreases, the posterior distributions' variance decreases.

the evidence of intractable models (Kass and Raftery, 1995). For example, the identity

$$\int \frac{1}{p\left(y\left|\theta,m\right.\right)}p\left(\theta\left|y,m\right.\right)\mathrm{d}\theta=\frac{1}{p\left(y\left|m\right.\right)}, \tag{15}$$

suggests using samples from the posterior distribution to estimate the marginal likelihood, an approach called the posterior harmonic mean estimator. Although an asymptotically unbiased estimator, this approach has several shortcomings, including a potentially infinite variance (Raftery et al., 2007). Gelman and Meng (1998) introduced in the statistical literature an alternative method to estimate marginal likelihoods called thermodynamic integration, originally proposed by Kirkwood (1935). In this approach, the difference in the normalization constants between two distributions is computed by constructing a continuous path between them. By integrating along this path, it is possible to compute differences in marginal likelihoods. More formally, thermodynamic integration is based on the equivalence (or a more general formulation thereof):

$$\ln p\left(y\left|m\right.\right)=\ln\int p\left(y\left|\theta,m\right.\right)p\left(\theta\left|m\right.\right)\mathrm{d}\theta-\ln\int p\left(\theta\left|m\right.\right)\mathrm{d}\theta \tag{16}$$

$$=\int_1^0 E_{p(y|\theta)^{1-t}p(\theta)}\ln p\left(y\left|\theta,m\right.\right)\,\mathrm{d}t \tag{17}$$

The path between prior and posterior is constructed by weighting the log likelihood function by the temperature $1-t$ with $t\in[0,1]$. A graphical illustration of this path is displayed in Fig. 7. In practice this is implemented by first selecting a finite set of temperatures $0=t_1<t_2<\cdots<t_{N-1},<t_N=1$, then drawing samples from the unnormalized power posteriors $p\left(y\left|\theta,m\right.\right)^{1-t}p\left(\theta\left|m\right.\right)$, and finally numerically computing the integral in Eq. (17) using quadratures. Although it has been shown that in practice this method has higher accuracy and precision than alternative sampling methods (Calderhead and Girolami, 2009), it requires evaluating N different power posteriors, with its accuracy increasing with N.

Samples from the power posteriors can be obtained using the Metropolis–Hastings algorithm. In this case, the likelihood of a sample drawn from the proposal distribution can be used for the acceptance step. Since the computation of the likelihood is the most expensive step, **mpdcm** greatly diminishes the run time of this algorithm.

4.2. Parallel tempering

If thermodynamic integration is implemented using the Metropolis–Hastings algorithm, it can be extended to improve the mixing of the Markov chains by allowing parallel chains to exchange information without changing their stationary distribution. This method is referred to as "population MCMC", "replica exchange MCMC", or "parallel tempering" (see Sengupta et al., 2015). The basic intuition underlying this approach is the idea of a set of Markov chains $n=1,\ldots,N$ exchanging information in a way reminiscent of biological evolution, where chains in the population "mutate" and "mate" (Laskey and Myers, 2003). This intuition is formalized by the product distribution

$$\pi_1\left(\theta_1\right)\pi_2\left(\theta_2\right),\ldots,\pi_{N-1}\left(\theta_{N-1}\right)\pi_N\left(\theta_N\right), \tag{18}$$

where $\theta_1,\ldots,\theta_N$ represent a population of parameters. Swendsen and Wang (1986) originally proposed to condition the distributions $\pi_1,\ldots,\pi_N$ with a temperature parameter in the same way as in thermodynamic integration, i.e.,

$$\pi_n\left(\theta\right)\propto p\left(y|\theta,m\right)^{1-t_n}p\left(\theta|m\right), \tag{19}$$

with $0=t_1<\cdots<t_N=1$. In order to propagate information between chains, an *exchange* operator swaps the parameters θ_i with θ_j with probability:

$$\min\left(1,\frac{p\left(y\left|\theta_j,m\right.\right)^{1-t_i}p\left(y\left|\theta_i,m\right.\right)^{1-t_j}}{p\left(y\left|\theta_j,m\right.\right)^{1-t_j}p\left(y\left|\theta_i,m\right.\right)^{1-t_i}}\right). \tag{20}$$

Since the exchange is reversible, the stationary distribution of each chain does not change. In practice, samples from each chain are drawn from the distribution defined in Eq. (18) and then samples from the population are exchanged according to Eq. (20). Samples drawn using this method can be used for either parameter estimation or for computing the evidence of a model through thermodynamic integration. Informally, since chains at higher temperature can explore the parameter space more freely, parallel tempering can dramatically improve the mixing of each particular chain (Calderhead and Girolami, 2009).

Pseudocode for parallel tempering using **mpdcm** is provided in Algorithm 2 (Altekar et al., 2004). The implementation generalizes path sampling with independent chains by adding a swapping step. Using path sampling and parallel tempering has already been suggested in the context of DCM (Sengupta et al., 2015). Our goal here is to show that these algorithms can be easily and efficiently implemented through **mpdcm**.

Algorithm 2:
Requires: $M\geq 1$ (Number of samples)
Requires: $S\geq 1$ (Number of swaps)
Requires: $0=t_1\leq\cdots\leq t_N=1$ (Temperature schedule)
for $m=0$ **to** $M-1$ **do**
  for $n=0$ **to** $N-1$ **do**
    Sample $\theta_n^{(m)}$ *from* $p\left(y|u,\theta_n\right)^{1-t_n}p(\theta_n)$
  end for
  for $n=0$ **to** $N-1$ **in** parallel **do**
    $Computel_n^{(m)}=p\left(y\left|u,\theta_n^{(m)}\right.\right)$
end for
  for $s=0$ **to** $S-1$ **do**
  Randomly select k *with* $1\leq k\leq N-1$
with probabilitymin $\left(1,\frac{p\left(y|u,\theta_k^{(m)}\right)^{1-t_{k+1}}p\left(y\left|u,\theta_{k+1}^{(m)}\right.\right)^{1-t_k}}{p\left(y\left|u,\theta_k^{(m)}\right.\right)^{1-t_k}p\left(y\left|u,\theta_{k+1}^{(m)}\right.\right)^{1-t_{k+1}}}\right)$, *swap samples* $\theta_k^{(m)}$ *and* $\theta_{k+1}^{(m)}$
  end for
end for
for $n=0$ **to** $N-1$ **do**
Compute $c_n=\frac{1}{M}\sum_{n=1}^{M}\ln l_n^{(m)}$
end for
return $\frac{1}{2}\sum_{n=1}^{N-1}(t_n-t_{n-1})(c_n+c_{n-1})$

Although the computational costs of parallel tempering are similar to those of path sampling with independent chains, information

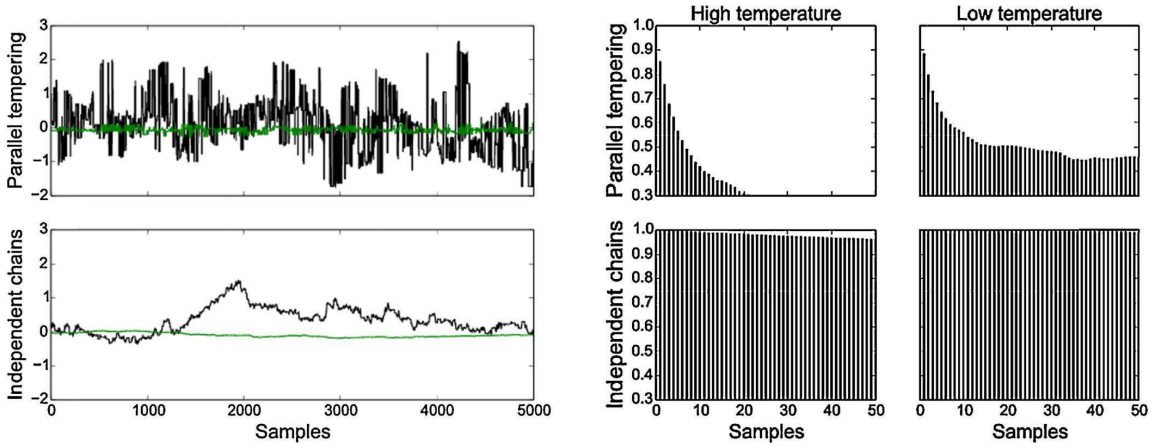


Fig. 8. Left: Samples from the same posterior distribution using independent chains and parallel tempering. Chains at low and high temperature are displayed in green and black respectively. Right: Autocorrelation function of the samples. In MCMC, samples are by definition correlated, although their stationary distribution corresponds to the target distribution. If samples are strongly correlated, more samples are necessary to estimate the moments of the target distribution. Since parallel tempering reduces the correlation between samples, the total number of necessary simulations diminishes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

exchange between chains makes the implementation of efficient parallel algorithms in computer grids far more challenging (Li et al., 2007). In particular, while path sampling can be implemented with completely parallel CPUs, parallel tempering requires the synchronization of the chains. Since **mpdcm** uses a single CPU and only the simulations are performed on a GPU, parallel tempering can be easily implemented. Fig. 8 displays samples of one connectivity parameter drawn using parallel tempering. The exchange operator effectively decorrelates samples drawn from each chain, increasing the statistical efficiency of this method at minimal computational costs.

Running parallel tempering using a purely CPU implementation is possible but the computational costs are much larger than with **mpdcm**. Assuming a DCM of 6 regions with 512 volumes, a total of 100 parallel chains, 40,000 iterations of the MCMC algorithm, and counting only the time required to perform simulations—based on results displayed in Fig. 4, it can be inferred that SPM would require over 22 days to finish the inversion of one session. For a multithreaded CPU implementation in C, the time required would be reduced to 6.3 h, and for **mpdcm** Euler's method, the time required would be reduced to 1.1 h.

Both thermodynamic integration as well as parallel tempering are implemented in **mpdcm** in the function tapas_*mpdcm_fmri_estimate.m*. Besides these two algorithms, we have implemented a maximum *a posteriori* estimator for single and multiple subjects and several other routines specified in the documentation.

5. Discussion

In this paper, we presented **mpdcm**, a toolbox aimed at speeding up simulations from the generative model that underlies DCM. Our implementation does not compromise the nonlinearities in the model and is nevertheless computationally efficient. We used two integration methods for generating simulations: Euler's method and a modified Runge–Kutta's fourth order method. The two integration schemes did not strongly differ from each other with regard to the simulated BOLD signal. They also did not strongly differ from the SPM's method based on Ozaki's scheme (Ozaki, 1992). However, they strongly differed in their performance (Fig. 4). Euler's method is more efficient because of the lower number of floating point operations and reduced memory complexity. Although more stable and precise methods exist, the characteristics of typical fMRI experiments and of DCM make Euler's method an acceptable alternative when precision is strongly outweighed by the need of a large number of simulations. Highly precise simulations can still be generated using Runge–Kutta's method without large additional computational time needed. Performance differences between a multithreaded CPU implementation of Euler's method and **mpdcm** reached one order of magnitude, similarly to benchmarks published before for other algorithms (Lee et al., 2010b).

Higher order integrators were not considered here, as integrators of order 5 or higher increase the memory needs of most algorithms. Low storage explicit Runge–Kutta methods are available (Williamson, 1980; Ketchenson, 2010), and we implemented Blum's (1962) modified Runge–Kutta's fourth order algorithm, which requires only 3 registers for each state variable. Although not presented here, our toolbox includes an adaptive size integrator based on Bogacki and Shampine (1989) third order, embedded algorithm. This scheme has several computational advantages, and adaptive size integrators tend to outperform fixed step implementations. However, by nature, the number of iterations of this algorithm is unknown at scheduling time, a characteristic that makes the allocation of GPU resources much less efficient.

The double precision floating point format as defined by the IEEE 754 standard (see Goldberg, 1991) is the default data type of several scientific applications. However, it is not always clear whether the advantages inherent to this format are necessary in all applications. Double precision floats provide a dynamic range of over 600 orders of magnitude, a range which is not needed in, for example, brain imaging. Moreover, not all applications require more precision than a small number of decimal places, while double precision format provides approximately 15 decimal places of precision. Regarding performance, in the case of ×86 architectures, double and single precision arithmetic is implemented by the same 80 bit floating point unit. Therefore, in ×86 architectures, the main performance difference between both formats is due to memory latencies and bandwidth. In contrast to this, in the case of consumer level GPUs the theoretical ratio between peak double and single precision FLOPS is often close to 1:24 or 1:32. We found that double precision arithmetic (compared to single precision) reduced

performance approximately by a factor of 3 in a GPU with a very low double to single floating point peak performance ratio (1:32). This difference was reduced to roughly 1:2 in a GPU designed for professional purposes.

The performance comparison between double and single precision indicates that **mpdcm** is largely memory bounded, as the difference in performance between the double and single precision performance are far from the differences between peak FLOPS. Importantly, we did not find any notable effects of rounding-off errors in the single precision implementation of **mpdcm,** with the highest difference being far below the signal to noise ratio that is typical of fMRI experiments. Single precision floating point arithmetic is thus sufficiently accurate and recommended for **mpdcm,** as the effect on the performance of Runge Kutta's method is significant. This is particularly important for consumer-level GPUs, where the peak single precision arithmetic performance is much larger than the double precision peak performance.

mpdcm is particularly attractive when several systems are simulated in parallel, because under those conditions performance gains of up to two orders of magnitude with respect to SPM and one order of magnitude with respect to a multithreaded *C* implementation become evident. In order to show how to exploit these large increases in performance, we presented an algorithm that can be easily optimized with our toolbox. In particular, we presented a GPU implementation of parallel tempering, a variant of MCMC that is known to improve the mixing properties of the Markov chains of the Metropolis–Hastings algorithm, reducing the computational costs with respect to thermodynamic integration with independent chains. Implementing parallel tempering using **mpdcm** is straightforward, as all the parallelization is managed internally at the GPU level. Although the computational gains obtained through our toolbox have been illustrated using sampling algorithms, variational methods that require numerical optimization can also profit, as numerical differentiation can be easily implemented in **mpdcm** (but see Sengupta et al., 2014). Moreover, while we stressed here that **mpdcm** can generate simulations from models with the same input and different parametrizations, it also supports generating simulations from different models *and* different inputs. This is important, as it facilitates inference in models that incorporate several data sets, such as hierarchical multi-subject models (Raman et al. unpublished results).

We have presented two sampling algorithms that exploit the massive parallelism allowed by **mpdcm.** Other sampling methods that make use of parallel simulations have been proposed. For example, multiple try MCMC (Liu et al., 2000) relies on proposing several samples instead of a single one, as usually done in the Metropolis–Hastings algorithm. More recently, Calderhead (2014) proposed an algorithm similar in spirit that extends the acceptance step by using a finite state Markov chain. The algorithms presented here have the advantage of allowing the estimation of the model evidence, and, in the case of parallel tempering, improving the mixing of all the chains. It is important to note that this type of algorithms can be potentially combined, in order to fully exploit the gains in parallel, simultaneous evaluations of the model. Thus, parallel simulations can be added to increase the statistical efficiency of an algorithm at a minimal cost.

Several authors have proposed before to use GPUs in the context of MCMC methods (for a review see Guo, 2012; Suchard et al., 2010). For example, Lee et al. (2010a) considered the improvement in mixing rate in Gaussian mixture models using parallel tempering and found a large increase in statistical efficiency. Jacob et al. (2011) proposed to use parallelism to increase the efficiency of an independent Metropolis–Hastings algorithm. In this variant of MCMC, the updates proposed do not depend on the former step of the chain. Thus, candidate updates can be precomputed in parallel. More generally, the increased computational power provided by GPU has opened the door to Bayesian methods based on sampling that were unfeasible before.

In summary, we envisage that future applications of DCM will require increasingly large computational resources, for example, when the likelihood landscape of a model is multimodal, or when implementing sampling methods for multi-subject analysis. Due to the wide availability of GPUs, algorithmic advances can be readily available in the absence of access to large computer grids, or when there is a lack of expertise to implement algorithms in such grids.

Software note

The **mpdcm** toolbox is available under the GPL license as part of the open source TAPAS software at www.translationalneuromodeling.org/software.

Acknowledgments

K.E.S. is supported by the René and Susanne Braginsky Foundation. K.E.S. and S.R. are supported by the Clinical Research Priority Program "Multiple Sclerosis" and "Molecular Imaging" at the University of Zurich. W.P. is supported by a core grant from the Wellcome Trust (091 593/Z/10/Z). B.S. is supported by the Wellcome Trust (088130/Z/09/Z).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jneumeth.2015.09.009.

References

Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics 2004;20(Feb (3)):407–15.
Blum EK. A modification of the Runge–Kutta fourth-order method. Math Comput 1962;16(78):176–87.
Butcher JC. Numerical methods for ordinary differential equations. 2nd ed. Chichester: John Wiley & Sons Ltd; 2008.
Buxton RB, Wong EC, Frank LR. Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. Magn Reson Med 1998;39(6):855–64.
Bogacki P, Shampine LF. A 3 (2) pair of Runge–Kutta formulas. Appl Math Lett 1989;2(4):321–5.
Calderhead B. A general construction for parallelizing Metropolis–Hastings algorithms. Proc Natl Acad Sci USA 2014;111(Dec (49)):17408–13.
Calderhead B, Girolami M. Estimating Bayes factors via thermodynamic integration and population MCMC. Comput Stat Data Anal 2009;53(12):4028–45.
Chumbley JR, Friston KJ, Fearn T, Kiebel SJ. A Metropolis–Hastings algorithm for dynamic causal models. Neuroimage 2007;38(Nov (3)):478–87.
Eklund A, Dufort P, Forsberg D, LaConte SM. Medical image processing on the GPU—past, present and future. Med Image Anal 2013;17(Dec (8)):1073–94.
Eklund A, Dufort P, Villani M, Laconte S. BROCCOLI: software for fast fMRI analysis on many-core CPUs and GPUs. Front Neuroinform 2014;8:8–24.
Friston K, Mattout J, Trujillo-Barreto N, Ashburner J, Penny W. Variational free energy and the Laplace approximation. Neuroimage 2007;34(1):220–34.
Friston KJ. Bayesian estimation of dynamical systems: an application to fMRI. Neuroimage 2002;16(Jun (2)):513–30.
Friston KJ, Harrison L, Penny W. Dynamic causal modelling. Neuroimage 2003;19(Aug (4)):1273–302.
Ferreira da Silva AR. Cudabayesreg: parallel implementation of a Bayesian multilevel model for fmri data analysis. J Stat Softw 2011;44(4):1–24 (10).
Gelman A, Meng X. L. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. Stat Sci 1998;13(2):163–85.
Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. London: Chapman and Hall/CRC; 2003.
Goldberg D. What every computer scientist should know about floating-point arithmetic. ACM Comput Surv (CSUR) 1991;23(1):5–48.
Guo G. Parallel statistical computing for statistical inference. J Stat Theory Pract 2012;6(3):536–65.
Hernandez M, Guerrero GD, Cecilia JM, Garcia JM, Inuggi A, Jbabdi S, et al. Accelerating fibre orientation estimation from diffusion weighted magnetic resonance imaging using GPUs. PLoS ONE 2013;8(4):e61892.

Jacob P, Robert CP, Smith MH. Using parallel computation to improve independent Metropolis–Hastings based estimation. J Comput Graph Stat 2011;20(3):616–35.
Jing Y, Zeng W, Wang N, Ren T, Shi Y, Yin J, et al. GPU-based parallel group ICA for functional magnetic resonance data. Comput Methods Programs Biomed 2015;119(Apr (1)):9–16.
Kass RE, Raftery AE. Bayes factors. J Am Stat Assoc 1995;90(430):773–95.
Ketchenson DI. Runge–Kutta's methods with minimum storage implementations. J Comput Phys 2010;229(5):1763–73.
Kirkwood JG. Statistical mechanics of fluid mixtures. J Chem Phys 1935;3(5):300–13.
Laskey KB, Myers JW. Population Markov chain Monte Carlo. Mach Learn 2003;50(1–2):175–96.
Lee A, Yau C, Giles MB, Doucet A, Holmes CC. On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. J Comput Graph Stat 2010a;Dec (19)(4):769–89.
Lee VW, Kim C, Chhugani J, Deisher M, Kim D, Nguyen AD, et al. Debunking the 100× GPU vs. CPU Myth: an evaluation of throughput computing on CPU and GPU. SIGARCH Comput Archit News 2010b;38(Jun (3)):451–60.
Li Y, Mascagni M, Gorin A. Decentralized replica exchange parallel tempering: an efficient implementation of parallel tempering using MPI and SPRNG. In: Proceedings of international conference on computational science and its applications (ICCSA). Kuala Lumpur: Springer; 2007. p. 507–19.
Liu JS, Liang F, Wong WH. The multiple-try method and local optimization in metropolis sampling. J Am Stat Assoc 2000;95(449):121–34.
MacKay DJC. Information theory, inference, and learning algorithms. Cambridge: Cambridge University Press; 2003.
Ozaki T. A bridge between nonlinear time series models and nonlinear stochastic dynamical systems: a local linearization approach. Stat Sin 1992;2(1):113–35.
Penny WD, Stephan KE, Mechelli A, Friston KJ. Comparing dynamic causal models. Neuroimage 2004;22(Jul (3)):1157–72.
Penny WD, Stephan KE, Daunizeau J, Rosa MJ, Friston KJ, Schoeld TM, et al. Comparing families of dynamic causal models. PLoS Comput Biol 2010;6(3):e1000709.
Raimondo F, Kamienkowski JE, Sigman M, Fernandez Slezak D. Cudaica: Gpu optimization of infomax-ica eeg analysis. Comput Intell Neurosci 2012;2012:2.
Raftery AE, Newton MA, Satagopan JM, Krivitsky PN. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In: Bernardo JM, Bayarri MJ, Berger JO, Dawid AP, Heckerman D, Smith AFM, West M, editors. Bayesian Statistics, 8. Oxford: Oxford University Press; 2007. p. 1–45.
Sengupta B, Friston KJ, Penny WD. Efficient gradient computation for dynamical models. Neuroimage 2014;98(Sep):521–7.
Sengupta B, Friston KJ, Penny WD. Gradient-free MCMC methods for dynamic causal modelling. Neuroimage 2015;112(May):375–81.
Shi L, Liu W, Zhang H, Xie Y, Wang D. A survey of GPU-based medical image computing techniques. Quant Imaging Med Surg 2012;2(Sep (3)):188–206.
S. Raman, L. Deserno, F. Schlagenhauf, K.E. Stephan. A hierarchical model for unifying unsupervised generative embedding and empirical Bayes, unpublished results.
Stephan KE, Weiskopf N, Drysdale PM, Robinson PA, Friston KJ. Comparing hemodynamic models with DCM. Neuroimage 2007;38(Nov (3)):387–401.
Stephan KE, Kasper L, Harrison LM, Daunizeau J, den Ouden HEM, Breakspear M, et al. Nonlinear dynamic causal models for fMRI. Neuroimage 2008;42(Aug):649–62.
Suchard MA, Wang Q, Chan C, Frelinger J, Cron A, West M. *Understanding* GPU programming for statistical computation: studies in massively parallel massive mixtures. J Comput Graph Stat 2010;19(Jun (2)):419–38.
Swendsen R, Wang J. Replica Monte Carlo simulation of spin-glasses. Phys Rev Lett 1986;57(Nov):2607–9.
Wang WJ, Hsieh IF, Chen CC. Accelerating computation of DCM for ERP in MATLAB by external function calls to the GPU. PLoS ONE 2013;8(6), e66599.
Williamson JH. Low-storage Runge–Kutta schemes. J Comput Phys 1980;35(1):48–56.

Chapter 3

In Chapter 2, we presented an implementation of DCM based on GPUs that is intended to support TI through population MCMC. In the present chapter, we now evaluate the merits of TI with respect to other methods to compute the model evidence, including other sampling schemes, as well as the VBL algorithm (Friston et al., 2007). The latter is the standard method used to compute the model evidence and invert DCMs, and it rests on the variational interpretation of Bayesian inference, which we discuss in some detail in this chapter. In addition to provide a numerical comparison between TI and VBL, we explore their theoretical connection from the perspective of statistical physics and thermodynamics, based on the classical presentation of (Jaynes, 1957).

Our main conclusion here is that TI in combination with population MCMC, can provide robust estimates of the model evidence that other state of the art algorithms cannot deliver. Thus, we argue that despite its computational cost, TI should be used when reliable estimates of the model evidence are paramount.

This chapter concludes Part I, in which we explored how to compare generative models, from a theoretical but also engineering perspective. The next Part is devoted to use the methods presented here to answer empirical questions in computational psychiatry through models of eye movement behavior.

Thermodynamic integration for dynamic causal models

Eduardo A. Aponte[1,*], *Sudhir Raman*[1], *Stefan Frässle*[1], *Jakob Heinzle*[1], *Will D. Penny*[2] *and Klaas E. Stephan*[1,2,3]

[1] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Switzerland.

[2] Wellcome Trust Centre for Neuroimaging, University College London, UK.

[3] Max Planck Institute for Metabolism Research, Cologne, Germany.

[*] Corresponding author: Eduardo A. Aponte aponte@biomed.ee.ethz.ch

Abstract

In generative modeling of neuroimaging data, such as dynamic causal modeling (DCM), one typically considers several alternative models, either to determine the most plausible explanation for observed data (Bayesian model selection) or to account for model uncertainty (Bayesian model averaging). Both procedures rest on estimates of the model evidence, a principled trade-off between model accuracy and complexity. In DCM, the log evidence is typically approximated using variational Bayes (VB) under the Laplace approximation (VBL). Although this approach is highly efficient, it makes distributional assumptions and can be vulnerable to local extrema. An alternative to VBL is Markov Chain Monte Carlo (MCMC) sampling, which is asymptotically exact but orders of magnitude slower than VB. This has so far prevented its routine use for DCM. This paper makes three contributions. First, we introduce a powerful MCMC scheme – thermodynamic integration (TI) – to neuroimaging and present a derivation that establishes a theoretical link to VB. Second, we present an implementation of TI for DCM that rests on population MCMC. Third, using simulations and empirical functional magnetic resonance imaging (fMRI) data, we compare log evidence estimates obtained by TI, VBL, and other MCMC-based estimators (prior arithmetic mean and posterior harmonic mean). We find that, in most cases, model comparison based on VBL gives reliable results, justifying its use in standard DCM for fMRI. Furthermore, we demonstrate that even for nonlinear models accurate estimates of the model evidence can be obtained with TI in acceptable computation time. This paves the way for using DCM in scenarios where the robustness of single-subject inference becomes paramount, such as differential diagnosis in clinical applications.

Keywords: model evidence, free energy, population MCMC, DCM, model comparison, fMRI

Introduction

Dynamic causal models (DCMs; Friston et al., 2003, reviewed in Daunizeau et al., 2011) are models for inferring unobservable neurophysiological processes – e.g., the effective connectivity between neuronal populations – from neuroimaging measurements such as functional magnetic resonance imaging (fMRI) or magneto-/electroencephalography (M/EEG) data. DCMs consist of two hierarchically related layers: a set of state equations describing neuronal population activity, and an observation model which links neurophysiological states to observed signals and accounts for measurement noise. Together with a prior distribution over model parameters, a DCM specifies a full generative or probabilistic forward model that can be inverted using Bayesian techniques.

Depending on the specific type of DCM, interactions between neuronal populations and the generation of measurable signals – e.g., blood oxygen level dependent (BOLD) signal in fMRI or scalp voltage fluctuations in EEG – are represented by different nonlinear equations. These nonlinearities prevent exact analytical inference and require the use of approximate or asymptotic inference techniques. To date, variational Bayes under the Laplace approximation (VBL; Friston et al., 2007) has been the method of choice for DCM, partially because of its computational efficiency.

In addition to inference on model parameters, an important scientific problem is the comparison of competing hypotheses that are formalized as different models. Under the Bayesian framework, model comparison is based on the evidence or marginal likelihood of a model. The evidence corresponds to the denominator or normalization constant from Bayes theorem and represents the probability of the observed data under a given model. It is a widely used score of model quality that quantifies the trade-off between model fit and complexity (Stephan et al., 2009).

Unfortunately, in most instances, it is not feasible to derive an analytical expression of the model evidence due to the intractable integrals that arise from the marginalization of the model parameters. Various asymptotical approximations exist, such as the Bayesian Information Criterion (BIC; Schwarz et al., 1978) and more recently the Widely Applicable Bayesian Information Criterion (Watanabe, 2013). Within

the framework of variational Bayes (VB), a lower bound approximation of the log model evidence (LME) is obtained as a byproduct of model inversion: the variational negative free energy (which we refer to as $-F_{VB}$ throughout this paper).

While highly efficient, model comparison based on the variational free energy has several potential pitfalls. For example, under the Laplace approximation as used in the context of DCM, there is no guarantee that $-F_{VB}$ still represents a lower bound of the LME (Wipf and Nagarajan, 2009). Furthermore, VB is commonly performed in combination with a mean field approximation, and the effect of this approximation on the posterior estimates can be difficult to predict (for discussion, see Daunizeau et al., 2011). Finally, in non-linear models, the posterior could become a multimodal density, a condition that makes difficult to apply gradient ascent methods regularly used in combination with the Laplace approximation.

For these reasons, Markov Chain Monte Carlo (MCMC) sampling has been explored as an alternative inference technique for DCM (Aponte et al., 2016; Chumbley et al., 2007; Penny and Sengupta, 2016; Raman et al., 2016; Sengupta et al., 2015; 2016). MCMC is particularly attractive for variants of DCM models in which Gaussian assumptions might be less adequate, such as nonlinear DCMs for fMRI (Stephan et al., 2008), DCMs of electrophysiological data (Moran et al., 2013a), or DCMs for layered fMRI signals (Heinzle et al., 2016). MCMC is also useful when extending DCM to more complex hierarchical models (Raman et al., 2016), in which the derivation of update equations for VB becomes difficult. MCMC does not entail any assumptions about the posterior distribution and is asymptotically exact. However, in practice, its computational cost leads to runtimes that are often prohibitively long for the datasets and models commonly encountered in neuroimaging. Furthermore, in contrast to VB, MCMC-based inversion of generative models does not provide an estimate of the model evidence for free.

While several MCMC strategies for computing the model evidence in neuroimaging applications have been explored (e.g., Friston et al., 2007; Penny, 2012), one particularly powerful and theoretically attractive MCMC variant that has not been investigated in great detail is thermodynamic integration (TI) (although see Penny and Sengupta, 2016). This method, like VB, rests on the concept of free energy and has

been proposed as gold standard for model evidence estimation (Calderhead and Girolami, 2009; Lartillot and Philippe, 2006). Despite strong theoretical advantages, so far, the computational costs of TI have prohibited its practical use in neuroimaging.

This technical note introduces TI to neuroimaging in general and DCM in particular, with three distinct contributions. First, we present the theoretical foundations of TI and demonstrate its theoretical link to VB. Second, we present an efficient implementation of TI that rests on population MCMC and parallelization using graphical processing units (GPUs). Third, we evaluate our TI scheme using simulations and empirical fMRI data. Specifically, we compare LME estimates obtained by TI to those by conventional sampling-based estimates of the LME (prior arithmetic mean and posterior harmonic mean) and VBL.

This paper is organized as follows: We begin with a brief review of DCM for fMRI to keep the paper self-contained. We then turn to model comparison, first reviewing conventional sampling-based estimators of the LME and subsequently presenting a derivation of TI that reveals its theoretical relation to VB. Using simulated data, we verify the accuracy of our TI implementation and its superiority compared to conventional sampling-based estimators of the LME. Finally, we compare the ranking of competing DCMs by TI and VBL for two empirical fMRI datasets – one frequently used dataset (Buchel and Friston, 1997) with nonlinear DCMs that may pose a particular challenge for VBL, and a more recent dataset (Frassle et al., 2016a).

Methods

Dynamic Causal Models

This paper focuses on fMRI data, and we therefore limit our discussion of DCM to the generation of BOLD signals (Friston et al., 2003; Stephan et al., 2008; 2007). In brief, DCM for fMRI is characterized by two layers: first, a set of ordinary differential equations that model the dynamics of interconnected neuronal states x and local hemodynamic states h. Second, the hemodynamic states are entered into a static nonlinear observation equation that relates venous blood volume and deoxyhemoglobin content to measured BOLD signal changes. In the following, we discuss only the most relevant equations, in order to

convey an understanding of the type of problem that model inversion in DCM implies.

The general form of the dynamics of the neuronal layer is

$$\frac{dx}{dt} = f(x, u, \theta_c), \tag{1}$$

where $x = (x_1, \dots, x_N)$ describes the neuronal states of N regions, $u = (u_1, \dots, u_M)$ represents the time series of M experimental manipulations or inputs, and θ_c the connectivity parameters that determine the neuronal dynamics. Using a second order Taylor expansion (see Stephan et al., 2008 for details), the dynamics f can be approximated as:

$$\frac{dx}{dt} = Ax + \sum_{j=1}^{\mathrm{M}} u_j B_j x + Cu + \sum_{i=1}^{\mathrm{N}} x_i D_i x\,. \tag{2}$$

The connectivity parameters θ_c can be divided into four subsets: The $N \times N$ matrix A describes endogenous connectivity strengths between regions. The set of $N \times N$ matrices $B = \{B_1, \dots, B_M\}$ encodes modulatory effects of inputs on connections between regions. The $N \times M$ matrix C describes the direct effects of driving inputs on regions. Finally, the $N \times N$ matrices $D = \{D_1, \dots, D_N\}$ denote second-order interactions between two regions that affect a third one. Linear DCMs use A and C matrices, bilinear DCMs contain at least one non-zero B matrix, and nonlinear DCMs contain at least one non-zero D matrix. Together $\theta_c = \{A, B, C, D\}$ fully describe the dynamics of the neuronal layer.

The hemodynamic model of DCM originates from the Balloon model proposed by Buxton et al. (1998) and extended by Friston et al. (2000) and Stephan et al. (2007). In brief, it describes how changes in neuronal states alter cerebral blood flow, which, in turn, affects venous blood volume and deoxyhemoglobin content. The model consists of a cascade of deterministic differential equations:

$$\frac{dh}{dt} = l(h, x, \theta_h), \tag{3}$$

where $h = (h_1, \dots, h_N)$ denotes four hemodynamic states used to model N regions. Detailed equations and the meaning of the hemodynamic parameters θ_h can be found in Stephan et al. (2007). It is worth noting that the hemodynamic equations are nonlinear and are solved by numerical approximations.

Finally, hemodynamic states enter a static nonlinear observation equation g with parameters θ_g that models the BOLD signal y:

$$y = g(h, \theta_g) + X_0\beta + \varepsilon. \quad (4)$$

The term X_0 is a matrix of confound regressors that accounts for constant terms and low frequency fluctuations. The Gaussian observation noise ε is characterized by the covariance matrix Π_ϵ^{-1}:

$$\varepsilon \sim N(0, \Pi_\epsilon^{-1}). \quad (5)$$

The precision matrix Π_ϵ is represented as a linear combination $\Pi_\epsilon = \sum_r \exp(\lambda_r)\, Q_r$. The precision components Q_r are positive definite and serve to account for regional differences in noise variance and temporal autocorrelation (Friston et al., 2003). Here, we assume that the time series have been whitened and therefore only account for region-specific variances. In this case, each Q_r is a diagonal matrix with diagonal elements belonging to region r set to 1, and 0 elsewhere.

To complete the generative model, the prior distributions of the parameters $\Theta = (\theta_c, \theta_h, \theta_g, \beta)$ and hyperparameters Λ need to be specified. Here, the priors have been largely matched to SPM8 release 5236 (http://www.fil.ion.ucl.ac.uk/spm), except for the scaling of the prior variance of the coefficients of the confound matrix X_0, which was adapted to the scaling of the data as explained in Appendix A. All parameters' prior distributions are Gaussian, and when positivity needs to be enforced, an adequate transformation function is used. A detailed specification of the priors is provided by Sup. Table 1.

Bayesian model comparison and selection

Bayesian inference involves the specification of a probabilistic or generative model m with data y and parameters θ. The model has two components: the prior density over θ, $p(\theta|m)$, and the likelihood function $p(y|\theta, m)$. These are combined to form the posterior distribution using Bayes' theorem. Conditioning on a model m, the posterior distribution is:

$$p(\theta|y, m) = \frac{p(y|\theta, m)p(\theta|m)}{p(y|m)}, \quad (6)$$

$$p(y|m) = \int p(y|\theta, m)p(\theta|m)d\theta. \quad (7)$$

The normalization constant in the denominator, $p(y|m)$, is known as the marginal likelihood or model evidence and corresponds to the likelihood of the data after marginalizing out the parameters of the model.

In practice, given the monotonicity of the logarithmic function, either the evidence or its logarithm can be used to score a set of candidate models $m_1, \ldots, m_n$ (Bayesian model comparison) and to identify the best model within the model space studied (Bayesian model selection; BMS). One common metric for assessing the relative goodness of two models is the Bayes factor (Kass and Raftery, 1995a):

$$B_{i,j} = \frac{p(y \mid m_i)}{p(y \mid m_j)}. \tag{8}$$

or, equivalently, the exponential of the difference in LME of two models.

BMS has gained an important role in neuroimaging, not only for DCM but also in other contexts requiring model comparison, such as EEG source reconstruction (Henson et al., 2009; Wipf and Nagarajan, 2009), or computational neuroimaging (Friston and Dolan, 2010; Stephan et al., 2015). Group-level BMS techniques exist which account for individual heterogeneity by treating the model as a random variable in the population (Rigoux et al., 2014; Stephan et al., 2009). Finally, Bayesian model averaging allows one to compute an average posterior over models (Penny et al., 2010), weighted by the posterior probability of each model. Critically, these approaches rely on an accurate estimate of each model's evidence.

As mentioned above, except for some special cases, the model evidence cannot be determined analytically, and one typically has to resort to approximations. One computationally efficient option is VB (for textbooks, see Koller and Friedman, 2009; MacKay, 2003), which provides a lower bound of the LME. An alternative, which we explore in detail here, is MCMC sampling. This family of methods is characterized by simulating a Markov process whose stationary distribution corresponds to the posterior distribution $p(\theta|y,m)$ (for a textbook reference, see Robert and Casella, 2013).

In the next sections, we describe standard sampling-based estimators of the model evidence and juxtapose these to the variational negative free energy approximation in VBL. We do not consider classical approximations to the LME such as BIC, since these have already been

evaluated in the context of DCM in previous work and were found to be less useful than the variational negative free energy (Penny, 2012).

Prior arithmetic mean estimator (AME)

Importance sampling is a Monte Carlo method for approximating the expected value of a random variable $h(\mathrm{X})$ under the density p by means of an auxiliary density function $w(X)$, which is required to be absolutely continuous with respect to p (Robert and Casella, 2013; Def. 3.9), or less formally, the auxiliary density w should share the same support as p to avoid zeros in the denominator:

$$\int h(x)p(x)dx = \int \frac{h(x)p(x)w(x)}{w(x)}\,dx. \tag{9}$$

From the strong law of large numbers, if this expected value exists, the process

$$\lim_{\mathrm{S}\to\infty} \frac{1}{\mathrm{S}} \sum_{i}^{S} h(x_i)\frac{p(x_i)}{w(x_i)}, \tag{10}$$

converges almost surely to Eq. 9 when the samples $x_1, \ldots, x_n$ have been drawn from the auxiliary distribution w.

In order to approximate the model evidence by importance sampling, the simplest choice of the auxiliary density is the prior distribution, $w = p(\theta \mid m)$. This results in the prior arithmetic mean estimator (AME):

$$p(y|m) = \int p(y|\theta, m)p(\theta|m)d\theta, \tag{11}$$

$$= \int p(y|\theta, m)p(\theta|m)\frac{p(\theta|m)}{p(\theta|m)}d\theta\,, \tag{12}$$

$$p_{AME} = \frac{1}{S}\sum_{i=1}^{S} p(y|\theta_i, m)\,. \tag{13}$$

where samples θ_i have been obtained from the prior distribution $p(\theta|m)$. Because samples of the likelihood $p(y|\theta, m)$ can greatly exceed the range of double precision floating point numbers, it is necessary to normalize the likelihood function in log space. In this case, Eq. 13 can be computed using the formula:

$$\ln p_{AME} = \ln \alpha - \ln S + \ln \sum_{i=1}^{S} \exp[\ln p(y|\theta_i\,, m) - \ln \alpha], \tag{14}$$

where $\alpha > 0$ is an arbitrary constant. In all analyses reported here, α was set to $\max_i p(y|\theta_i, m)$.

A serious shortcoming of AME is that in the great majority of situations most samples drawn from the prior have very low likelihood. Therefore, an extremely large number of samples is required to ensure that high likelihood regions of the parameter space are taken into account by the estimator; otherwise, the model evidence will be underestimated.

Posterior harmonic mean estimator (HME)

The second choice for the auxiliary density is the posterior distribution, which results in the posterior harmonic mean estimator (HME). This estimator has received divergent appraisals in the literature as a method for computing the LME (for example Kass and Raftery, 1995a, Wolpert and Schmidler, 2012). Re-expressing the model evidence, the HME can be derived as follows:

$$\begin{aligned} \frac{1}{p(y|m)} &= \int \frac{p(\theta|m)}{p(y|m)} d\theta, \\ &= \int \frac{p(y|\theta,m)p(\theta|m)}{p(y|\theta,m)p(y|m)} d\theta\,, \\ &= \int \frac{p(\theta|y,m)}{p(y|\theta,m)} d\theta, \end{aligned} \tag{15}$$

$$p_{HME} = \left(\frac{1}{S} \sum_{i=1}^{S} \frac{1}{p(y|\theta_i, m)} \right)^{-1}. \tag{16}$$

Here, samples θ_i have been drawn from the posterior distribution $p(\theta|y,m)$.

In order to avoid numerical instabilities, it is again necessary to normalize in log space, such that Eq. 16 can be computed using the formula

$$\ln p_{HME} = \ln S + \ln \alpha - \ln \sum_{i=1}^{S} \exp[-\ln p(y|\theta_i, m) + \ln \alpha]\,. \tag{17}$$

Here, $\ln \alpha$ was chosen to be the maximum of $-\ln p(y|\theta_i, m)$.

A disadvantage of HME is that its variance might be infinite when the likelihood function is not heavy-tailed (Raftery et al., 2006), which has serious consequences for the convergence rate of a wide variety of

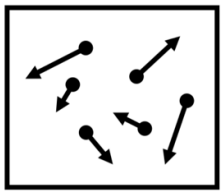
models (Wolpert and Schmidler, 2012). A second problem is that the samples used for HME are obtained from the posterior distribution only. This leads to the opposite behavior as for AME: because the contribution of the prior to the LME might not be appropriately accounted for, the HME tends to overestimate the model evidence, a behavior that can be difficult to diagnose (Lartillot and Philippe, 2006). Several improvements of the HME have been proposed to account for this shortcoming (for example, Raftery et al., 2006).

Thermodynamic Integration (TI)

This section introduces TI from a statistical physics perspective, focusing on free energy as a concept that has a direct relation to the log model evidence in statistics. This provides a link to the variational Bayes approach to approximate the log evidence which is used in DCM and which is described in the following section.

Free energy: A perspective from physics

In thermodynamics, the analog of the model evidence is the so-called partition function Z of a system that consists of an ensemble of particles in thermal equilibrium. A classical discussion of the relationships presented here can be found in Jaynes (1957) and a more modern perspective in Ortega and Braun (2013). For example, the system can consist of an ideal monoatomic gas, in which the energy of a single particle is a function of its kinetic energy (see Fig. 1). If the system is large enough, the energy of a single particle can be treated as a continuous random variable. The total energy of such a system can then be approximated as the product of the number of particles times the probability of a particle being in a certain energy level, times the energy associated with that level.



Kinetic energy: $\phi(\theta) = \frac{|\theta|^2 m}{2}$

Internal energy: $U = \int q(\theta)\phi(\theta)d\theta$

Shannon entropy: $S = \int q(\theta)\ln q(\theta)d\theta$

Partition function: $Z = \int \exp\left(-\frac{|\theta|^2 m}{2kT}\right)d\theta$

Equilibrium dist.: $q(\theta) = \frac{1}{Z}\exp\left(-\frac{|\theta|^2 m}{2kT}\right)$

Figure 1:. In a monoatomic ideal gas at temperature T, the energy of a single particle is a function of its velocity θ and mass m. We assume that there is a large number of particles, such that their velocity can be treated as a continuous random variable. The total energy of the system is the expected energy of a single particle times the number of particles. It is possible to show that once the system settles in equilibrium, the probability that a particle attains a certain velocity is Gaussian distribution. Note that at higher temperatures, the probability of higher kinetic energies increases, leading to higher uncertainty about it, i.e., higher Shannon entropy.

The internal energy U corresponds to the expected energy of the system. This is given by the potential energy ϕ associated with each possible configuration θ that the system can occupy, times the probability q that the system occupies that state

$$U \stackrel{\text{def}}{=} \int q(\theta)\phi(\theta)d\theta\,. \tag{18}$$

The Shannon entropy S of q is:

$$S \stackrel{\text{def}}{=} -\int q(\theta)\ln q(\theta)\,d\theta\,. \tag{19}$$

If the system is not perturbed by an external force, the third law of thermodynamics prescribes that its entropy can only increase or stay constant. Thus, the system is at equilibrium when the associated Shannon entropy is maximized, subject to the constraint that the system's internal energy is constant and equal to U, and

$$\int q(\theta)d\theta = 1. \tag{20}$$

The maximization problem under the above-mentioned conditions can be solved through a variational Lagrangian with two constrains (represented by the Lagrange multipliers λ_1 and λ_2):

$$\frac{\delta}{\delta q}[-\int q(\theta)\ln q(\theta)\,d\theta - \lambda_1(\int q(\theta)d\theta - 1) - \lambda_2(\int q(\theta)\phi(\theta)d\theta - U)] = 0. \quad (21)$$

Noting that

$$-\frac{\delta}{\delta q}\int q(\theta)\ln q(\theta)\,d\theta = -1 - \ln q(\theta)\,, \quad (22)$$

$$-\frac{\delta}{\delta q}\lambda_1\left(\int q(\theta)d\theta - 1\right) = -\lambda_1 \quad (23)$$

$$-\frac{\delta}{\delta q}\lambda_2\left(\int q(\theta)\phi(\theta)d\theta - U\right) = -\lambda_2\phi(\theta). \quad (24)$$

Solving the Lagrangian yields

$$\ln q(\theta) = -\lambda_1 - \lambda_2\phi(\theta) - 1, \quad (25)$$

$$q(\theta) = \frac{1}{\exp(\lambda_1 + 1)}\exp -\lambda_2\phi(\theta)\,. \quad (26)$$

The term $\exp(\lambda_1 + 1)$ is the normalization constant of $q(\theta)$. The term λ_2 is an energy/information conversion factor, and constitutes the statistical physics' definition of temperature (Jaynes, 1957, S. J. Blundell and K. M. Blundell, 2009 section 4.4):

$$\frac{1}{kT} \stackrel{\text{def}}{=} \lambda_2, \quad (27)$$

where, k is the Boltzmann constant. This term is more commonly represented with the symbol β, and here we have used the symbol λ to highlight its derivation as a Lagrange multiplier (Ortega and Braun, 2013). We can write:

$$q(\theta) = \frac{1}{Z}\exp\left(-\frac{\phi(\theta)}{kT}\right), \quad (28)$$

$$Z \stackrel{\text{def}}{=} \int \exp\left(-\frac{\phi(\theta)}{kT}\right)d\theta\,, \quad (29)$$

where Z is referred as the partition function of the system.

In a closed system, F_H is called the Helmholtz free energy, which is defined as the difference between the internal energy U of the system and its Shannon entropy S times the information-energy gain kT:

$$F_H \stackrel{\text{def}}{=} U - kTS. \quad (30)$$

Substituting the definition of the internal energy and Shannon entropy (Eq. 18 and 19), as well as the definition of q (Eq. 27) into Eq. 30, the Helmholtz free energy is

$$F_H = \int q(\theta)\phi(\theta)d\theta + kT \int q(\theta) \ln q(\theta) \; d\theta \,. \tag{31}$$

$$= \int \frac{1}{Z} \exp\left(-\frac{\phi(\theta)}{kT}\right)\phi(\theta)d\theta + kT \int \frac{1}{Z} \exp\left(-\frac{\phi(\theta)}{kT}\right)\left(-\frac{\phi(\theta)}{kT} - \ln Z\right)d\theta, \tag{32}$$

$$= -kT \ln Z. \tag{33}$$

It readily follows that

$$-\frac{F_H}{kT} = \ln Z \,. \tag{34}$$

Free energy: A perspective from statistics

The physical concept of free energy described above plays an important role in statistics. More specifically, the statistical concept of free energy that is analogous to the physical concept arises when considering the negative log joint probability $-\ln p(y, \theta|m)$ as proportional to the energy function (compare Neal and Hinton, 1998):

$$\phi(\theta) = -kT \ln p(y|\theta, m)p(\theta|m) = -kT \ln p(y, \theta|m) \,. \tag{35}$$

Inserting the expression for ϕ in Eq. 34 into Eq. 28, one can show that the equilibrium distribution of the system is the posterior distribution. Specifically, the free energy can be written as

$$F_H = -kT \int q(\theta) \ln p(y, \theta|m) \, d\theta + kT \int q(\theta) \ln q(\theta) \, d\theta, \tag{36}$$

$$\frac{-F_H}{kT} = \int q(\theta) \ln p(y, \theta|m) \, d\theta - \int q(\theta) \ln q(\theta) \, d\theta, \tag{37}$$

$$q(\theta) = \frac{1}{Z} \exp\left(-\frac{\phi}{kT}\right) = \frac{1}{Z} \exp(\ln p(y, \theta|m) \,. \tag{38}$$

By setting the term kT to unity in Eq. 37 one can move from a physical perspective of free energy (expressed in Joules) to a statistical formulation (expressed in information units of bytes in natural base) . This is the common convention in the statistical literature, and thereby, all quantities become unit-less information theoretic terms. In the following, we will follow this convention and, for notational consistency, explicitly show the sign of the negative free energy.

Under the choice of the energy function in Eq. 35, the partition function corresponds to the normalization constant of the joint probability

$p(y, \theta|m)$. From this, it follows that the negative free energy is equal to the LME:

$$-F_H = \ln p(y|m)\,. \tag{39}$$

For generality, we will treat the free energy as a functional of a possibly un-normalized log density, such that

$$-F_H[\phi] = \ln \int \exp -\phi(\theta)\, d\theta\ . \tag{40}$$

Thermodynamic integration

Thermodynamic integration (TI; Kirkwood, 1935, more recently see Gelman and Meng, 1998) was initially introduced to compute the differences in free energy between two states of a system by constructing a smooth transition between them. Here, we focus on the transition between energy states corresponding to the log prior density $\phi_0(\theta) = -\ln\, p(\theta|m)$ and the log posterior $\phi(\theta) = -\ln p(y|\theta, m) - \ln p(\theta|m)$. Under the assumption that $\int p(\theta|m)d\theta = 1$, and substituting ϕ_0 and ϕ into Eq. 39, it follows that

$$F_H[\phi] - F_H[\phi_0] = -\ln \int p(y|\theta, m)p(\,\theta|m)d\theta\ - \ln \int p(\theta|m)d\theta\,. \tag{41}$$

The goal is to compute the difference in free energy by integrating along a piecewise differentiable path connecting prior and posterior. A transition between $F[\phi]$ and $F[\phi_0]$ can be constructed by the power posteriors ϕ_t:

$$\phi_t(\theta) = -\ln p(y|\theta, m)^t - \ln p(\theta|m)\,, \tag{42}$$

with $t \in [0,1]$, such that $\phi_1 = \phi$. Under mild assumptions, it can be shown that:

$$-\ln p(y|m) = F_H[\phi] - F_H[\phi_0], \tag{43}$$

$$= \int_{t=0}^{t=1} \frac{\partial}{\partial t} F_H[\phi_t]dt, \tag{44}$$

$$= \int_{t=0}^{t=1} \frac{\partial}{\partial t} \ln \int p(y|\theta, m)^t p(\theta|m)\, d\theta\, dt, \tag{45}$$

$$= \int_{t=0}^{t=1} \int \frac{p(y|\theta, m)^t p(\theta|m)}{Z_t} \ln p(y|\theta, m)\, d\theta dt, \tag{46}$$

$$= \int_{t=0}^{t=1} \mathrm{E}[\ln p(y|\theta,m)]_{p(\theta|y,t,m)}\, dt, \tag{47}$$

which we refer to here as the fundamental TI equation. In Eq. 47 we have used the notation

$$Z_t = \int p(y|\theta,m)^t p(\theta|m) d\theta, \tag{48}$$

$$p(\theta|y,t,m) = \frac{p(y|\theta,m)^t p(\theta|m)}{Z_t}. \tag{49}$$

In practice, the expected value $\mathrm{E}[\ln p(y|\theta,m)]_{p(\theta|y,t,m)}$ can then be estimated by sampling from the power posterior

$$\lim_{S\to\infty} \frac{1}{S}\sum_{i=1}^{S} \ln p(y|\theta_i,m) = \mathrm{E}[\ln p(y|\theta,m)]_{p(\theta|y,t,m)}, \tag{50}$$

where samples θ_i have been drawn from $p(\theta|y,t,m)$. The integral over t in Eq. 47 can be computed through, for example, a quadrature rule using a predefined set of values for t $0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = 1$. This yields

$$\frac{1}{2}\sum_{i=1}^{N-1} (t_{i+1} - t_i)\left(\mathrm{E}[\ln p(y|\theta,m)]_{p(\theta|y,t_{i+1},m)} - \mathrm{E}[\ln p(y|\theta,m)]_{p(\theta|y,t_i,m)}\right). \tag{51}$$

The optimal schedule in terms of minimal variance of the estimator and minimal error introduced by this discretization in the context of linear models has been outlined by Gelman and Meng (1998) and Calderhead and Girolami (2009).

TI can be seen as a path integral $F(l(0)) = F[\phi_0]$, $F(l(1)) = F[\phi]$, where l is a function such that

$$F_H[\phi] - F_H[\phi_0] = \int_0^1 \frac{\partial F}{\partial l}\frac{dl}{dt} dt = \int_0^1 \frac{\partial F}{\partial l} p(t) dt = \mathrm{E}\left[\frac{\partial F}{\partial l}\right]_{p(t)}. \tag{52}$$

Here, we have assumed that the derivative $\frac{dl}{dt} = p$ is a strictly positive density over the interval [0,1]. Thus, the selection of an optimal schedule is equivalent to the selection of an optimal importance distribution p (Calderhead and Girolami, 2009; Gelman and Meng, 1998). Under this perspective, TI is an expected value over a set of distributions ranging from the prior to the posterior. This is in contrast to AME and HME which

represent the opposite extremes of this spectrum (Gelman and Meng, 1998b; Penny and Sengupta, 2016).

Notably, the TI equation can also be understood in terms of the definition of the free energy by noting that the latter can be written as the sum of an expected log likelihood and a cross entropy term

$$-F_H(t) = t \int p(\theta|y,t) \ln\, p(y|\theta)\, d\theta - \int p(\theta|y,t) \ln \frac{p(\theta|y,t)}{p(\theta)} d\theta\,, \quad (53)$$

$$-F_H(t) = tA(t) - S(t). \quad (54)$$

The term $A(t) = -\partial F_H/\partial t$ is sometimes referred to as the accuracy of the model (for example Stephan et al., 2009), while the second term is the KL divergence between posterior and prior, also referred to as the complexity term. Note that Eq. 54 is typically presented in the literature only for the case of $t = 1$.

The fundamental TI equation (Eq. 47) and Eq. 54 are closely related as represented schematically in Fig. 2. For a given t, the negative free energy can be interpreted as the signed area below the curve $A(t) = -\partial F_H/\partial t$, whereas the term $t \times A(t)$ is the rectangular area below the constant line given by $A(t)$. Comparison with Eq. 53 shows that the area $tA(t) + F_H(t)$ is the KL divergence between prior and corresponding power posterior.

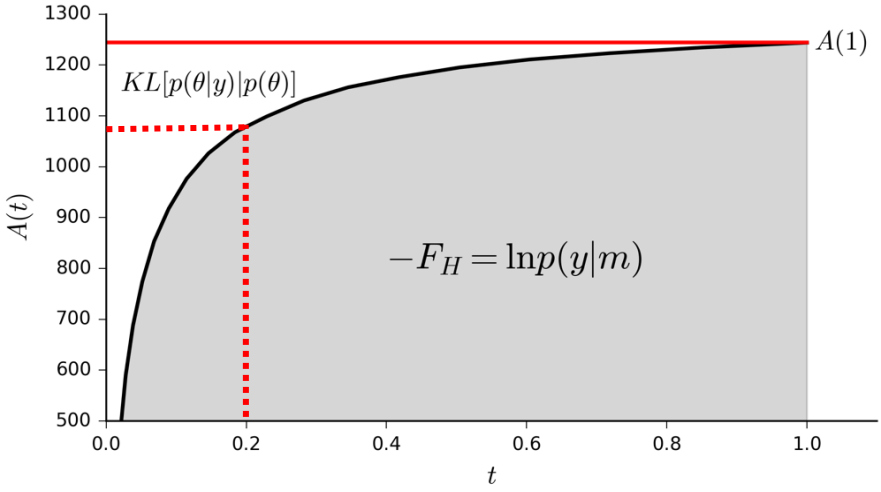


Figure 2:. Graphical representation of the TI equation. The free energy is equal to the *signed* area below $A = \partial F_H/\partial t$, and thus the signed area represented by $A(1) + F_H$ is equal to the KL divergence between posterior and prior. The same relation holds for each $t \in [0,1]$.

This relationship holds because the free energy is a convex function with respect to t, or equivalently, $A(t)$ is a monotonically increasing function, as it can be shown that

$$\frac{\partial A(t)}{\partial t} = \text{Var}[\ln p(y|\theta)]_{p(\theta|y,\,t)} > 0. \tag{55}$$

While from a theoretical point of view TI represents a very appealing choice for computing the model evidence, its practical utility can be limited due to the obvious computational disadvantage of requiring samples not just from a single but from an ensemble of distributions (one for each value of t in the integration of Eq. 51). Arguably, this computational burden has so far prevented routine use of TI in neuroimaging. Below, we present an efficient population MCMC implementation that uses parallelization and GPUs to overcome this bottleneck.

Variational Bayes

Variational Bayes (VB) is a general approach to transform intractable integrals into tractable optimization problems. Importantly, this optimization method simultaneously yields an approximation to the posterior density and a lower bound to the LME.

The fundamental equality which underlies VB is based on introducing a tractable density q to approximate the posterior $p(\theta|y,m)$. This leads to a similar expression of

$$-F_H = \ln p(y|m) = \int q(\theta) \ln p(y|m) \frac{q(\theta)}{q(\theta)} d\theta\,, \tag{56}$$

$$= \int q(\theta) \ln \frac{p(y,\theta|m)q(\theta)}{p(\theta|y,m)q(\theta)} d\theta\,, \tag{57}$$

$$= \underbrace{\int q(\theta) \ln p(y|\theta,m)\, d\theta}_{Approx.\ accuracy} - \underbrace{\int q(\theta) \ln \frac{q(\theta)}{p(\theta|m)} d\theta}_{Approx.\ complexity} + \underbrace{\int q(\theta) \ln \frac{q(\theta)}{p(\theta|y,m)} d\theta}_{Error}\,. \tag{58}$$

The last term in Eq. 58 is the KL divergence or error between the unknown posterior density and the approximate density q. Given that the KL divergence is never negative, the first two terms in Eq. 58 represent a lower bound of the log evidence $-F_H$, and in the following we will refer to it as the variational free energy $-F_{VB}$. Eq. 58 can be rewritten as

$$-F_H - \mathrm{KL}(q(\theta)||p(\theta|y,m)) = -F_{VB} \tag{59}$$

It may seem confusing that the term 'negative free energy' is sometimes used in the literature to denote the logarithm of the partition function Z, as we have done above, and sometimes to refer to the lower bound $-F_{VB}$. This is justified because the variational free energy $-F_{VB}$ becomes identical to the negative free energy $-F_H$ when the approximate density q equals the posterior and hence their KL divergence becomes zero. In this special case

$$\max_{\mathrm{q}} -F_{VB}[q] = -F_H. \tag{60}$$

To maintain consistency in the notation, we will distinguish $-F_H$ and $-F_{VB}$ throughout the paper.

VB aims at reducing the KL divergence between q and the posterior density by maximizing the lower bound $-F_{VB}$ as a functional of q:

$$-F_{VB}[q] = \int q(\theta) \ln p(y|\theta,m)\, d\theta - \int q(\theta) \ln \frac{q(\theta)}{p(\theta|m)} d\theta. \tag{61}$$

Different VB algorithms are defined by the particular functional form used for the approximate posterior. In the next section, we present VB under the Laplace approximation (Friston et al., 2007), as commonly used in DCM.

Note that one can conceptualize TI as selecting the power distributions as the approximate posterior

$$q_t(\theta) = \frac{1}{Z_t} p(y|\theta, m)^t p(\theta|m). \tag{62}$$

However, instead of directly computing the Helmholtz free energy associated with them, the derivatives

$$\frac{\partial F[q_t]}{\partial t}, \tag{63}$$

are estimated using sampling and are used for numerical integration.

Variational Bayes under the Laplace approximation for DCM

Commonly, in order to maximize $-F_{VB}$, a mean field approximation of q is used. In other words, the distribution q is assumed to factorize into different sets of parameters, each of which defines a more tractable optimization problem. In the case of DCM, q is assumed to have the form:

$$q(\Theta, \Lambda) = q(\Theta)q(\Lambda), \tag{64}$$

i.e., the parameters $\Theta = (\theta_c, \theta_h, \theta_g, \beta)$ and the hyperparameters Λ are assumed to be conditionally independent. The functional $-F_{VB}$ can be optimized iteratively with respect to Θ and Λ converging to a maximum $-F_{VB} \leq \ln p(y|m)$ (Koller and Friedman, 2009). This mean field approximation yields two update equations:

$$\ln q(\Theta) = \int q(\Lambda) \ln p(y, \Theta, \Lambda) d\Lambda + c_\Theta, \tag{65}$$

$$\ln q(\Lambda) = \int q(\Theta) \ln p(y, \Theta, \Lambda) d\Theta + c_\Lambda. \tag{66}$$

where c_Θ and c_Λ are constants with respect to Θ and Λ, respectively. Analytical update equations usually require distributional assumptions about q (but see instances of "free-form" VB, e.g. Stephan et al., 2009). In DCM, it is typically assumed that all terms are Gaussian (but see Raman et al., 2016 who used conjugate priors for the noise terms).

Despite the mean field approximation, the integrals in Eq. 65 and 66 cannot be solved analytically because of the nonlinearities of the forward model (Eqs. 1-4). This problem is circumvented by approximating the log of the unnormalized posterior with a second order Taylor expansion on a local maximum, or equivalently, the unnormalized posterior is assumed to be Gaussian. To obtain a maximum, it is necessary to

optimize the objective function $\ln p(y, \Theta, \Lambda)$, usually through a gradient ascent scheme (but see Lomakina et al., 2015 for an alternative based on Gaussian processes). This approach is called the Laplace approximation (Friston et al., 2007) and underlies other methods such as BIC (Schwarz, 1978) or a more common approach where the normalization constant of an approximate, tractable posterior is directly used (Kass and Raftery, 1995b). As a consequence of this approximation, the variational free energy is no longer guaranteed to represent a bound on the log evidence (Wipf and Nagarajan, 2009). In Appendix 2, we present a simplified version of the derivation of the VBL estimate of the free energy in Friston et al. (2007) and an explicit expression for the accuracy term.

Implementation

In this section, we describe the implementation for each of the estimators of the LME described above. Open source code is available in the TAPAS software package (www.translationalneuromodeling.org/software).

MCMC

TI was implemented by obtaining samples from the power posterior distributions $p_i(\theta|y,m) \propto p(y|\theta,m)^{t_i} p(\theta|m)$, with $10^{-5} = t_0 < t_1, \ldots, < t_N = 1$. The temperature schedule obeyed a fifth order power rule as suggested by Calderhead and Girolami (2009). Samples from each of the chains were drawn using the Metropolis-Hastings (MH) algorithm, with a Gaussian kernel as proposal distribution for the connectivity, hemodynamic and forward model parameters $(\theta_c, \theta_h, \theta_g)$. Following Shaby and Wells (2010), the covariance of the proposal distribution was modified during the burn-in phase to resemble the covariance of the posterior distribution. The coefficients of the confound matrix X_0 (see Eq. 4) were sampled using a Gibbs step.

The hyperparameters Λ were sampled using a 'pseudo' Gibbs step, by noting that if the prior was defined to be a Gamma distribution, its conditional posterior is again a Gamma distribution, from which samples can be easily obtained. Thus, one can replace the Gaussian prior of the log precision component by a log normal distribution and approximate it by a Gamma distribution $q(\Lambda)$ with matched moments (Raman et al., 2016), in order to obtain an analytical posterior of the form $q(\Lambda|y,\Theta) \propto p(y|\Theta,\Lambda)p(\Theta)q(\Lambda)$. This last distribution can be used to obtain samples

from Λ. To account for this approximation, a MH step can be used as acceptance criterion for each proposed sample Λ^*. The corresponding ratio is

$$\frac{p(y|\Theta,\Lambda^*)p(\Theta)p(\Lambda^*)q(\Lambda|\theta,y)}{p(y|\Theta,\Lambda)p(\Theta)p(\Lambda)q(\Lambda^*|\theta,y)} = \frac{p(y|\Theta,\Lambda^*)p(\Theta)p(\Lambda^*)p(y|\Theta,\Lambda)p(\Theta)q(\Lambda)}{p(y|\Theta,\Lambda)p(\Theta)p(\Lambda)p(y|\Theta,\Lambda^*)p(\Theta)q(\Lambda^*)}, \quad (67)$$

$$= \frac{p(\Lambda^*)q(\Lambda)}{p(\Lambda)q(\Lambda^*)}. \quad (68)$$

To further enhance the efficiency and convergence properties of the algorithm, we adopted a population MCMC approach in which neighboring chains were allowed to interact by means of a "swap" accept-reject (AR) step (Swendsen and Wang, 1986). In brief, population MCMC defines a joint product distribution

$$\prod_{i=0}^{N} p(\theta_i|y,t_i,m) = \prod_{i=0}^{N} \frac{p(y|\theta_i,m)^{t_i}p(\theta_i|m)}{Z_i}, \quad (69)$$

where N is the number of distributions or chains. The goal is to obtain samples from this distribution by two types of AR steps: First, local steps are used to sample parameters θ_i from $p(\theta_i|y,t_i,m)$. Second, samples are obtained using the swapping step in which a set of neighboring parameters θ_i, θ_{i+1} are randomly chosen and then exchanged between chains with probability:

$$\min(1, (p(y|\theta_{i+1},m)^{t_i}p(\theta_{i+1}|m)/((p(y|\theta_i,m)^{t_{i+1}}p(\theta_i|m))). \quad (70)$$

This AR step does not change the stationary distribution of any of the chains.

Population MCMC can be easily parallelized (Aponte et al., 2016), as each of the chains is independent of the rest of the ensemble. Swapping steps need to be performed serially but, assuming that the likelihood and prior functions have been already evaluated, this method increases the efficiency of the sampling scheme while only inducing negligible computational costs (for example Aponte et al., 2016; Calderhead and Girolami, 2009). Intuitively, the increase in efficiency is achieved by exploring the sampling space in a way comparable to simulated annealing, i.e., allowing some of the chains to explore the parameter space more freely by relaxing the likelihood function.

Since TI requires samples from both the prior and the posterior distribution, the same sampling algorithm can be used for computing all

three sampling-based estimators (TI, AME, HME). This ensures that any observed differences between estimators are not simply due to differences in the implementation of the samplers.

We assessed the convergence of our sampling scheme using the Gelman-Rubin's potential scale reduction factor (PSRF; Gelman and Rubin, 1992) as diagnostic. This method tests parameter-wise convergence by comparing the variance of segments of the chains. A score, or $\hat{R}$ statistic below 1.1 is a commonly accepted criterion for convergence. To compute this score, the samples of the log likelihood of the first and last third section of each chain after the burn-in phase were compared.

Variational Bayes under the Laplace approximation (VBL)

The VBL algorithm used here was the implementation available in the software package SPM8 (release 5236), which employs a gradient ascent scheme to optimize the marginal distributions $q(\Theta)$ and $q(\Lambda)$ (Friston et al., 2007). This algorithm was initialized at the prior mean of the parameters and hyperparameters if not stated otherwise.

To minimize any differences between our sampling-based DCM inversion approach and that of SPM, we used the same 4th order Runge-Kutta scheme for integrating the DCM state equations (as required for evaluating the likelihood) in both TI and VBL.

Integration of the dynamical system

The computationally most intensive part of DCM is the evaluation of its likelihood function, because it requires the integration of the neuronal and hemodynamic state equations in order to predict the BOLD signal given a set of connectivity parameters. Here, we relied on the mpdcm toolbox (Aponte et al., 2016) which parallelizes the evaluation of the likelihood in DCM in two ways: first, the neuronal and hemodynamic states of each region are computed in parallel, i.e., the functions f_i and g_i are simultaneously evaluated region-wise. Second, integration can be performed for several sets of parameter values simultaneously. The integrator used here was the standard 4th order Runge-Kutta explicit method. The numerical accuracy of our implementation was verified in previous work (Aponte et al., 2016).

Computational environment

To provide the reader with an impression of the computation time required by TI using the mpdcm toolbox, we report the computation time of the sampling algorithm for the DCMs investigated here. A more thorough evaluation can be found in Aponte et al. (2016). The computation time depends on the exact hardware used and therefore the values presented here should only be understood as a rough guide. Computation time was measured on a machine with the following specification: Ubuntu 15.04, Linux kernel 2.6.32, MATLAB 8.3.0 and the CUDA toolkit 8.0. The chip set was an Intel i7-4770K and a NVIDIA GTX 1080 graphics card.

Simulations

Linear Models

In order to evaluate the accuracy of AME, HME, and TI for a situation where the ground truth is known, we first compared the estimates from a Bayesian linear regression model whose evidence can be computed analytically. These models are defined with the following prior and likelihood functions:

$$\begin{aligned} p(\theta) &= N\left(\theta; 0, \Pi_p^{-1}\right), \\ p(y \mid \theta) &= N(y; X\theta, \Pi_e^{-1}), \end{aligned} \tag{71}$$

where θ is the $[p \times 1]$ vector of regression coefficients, y is the $[M \times 1]$ vector of data points, X is the $[M \times p]$ design matrix, and Π_p^{-1} and Π_e^{-1} are the covariance matrices of the prior and errors, respectively. The LME is given by

$$\ln \int p(y|\theta)p(\theta)d\theta =$$

$$-0.5\ln|\Pi| - 0.5N\ln 2\pi + 0.5\ln|\Pi_e| + 0.5\ln\left|\Pi_{\mathrm{p}}\right| - 0.5y^T\Pi_e y + 0.5\,\eta^T\Pi\eta \tag{72}$$

$$\Pi = \Pi_p + X^T\Pi_e X, \tag{73}$$

$$\eta = \Pi^{-1}X^T\Pi_e y. \tag{74}$$

For our simulations, we chose $M = 100$, $\Pi_p^{-1} = 16I_p$ and $\Pi_e^{-1} = 10I_e$, where I_p and I_e are the corresponding identity matrices. The design matrix was chosen to have a block structure equivalent to a design for a one-way ANOVA with p levels (for the values of p that do not exactly divide by M, the excess data points were assigned to the last cell).

Synthetic data was generated by sampling from the generative model defined in Eq. 68.

DCM: Simulated data

In the first experiment, we used simulated data from 5 DCMs (linear: model 1; bilinear: models 2 to 4; nonlinear: model 5) with two inputs (u1 and u2). The DCMs are displayed in Fig. 3, and are available for download at https://bitbucket.org/aponteeduardo/aponte_et_al_2016. The parameters were chosen to maximize qualitative differences between the signals generated from them. The BOLD signal data was simulated assuming a repetition time (TR) = 2s and 720 scans per simulation. The driving inputs were entered with a sampling rate of $2.0Hz$, such that each simulation required 2880 integration steps. Simulated time series were corrupted with Gaussian noise yielding a signal-to-noise ratio (SNR) of 1.0. Here, SNR was defined as the ratio of signal standard deviation to noise standard deviation (Welvaert and Rosseel, 2013). This means that our simulated data contained identical amounts of noise and signal, representing a relatively challenging SNR scenario.

We generated 40 different datasets with different instantiations of Gaussian noise, such that the underlying time series remained constant for each model. We then counted how often the data-generating model obtained the highest model evidence and compared the ensuing values across the different estimators (i.e., AME, HME, TI, VBL). Notably, the absolute value of the log evidence of a given model is irrelevant for model scoring; instead, its difference to the log evidence of other models is decisive.

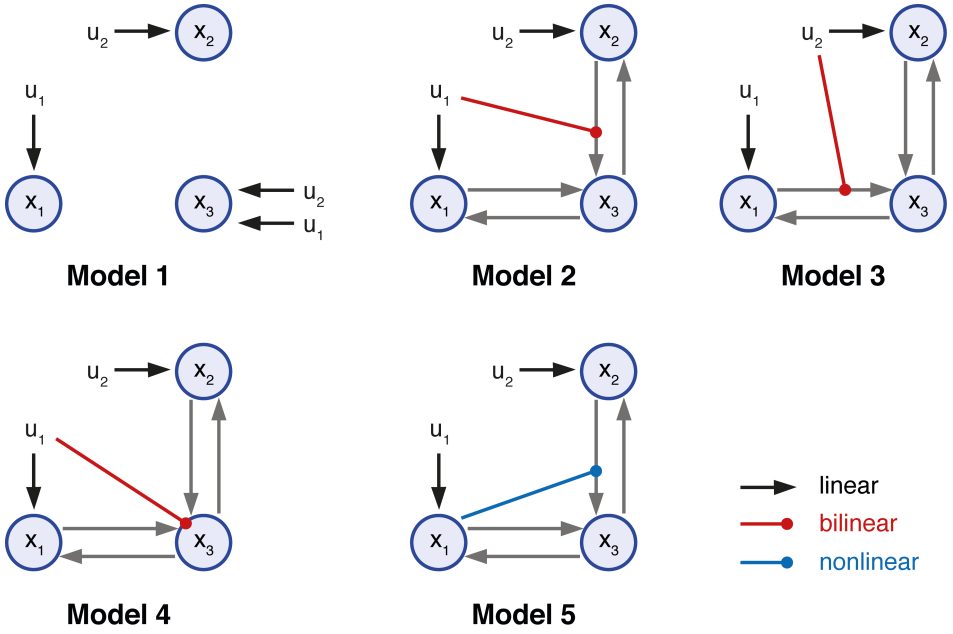


Figure 3: Illustration of the five simulated 3-region DCMs used for cross-model comparison. Self-connections are not displayed. The variables u_1 and u_2 represent two different experimental conditions or inputs. All models represented different hypotheses of how the neuronal dynamics in area x_3 could be explained in terms of the two driving inputs and the effects of the other two regions x_1 and x_2. Model m_1 can be understood as a 'null hypothesis' in which the activity of all the areas can be explained by the driving inputs. Models m_2 and m_3 correspond to two forms of bilinear effect on the forward connection of areas x_1 and x_2. Model m_4 represents the hypothesis that input u_1 affects the self-connection of area x_3 (not displayed). Model m_5 represents a non-linear interaction between regions x_1 and x_2. Endogenous connections are depicted by gray arrows, driving inputs by black arrows, bilinear modulations by red arrows and nonlinear modulations by blue arrows

Empirical data: Attention to motion

In order to compare VBL and TI using empirical data, we used the "attention to motion" fMRI dataset (Buchel and Friston, 1997) that has been analyzed in numerous previous methodological studies (e.g., Friston et al., 2003; Marreiros et al., 2008; Penny et al., 2004a; 2004b; Stephan et al., 2008). In brief, Buchel and Friston (1997) investigated the effect of attention on motion perception; in particular, the authors examined attentional effects on the connectivity between primary visual cortex (V1), motion-sensitive visual area (V5) and posterior parietal cortex (PPC). There were four conditions (all under constant fixation): fixation only (F), presentation of stationary dots (S), passive observation

of radially moving dots (N), or attention to the speed of these dots (A). Four sessions were recorded and concatenated yielding a total of 360 volumes ($T_E = 40ms,\ TR = 3.22s$). Three inputs were constructed using a combination of the three conditions: $stimulus = S + N + A$, $motion = N + A$, $attention = A$. Driving inputs were resampled at $0.8Hz$, requiring a total of 1440 integration steps. Further details of the experimental design and analysis can be found in Buchel and Friston (1997).

One reason for selecting this dataset is that Stephan et al. (2008) previously demonstrated that a nonlinear model had higher evidence than comparable bilinear models (Fig. 4). This case is of interest for evaluating the quality of different LME estimators, as one would expect that the introduction of nonlinearities represents a challenging case for VBL.

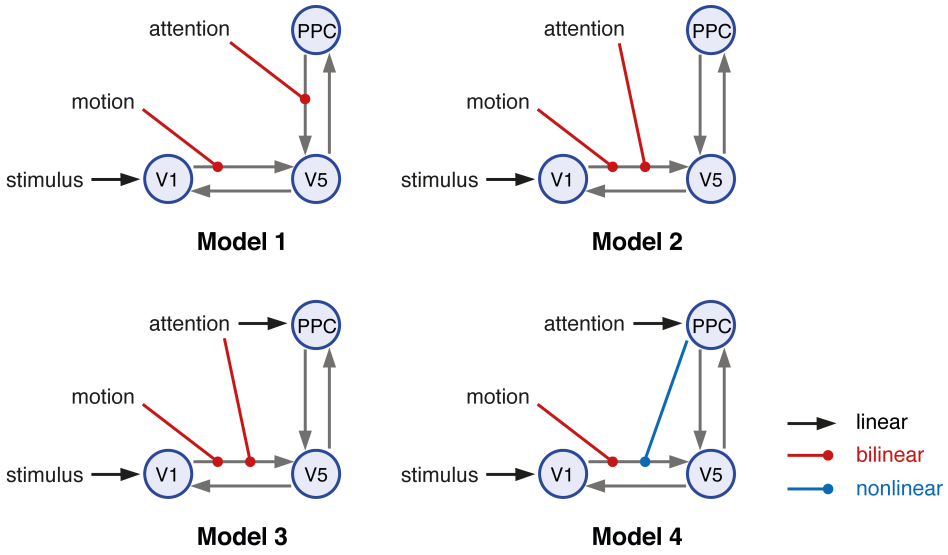


Figure 4: Illustration of the four models used in (Stephan et al., 2008) representing different hypotheses of the putative mechanisms underlying attention-related effects in the motion-sensitive area V5. The first three models are bilinear whereas the fourth model is a nonlinear DCM. Endogenous connections are depicted by gray arrows, driving inputs by black arrows, bilinear modulations by red arrows and nonlinear modulations by blue arrows. Inhibitory self-connections are not displayed. V1: primary visual area, V5 = motion sensitive visual area, PPC: posterior parietal cortex.

Empirical data: Face perception

Additionally, we analyzed fMRI data from a single representative subject participating in a face perception paradigm described in detail in Frassle et al. (2016a). This dataset differs in complexity from Buchel and Friston

(1997) in several ways: it consists of almost three times as many scans (940), and the DCMs contain twice the number of regions (6) and nearly three times more free parameters (shown in Figure 4, models m1, m2 and m_4 possessed 28 connectivity parameters, model m3 contained 36 connectivity parameters), and a much lower SNR. This dataset is a useful candidate for evaluating sampling methods for inversion of DCMs since a previous analysis suggested possible instabilities of the VBL estimates for challenging scenarios where the number of network nodes and free parameters is high (Frassle et al., 2015).

In brief, subjects viewed either faces (F), objects (O), or scrambled (i.e., Fourier-randomized) images (S) in the left (LVF) or right visual field (RVF) in a block design, under central fixation. This study examined hemispheric lateralization in the human brain by probing intra- and interhemispheric integration in the core face perception network. The network comprised bilateral occipital face area (OFA; Puce et al., 1996) and fusiform face area (FFA; Kanwisher et al., 1997), serving as the key regions for face processing (Haxby et al., 2000), as well as left and right primary visual cortex (V1), representing the visual input regions of the network. A total of 940 scans were acquired, with TR = 1450 ms. Here, we tested the four DCMs displayed in Fig. 5. For all DCMs, non-zero entries in the endogenous connectivity (A-matrix) and driving inputs (C-matrix) were identical. Driving inputs u, representing the visual stimulation in either the left or right visual field, entered the contralateral V1 and were sampled at four times the frequency of the TR; thus 3760 integration steps were performed for each simulation. Forward and backward intra-hemispheric endogenous connections were assumed between V1 and OFA, and between OFA and FFA. Furthermore, reciprocal inter-hemispheric connections were set among the homotopic face-sensitive regions. Critically, models differed with regard to the experimental conditions that were allowed to modulate both intra- and interhemispheric connections, implementing different hypotheses of how hemispheric lateralization in the face perception network could arise from the functional integration within and between hemispheres. A comprehensive description of the experimental design and analysis can be found in Frassle et al. (2016b; 2016c).

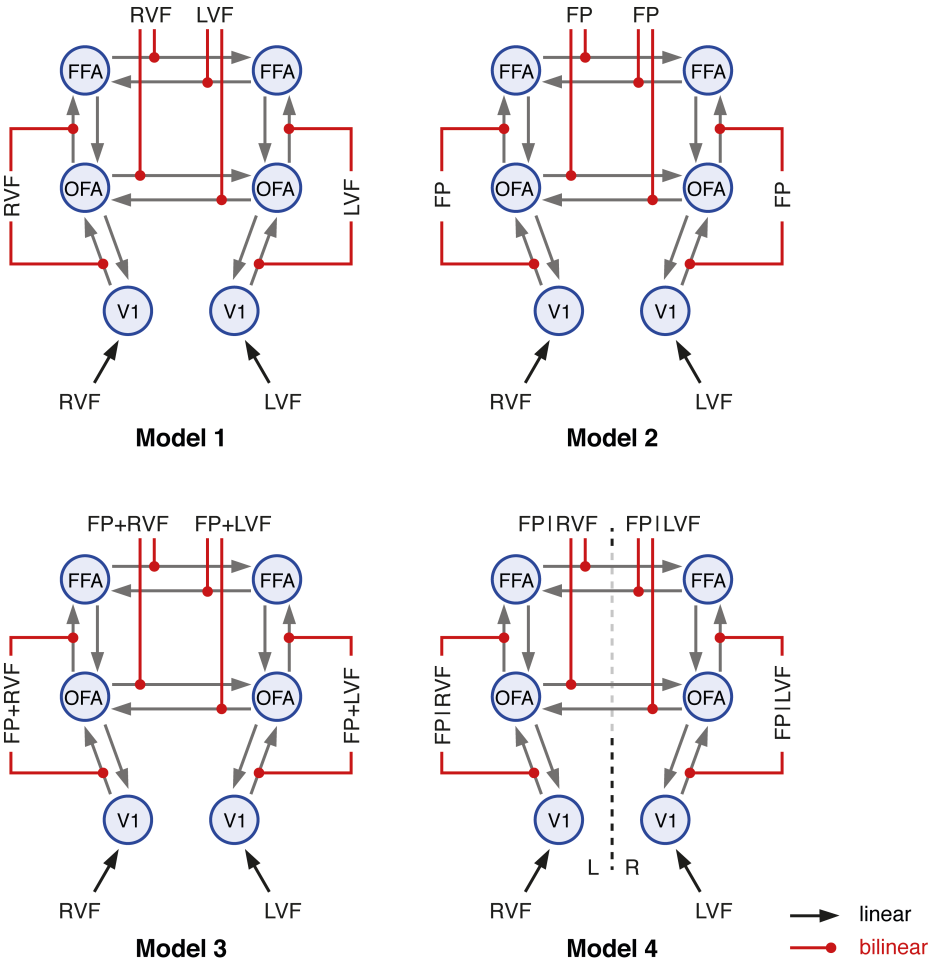


Figure 5: Four different models used in Frässle et al. (2016a, 2016b), representing different hypotheses of the putative mechanisms underlying hemispheric lateralization in the face perception network. Endogenous connections are depicted by gray arrows, driving inputs by black arrows, and bilinear modulations by red arrows. Inhibitory self-connections are not displayed. V1: primary visual area, OFA: occipital face area, FFA: fusiform face area. L: left hemisphere, R: right hemisphere. LVF: left visual field, RVF: right visual field, FP: face perception. FP+RVF: Face perception and right visual field stimulation. FP+LVF: Face perception and left visual field stimulation.

Results

Synthetic data: linear models

In this analysis, we computed the LME of a general linear model (with a varying number p of regressors) using TI, AME and HME, and compared the results against the analytically computed LME.

Varying p from 2 to 32 in steps of 2, we repeated the data generation process 10 times. For each of these runs, values of the regression parameters θ were drawn from the prior, and observations y were generated according to the likelihood. The TI approximation to the model evidence was computed using $J = 64$ chains with a 5th order annealing schedule. We then computed AME based on the samples from the prior density ($t = 0$), and HME based on samples from the posterior ($t = 1$). Fig. 6 shows the error in the LME estimates as a function of the number of model parameters for the three approaches. Consistent with previous reports, we found that HME overestimated the LME, while AME underestimated it Lartillot and Philippe (2006). Only TI provided good estimates over the full range of models. However, for a large number of model parameters, we observed a small bias in that TI estimates began to slightly overestimate the log evidence. This suggests that our TI implementation may require a larger number of chains.

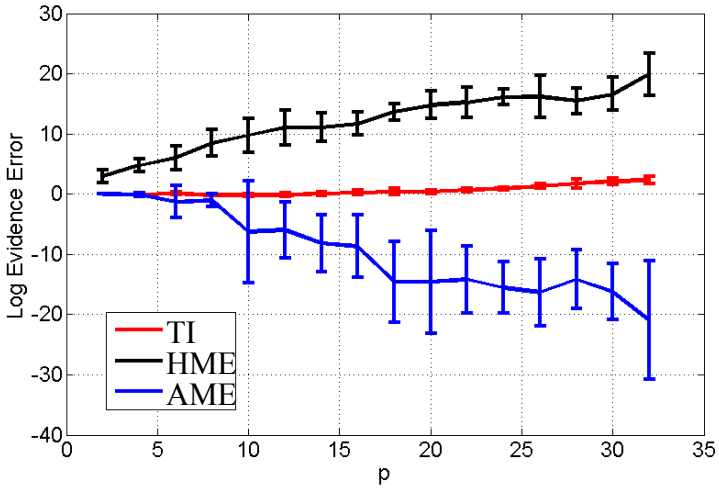


Figure 6: Error in estimating the log evidence of linear models for three different sampling approaches. The curves show mean and standard deviation (error bars) over ten runs at each value of p (number of GLM parameters) for thermodynamic integration (TI), posterior harmonic mean estimator (HME) and prior arithmetic mean estimator (AME).

Synthetic data: DCM

In a pretesting phase, we found that TI generated stable estimates of the LME using 64 chains. All simulations were executed with a burn-in phase of 1×10^4 samples, followed by 1×10^4 kept samples. Exemplary run times of the algorithm are show in Table 1.

Computation time: Synthetic data					
	$\boldsymbol{m_1}$	$\boldsymbol{m_2}$	$\boldsymbol{m_3}$	$\boldsymbol{m_4}$	$\boldsymbol{m_5}$
Samples/second	240.4	200.1	200.0	199.2	186.9

Table 1: Computational time in simulations per second. Each simulation required 2880 integration steps. Simulations were performed on a CPU.

We evaluated the convergence of the MCMC algorithm by examining the samples of the log likelihood of all chains. We found that the $\hat{R}$ statistic was below 1.1 in all but a few instances (less than 0.002% of chains). Estimated log model evidences are displayed in Figs. 7 and 8. Consistent with the linear model analysis in the previous section, the HME was

always higher and the AME always lower than the TI estimate of the LME. VBL estimates were clearly close to the TI estimate.

We then examined how often the data-generating model was identified correctly by model comparison, i.e., how often it showed the largest LME of all models. Of all estimators, AME failed most frequently to detect the data-generating model (Table 2). HME identified the correct model more consistently (Table 3). Both VBL and TI displayed a similar behavior (Tables 4 and 5), although model m_5 was identified slightly more consistently through VBL. However, as displayed in Fig. 8, according to both inversion schemes, the data generating model was consistent with the model with the highest LME.

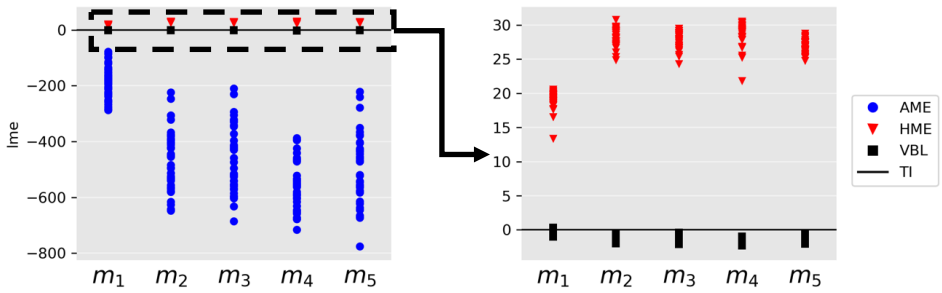


Figure 7: Estimated LME for all models relative to TI when inverted with the corresponding data-generating model under $SNR = 1$ for 40 different models. Right panel zooms into the left panel. Red triangles correspond to the HME, blue circles to the AME, and black squares to VBL. HME was always higher and AME always lower than the TI estimate. All LME estimates are shown after subtracting the TI-based estimate for the same model.

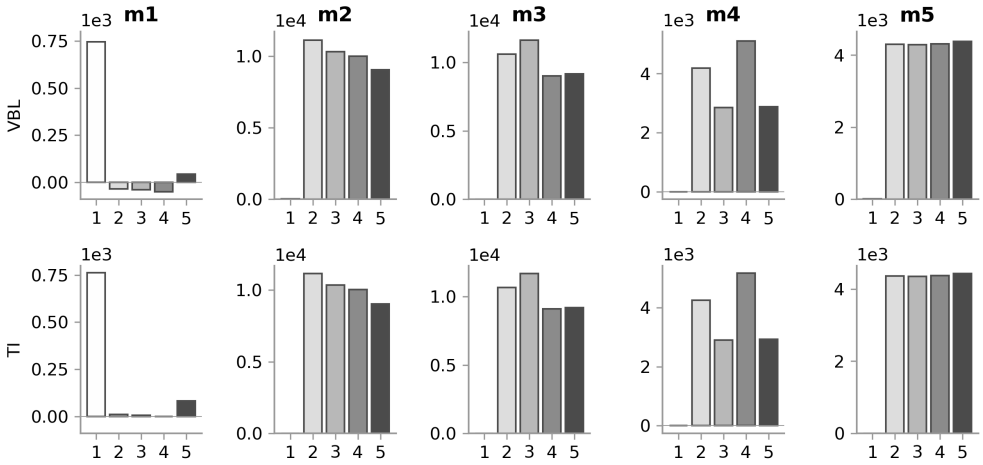


Figure 8: Estimated LME using VBL (Top) and using TI (Bottom). The panels display the LME (summed across 40 simulations) for each model given data generated by one model (indicated above each panel). Results were normalized such that columns (VBL and TI) share the same base line and can be directly compared. VBL estimates were always higher than TI, although qualitatively the results were similar. For both VBL and TI the data-generating model obtained the highest LME.

AME: Synthetic data						
		Generation				
		m_1	m_2	m_3	m_4	m_5
Inversion	m_1	39	14	10	34	29
	m_2	1	13	11		3
	m_3		6	11	3	
	m_4		3	4	2	5
	m_5		4	4	1	3

Table 2: Cross-model comparison results for AME in the case of synthetic data (SNR = 1). The row label indicates the data-generating model, the column index is the inferred model.

HME: Synthetic data						
		Generation				
		m_1	m_2	m_3	m_4	m_5
Inversion	m_1	39				
	m_2		40			12
	m_3			40		12
	m_4				40	4
	m_5	1				12

Table 3: Cross-model comparison results for HME in the case of synthetic data (SNR = 1). The row label indicates the data-generating model, the column index is the inferred model.

VBL: Synthetic data						
		Generation				
		m_1	m_2	m_3	m_4	m_5
Inversion	m_1	40				
	m_2		40			
	m_3			40		
	m_4				40	1
	m_5					39

Table 4: Cross-model comparison results for VBL in the case of synthetic data (SNR = 1). The row label indicates the data-generating model, whereas the column index is the inferred model.

TI: Synthetic data						
		Generation				
		m_1	m_2	m_3	m_4	m_5
Inversion	m_1	40				
	m_2		40			
	m_3			40		
	m_4				40	2
	m_5					38

Table 5: Cross-model comparison results for TI in the case of synthetic data (SNR = 1). The row label indicates the data-generating model, the column index is the inferred model.

Empirical data

Because our previous results clearly demonstrated the inferiority of the AME and HME, in the following, we limit our analysis to TI and VBL.

Attention to Motion

For the attention to motion dataset, 16×10^3 samples were collected from 64 chains, from which 8×10^3 were discarded in the burn-in phase. A summary of the computational times can be found in Table 6. We found that, across all four models, only in one chain the PSRF score of the samples of the log likelihood was above 1.1, indicating the convergence of the algorithm.

Computation time: Attention to motion				
	$\boldsymbol{m_1}$	$\boldsymbol{m_2}$	$\boldsymbol{m_3}$	$\boldsymbol{m_4}$
Simulations/second	1077.4	1084.7	1090.3	1032.3

Table 6: Computational time in terms of simulations per second of the sampler used to investigate the models reported in Stephan et al. (2008). Each simulation required 1440 integration steps. Simulations were performed in GPU.

Table 7 summarizes the evidence estimates obtained with TI and VBL. In comparison to the results reported by Stephan et al. (2009), three findings are worth highlighting. First, the VBL algorithm reproduced the same ranking of models reported in Stephan et al. (2008), although Stephan et al. used an earlier version of the VBL algorithm with different prior parameters and a different integration scheme. Moreover, our TI implementation produced the same ranking as the one obtained under VBL.

Second, the difference between the VBL free energy estimates and the TI estimates varied considerably across models. To investigate this variability, we compared the log likelihood of the sample with the highest posterior probability (i.e., the MAP) obtained using MCMC and the likelihood of the convergence point obtained using VBL, as this term explicitly enters the accuracy term. Results are summarized in the lower section of Table 7. Again, large discrepancies were clear in model m_4 (>40 log units), and this difference was also observable in the accuracy estimates.

Attention to motion dataset				
Log model evidence				
	m_1	m_2	m_3	m_4
VBL	-1790.0	-1778.6	-1776.6	-1774.8
TI	-1772.6	-1761.1	-1757.8	-1729.1
Accuracy				
	m_1	m_2	m_3	m_4
VBL	-1547.6	-1538.5	-1531.6	-1530.7
TI	-1525.6	-1520.2	-1511.8	-1483.5
Log likelihood at MAP				
	m_1	m_2	m_3	m_4
VBL	-1514.3	-1505.3	-1497.9	-1496.6
TI	-1504.6	-1498.4	-1490.6	-1459.4

Table 7: Log model evidence, accuracy and log likelihood at the MAP estimate using both TI and VBL.

m_1	m_2	m_3	m_4
Stephan et al. 2008.			
0.0	3.1	5.6	13.6
VBL			
0.0	11.4	13.4	15.2
TI			
0.0	11.5	14.8	43.5

Table 8: Results of model comparison, in terms of log evidence differences with respect to the worst model (m_1), from Stephan et al (2008)

Third, while VBL detected the most plausible model, the findings from this dataset suggest that VBL-based inversion of DCMs might not always be fully robust. In particular, the difference between the algorithms could be attributed to the VBL algorithm converging to a local extremum. To assess the differences between TI and VBL more systematically, we repeated the simulations 10 times with different starting positions of the inversion scheme, sampled from the prior density. Fig. 9 displays the predicted BOLD signal time series and Fig. 10 the estimated model evidence. VBL estimates of the LME and the predictive fits displayed a much larger variance than the TI estimates. This difference could be traced back to the estimates of the accuracy as shown in Appendix 3.

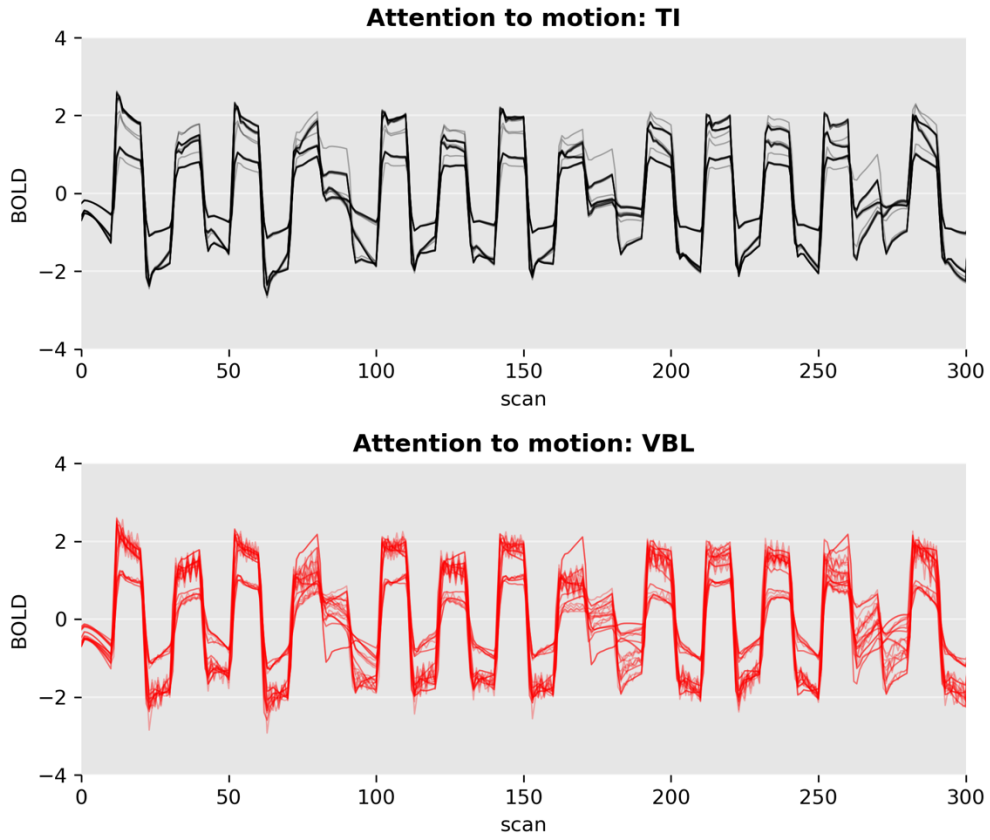


Figure 9. Comparison of 10 fits (MAP) of model m_4 between TI and VBL for the "attention to motion" dataset from Buchel and Friston (1997). Both estimates are qualitatively similar, but VBL fits display higher variability.

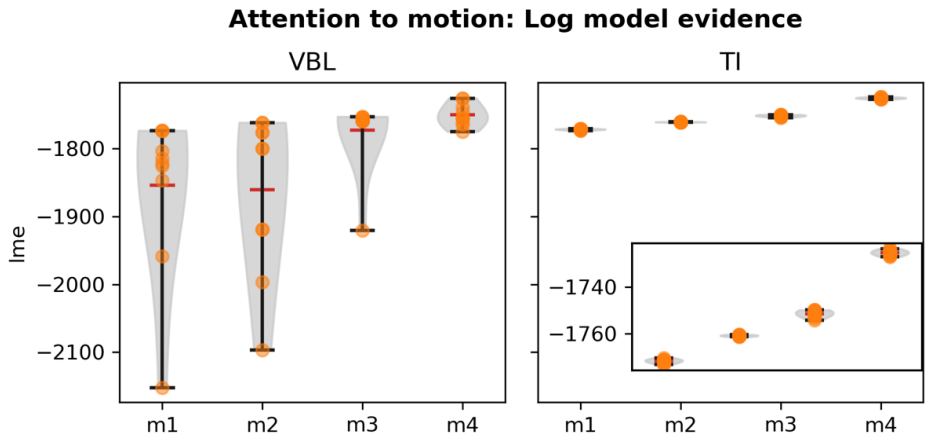


Figure 10: Estimates of the LME attention to motion dataset after initializing VBL and TI from 10 different expansion points (yellow points) drawn from the prior. TI estimates show much lower variability as compared to VBL estimates. The inset on the right panel zooms in the TI estimates.

Face perception

For the face perception dataset, the same number of iterations (16×10^3), and discarded burn-in samples (8×10^3) were used, but the number of chains was increased to 96. This has the advantage of better leveraging the parallel architecture of a GPU. Representative computation times for the four models are displayed in Table 9. We found that for all but 3 chains (<1%), the $\hat{R}$ was below 1.1, indicating the convergence of the algorithm. The difference between TI and VBL estimates of the LME, accuracy and log likelihood at the MAP are shown in Table 10. Again, differences between the estimates were apparent, but more importantly there were differences in the actual ranking of the models. In particular, while VBL favored model m_1, TI favored model m_3. Interestingly, although the estimates of the LME showed large differences, the estimates of the accuracy were similar for the two inversion schemes.

Computation time: Face perception				
	m_1	m_2	m_3	m_4
Simulations/second	1182.0	1216.7	1171.3	1184.3

Table 9: Computational time in simulations per second. A total of 2820 integration steps per simulation were performed in a GPU.

Face perception dataset				
	m1	**m2**	**m3**	**m4**
Log model evidence				
VBL	-3060.8	-3088.4	-3071.6	-3065.8
TI	-3051.5	-3059.7	-3032.9	-3051.8
Accuracy				
VBL	-2303.6	-2313.7	-2279.6	-2307.4
TI	-2309.9	-2317.9	-2279.8	-2303.2
Likelihood at MAP estimate				
VBL	-2223.3	-2233.3	-2194.9	-2226.5
TI	-2276.5	-2281.1	-2236.5	-2268.4

Table 10: Log model evidence, accuracy and log likelihood at the MAP estimate using both TI and VBL for the four DCMs of the "face perception" dataset

To better understand the observed differences between the algorithms, we repeated the simulations 10 times by sampling the starting points of the algorithms from the prior, after scaling its variance by $1/\sqrt{10}$. The scaling was used because of numerical instabilities encountered with the VBL algorithm. Results for the LMEs under the different starting positions are shown in Fig. 11. Consistent with the attention to motion dataset, the variance of the VBL estimates was much higher than the TI estimates.

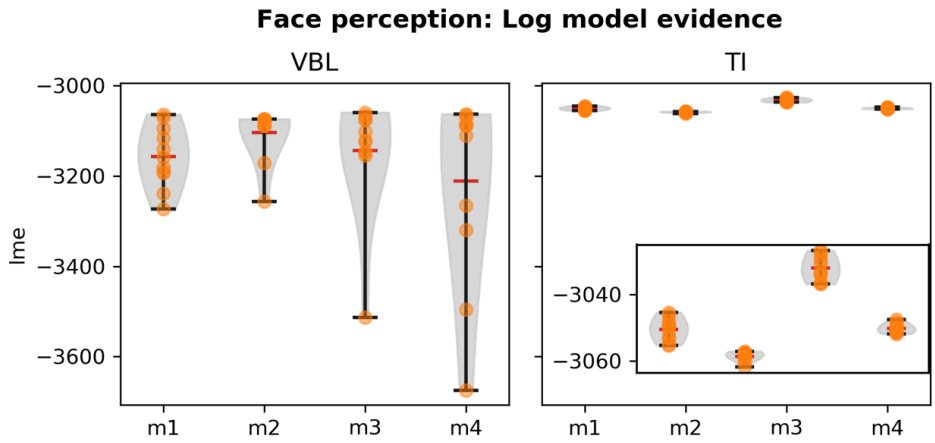


Figure 11: Estimates of the LME in the face perception dataset after initializing VBL and TI from 10 different expansion points drawn from the prior. Clearly, TI estimates show lower variability. The inset on the right panel zooms in the TI estimates.

This was also apparent when inspecting the predicted BOLD signal time series from TI and VBL for the model with the highest score (m_3) under the different initializations of the two algorithms (Fig. 12). Both VBL and TI generated qualitatively similar predictive fits; however, TI yielded more consistent results, suggesting that gradient ascent optimization might lead to highly variable estimates.

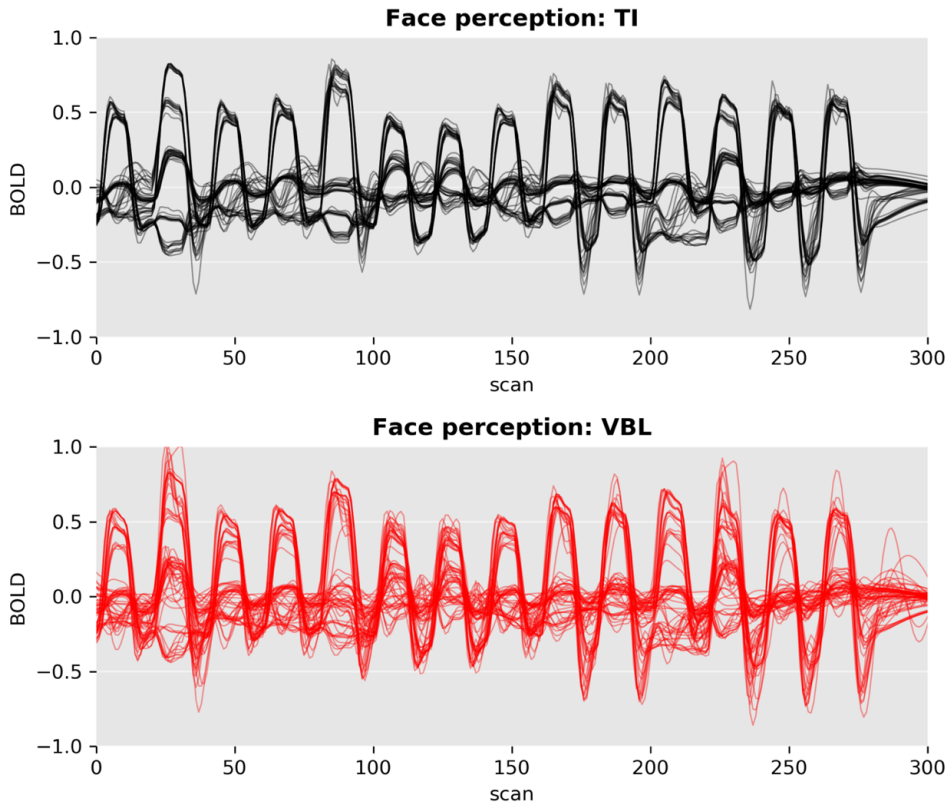


Figure 12. Comparison of the fits of model m3 in Frassle et al., (2016c) for different starting points of the algorithm. Displayed are the predicted time course of 6 regions. Although both estimates are qualitatively similar, the fits obtained under VBL display a much higher variability.

Discussion

In this technical note, we have described in detail various options for approximating the model evidence of probabilistic generative models. In brief, our analyses gave three main results. First, we replicated previous reports (for example Lartillot and Philippe, 2006) that HME and AME exhibit inadequate performance and are not well-suited for estimating the LME. Having said this, it is important to note that variants of both estimators have been proposed that aim at solving some of the known problems (Penny and Sengupta, 2016).

Second, TI provided robust estimates of LME, with superior performance compared to other estimators. It therefore represents a promising method for particularly challenging generative models, such as DCMs of electrophysiological data (Penny and Sengupta, 2016; Sengupta et al., 2016; 2015) or hierarchical models of DCM (Raman et al., 2016).

Third, although VBL was robust in most instances, we found evidence for variability in the estimates due to local optima in the objective function – especially for challenging scenarios where the number of network nodes and free parameters is high. While its computational efficiency and relatively robust performance justify VBL as a default choice for standard applications of DCM for fMRI, sampling-based approaches like TI might become the method of choice when the robustness and validity of single-subject inference is paramount. For example, the utility of generative models for clinical applications, such as differential diagnosis based on model comparison or prediction of individual treatment responses (Stephan et al., 2017), heavily depends on our ability to draw reliable and accurate conclusions from model estimates.

Comparison VBL / TI

Given the widespread use of VBL, its comparison to TI is of particular interest. In simulated data, both implementation yielded similar results regarding model estimates and cross-model comparison. More specifically, although VBL estimates were higher when compared to TI, both methods were qualitatively similar and lead to similar conclusions in terms of model comparison.

For the two empirical datasets tested in this paper, LME estimates differed considerably between VBL and TI. For the "attention to motion" dataset, this was most likely due to the optimizer used in VBL converging to a local optimum. This can be seen by different initializations (i.e., starting positions randomly sampled from the prior) generating higher variability in the VBL estimates as compared to TI.

This last observation should be seen in the light of several important aspects. First, the VBL algorithm used here is based on gradient optimization (Friston et al., 2007), and thus is intrinsically susceptible to local minima. This problem can be ameliorated by initializing the optimizer from different starting points or using global optimization methods (see (Lomakina et al., 2015)). Second, our results are not directly comparable to previous findings in the literature (Chumbley et al., 2007) as the integrator used here (Aponte et al., 2016) fully accounts for the nonlinearities of the Balloon-model. This might result in a more difficult posterior landscape than for the integrator routinely used in DCM (Friston et al., 2003), which uses several linear approximations.

Third, the lower variability of the results obtained using MCMC reflects a trade-off between computation time and variance of the estimators. More specifically, it is not surprising that a computationally far more intensive approach like TI generates more consistent estimates as compared to the highly efficient VBL algorithm. Finally, MCMC-based methods are also susceptible to failures in convergence, and thus sampling does not constitute a universal solution to the variance in the estimates. Instead, both VBL and MCMC estimates have to be carefully examined for convergence, using over-dispersed initialization points (Gelman et al., 2003).

Thermodynamic integration

Thermodynamic integration has received relatively little attention in the neuroimaging and cognitive science community until now (but see for example (Aponte et al., 2016; Penny and Sengupta, 2016). As mentioned before, this is arguably due to the high computational burden induced by simulating from an ensemble of chains. Although the TI estimator is a computationally expensive method for computing the LME, the inherent parallel nature of this technique can be exploited to obtain estimates in reasonable time. In particular, as suggested by Calderhead and Girolami (2009), combining TI with population MCMC tends to increase the efficiency of the sampler and can tackle multimodal probability landscapes. This observation has been also reported by (Ballnus et al., 2017), which provided benchmark evidence that multi-chain methods can efficiently explore the posterior landscape of dynamical systems comparable to DCM, and that the computational burden is offset by increased sampling efficiency. Here, we have shown that advances in hardware allow obtaining as many as 10^5 samples of realistic DCMs in only few minutes. Thus, our approach provides an alternative to VBL that is available to the community as open source software. We anticipate that further advances in sampling algorithms, specialized hardware, and software implementations will further reduce the computational time required to obtain highly accurate MCMC estimates of the model evidence. This will facilitate identification of a wide variety of nonlinear models, for which approximate methods as VBL are not adequate. Currently, the mpdcm toolbox (Aponte et al., 2016) supports parallelization both with Nvidia GPUs and multithreading in CPUs that does not require specialized hardware.

One of the key advantages of establishing an MCMC framework for model comparison using TI is the estimation of more complex hierarchical extensions of DCM (e.g. (Raman et al., 2016)). While the derivation of the VB equations for such an extension is difficult, inference using the MCMC framework would be possible. Other important applications are cases where a multi-model posterior distribution may pose a greater challenge for inference, e.g., due to local minima. Examples beyond those discussed above include DCMs for layered fMRI (Heinzle et al., 2016) or electrophysiological signals (Kiebel et al., 2009) (Penny and Sengupta, 2016) and, in particular, conductance-based DCMs (Moran et al., 2013b). Along these lines, it has been shown for DCM for EEG that gradient-based sampling-based methods outperform other, more conventional techniques, such as the Metropolis Hastings algorithm used here (Sengupta et al., 2016; 2015). However, note that we have used a combination of different methods, such as population MCMC, adaptive MCMC, and Gibbs sampling.

A promising method to reduce the computational demands imposed by TI is the Widely Applicable Bayesian Information Criterion (Watanabe, 2013). This method combines the observation that, according to the intermediate value theorem there exists an optimal temperature t^* such that

$$-F_H(1) = A(t^*) \tag{75}$$

Advanced mathematical methods provide a generally valid asymptotic approximation

$$t^* \approx \ln\frac{1}{n} \tag{76}$$

where n is the number of observations. This asymptotic approximation is valid even when a model does not satisfy the regularity conditions required by Laplace-based methods and traditional asymptotic approximations. This result can be particularly useful in the case of hierarchical models in which the computational burden induced by TI is still prohibitive.

Summary

Here we compared estimates of the model evidence based on VBL and TI. Our results suggest that VBL provides comparable results to TI in many instances, although it can be susceptible to local minima.

Sampling-based methods are computationally much more expensive but are less susceptible to the above problems. Both methods should be examined carefully for signs of convergence failures. In summary, sampling-based approaches have great potential for applications where regularity conditions are not satisfied, as in hierarchical models, or when the prior density does not satisfy any conjugacy property, and when sufficient computational resources are available.

Software note

The method described in this paper is available as part of the mpdcm toolbox in the open source TAPAS software (www.translationalneuromodeling.org/software).

Acknowledgements

The authors acknowledge support by the René and Susanne Braginsky Foundation (KES), the Clinical Research Priority Program “Multiple Sclerosis” (KES, SR) and “Molecular Imaging” (KES) at the University of Zurich, the Wellcome Trust (WP: core grant 091 593/Z/10/Z; BS: 088130/Z/09/Z), the ETH Zurich Postdoctoral Fellowship Program and the Marie Curie Actions for People COFUND Program (SF).

Appendix

Appendix 1

In the SPM version used here (5236), BOLD signals y are rescaled with respect to their ℓ_∞ norm, such that

$$\left|\left|y\right|\right|_\infty = 4. \tag{A.1}$$

The confound matrix X_0 usually consists of cosine functions that account for baseline effects and low frequency components and can be imagined as implementing a noise model of the residuals. We assume N observations such that data from a region is $y[t],\ t = 0, \dots, N-1$, and the components of $X_0 = [x_K, \dots, x_{M-1}]^T$, $K > 0$ are

$$x_k[t] = \cos(\frac{2\pi kt}{N}).$$

In this case $X_0^T X_0$ is a diagonal matrix. The diagonal elements are given by

$$\sum_{n=0}^{N-1} \cos\left(\frac{2\pi\omega n}{N}\right)^2 = \frac{1}{4}\sum_{n=0}^{N-1}\left(\exp\frac{in\omega 2\pi}{N} + \exp -\frac{in\omega 2\pi}{N}\right)^2, \tag{A.2}$$

$$= \frac{1}{4}\sum_{n=0}^{N-1}\left(\exp\frac{in\omega 4\pi}{N} + 2 + \exp -\frac{in\omega 4\pi}{N}\right), \tag{A.3}$$

$$= \frac{N}{2} + \left(\frac{1-\exp i\omega 4\pi}{1-\exp\frac{i\omega 4\pi}{N}}\right) + \left(\frac{1-\exp -i\omega 4\pi}{1-\exp -\frac{i\omega 4\pi}{N}}\right) \tag{A.4}$$

$$= \frac{N}{2} \tag{A.5}$$

Thus,

$$X_0^T X_0 = \frac{N}{2} I. \tag{A.6}$$

The posterior variance of the regressors conditioned on the predictions from DCM, the variance of the error σ_c^2, and the prior variance σ_0, is

$$(\sigma_c^{-2} X^T X + \sigma_0^{-2} I)^{-1} = \left(\frac{\sigma_c^{-2} N}{2} I + \sigma_0^{-2} I\right)^{-1}$$

$$= \left(\frac{N}{2\sigma_c^2} + \frac{1}{\sigma_0^2}\right)^{-1} I.$$

To derive the prior variance of the signal predicted by $X_0\beta$, we note that for the predicted signal y:

$$E[y[t]^2] = E\left[\left(\sum_{\omega=K}^{M-1} \beta_\omega \cos\frac{2\pi t\omega}{N}\right)^2\right], \tag{A.7}$$

$$= E\left[\sum_{\omega,k=K}^{M-1} \beta_\omega\beta_k \cos\frac{2\pi t\omega}{N}\cos\frac{2\pi tk}{N}\right]. \tag{A.8}$$

Because the coefficients are assumed to be uncorrelated and to have zero mean, it follows that

$$= \sum_{\omega=K}^{M-1} Var(\beta_\omega)\cos^2\left(\frac{2\pi t\omega}{N}\right), \tag{A.9}$$

$$= \sigma_0^2\left(\sum_{\omega=0}^{M-1}\cos^2\left(\frac{2\pi t\omega}{N}\right) - \sum_{\omega=0}^{K-1}\cos^2\left(\frac{2\pi t\omega}{N}\right)\right). \tag{A.10}$$

Assuming that $2Mt/N$ is an integer, if follows that

$$= \sigma_0^2\left(\frac{M}{2} - \sum_{\omega=0}^{K-1}\cos^2\left(\frac{2\pi t\omega}{N}\right)\right). \tag{A.11}$$

Trivially it follows that

$$\frac{\sigma_0^2(M-K)}{2} \leq E[y[t]^2] = Var(y[t]) \leq \frac{\sigma_0^2 M}{2}. \tag{A.12}$$

This constitutes an approximation to the prior variance of the signal. Although in the SPM implementation of DCM used here, σ_0^2 is set to 10^8, here we use a more pragmatic value $\sigma_0 = \left||y|\right|_\infty = 4$. From Eq. A.12, it can be seen that this constitutes a more conservative prior variance than the SPM implementation, but still liberal enough to a priori easiliy account for the totality of the variance in the data.

Appendix 2

The expression for the variational negative free energy can be derived by noting that Eq. 61 can be written as an energy term plus an entropy term

$$-F_{VB} = \mathrm{E}[\ln p(y,\theta)]_{q(\theta)} - \mathrm{E}[\ln q(\theta)]_{q(\theta)}. \tag{A.13}$$

For simplicity, in the rest of this section, we collapse parameters Θ and hyperparameters Λ into a d-dimensional vector θ, assuming that a maximum has been obtained. Also, we assume that all densities are

conditioned on model m, and make this assumption implicit. Moreover, we assume that the prior distribution of parameters θ is a Gaussian distribution centered at θ_0 with covariance Π_0^{-1}.

According to the Laplace approximation, $q(\theta)$ is a Gaussian distribution with mean $\theta^* = \arg\max_{\theta} p(y, \theta)$ and variance

$$\Pi = -\frac{\partial^2 \ln p(y,\theta)}{\partial \theta^2} = \Pi_0 - \frac{\partial^2 \ln p(y|\theta)}{\partial \theta^2}. \tag{A.14}$$

We denote the negative Hessian of the likelihood or observed Fisher information in the following as Π_L.

The energy term in Eq. A.13 is approximated using the Laplace method, which yields

$$E[\ln p(y,\theta)]_{q(\theta)} \approx \ln p(y,\theta^*) - \frac{1}{2}\mathrm{E}[(\theta^* - \theta)'\Pi(\theta^* - \theta)]_{q(\theta)}, \tag{A.15}$$

$$= \ln p(y,\theta^*) - \frac{1}{2} tr\big(\Pi \mathrm{E}[(\theta^* - \theta)(\theta^* - \theta)']_{q(\theta)}\big), \tag{A.16}$$

$$= \ln p\,(y,\theta^*) - \frac{1}{2} tr(\Pi\Pi^{-1}) = \ln p(y,\theta^*) - \frac{1}{2} d. \tag{A.17}$$

where tr denotes the trace operator.

The last term in Eq. A.13 the entropy of a Gaussian distribution, which is given by:

$$-E[\ln q(\theta)]_{q(\theta)} = \frac{1}{2}(d \ln 2\pi + \ d - \ln|\Pi|). \tag{A.18}$$

where Π is the precision of q.

Plugging Eqs. A.17 and A.18 into Eq. A.13, the variational free energy is given by

$$-F_{VB} = \ln p(y,\theta^*) + \frac{1}{2}(d \ln 2\pi - \ln|\Pi|). \tag{A.19}$$

The first term on the right of Eq. A.19 can be expanded to obtain the full expression:

$$\ln p(y,\theta^*) = \ln p(y|\theta^*) + \ln p(\theta^*), \tag{A.20}$$

$$= \ln p(y|\theta^*) - \frac{1}{2} d \ln 2\pi + \frac{1}{2}\ln|\Pi_0| - \frac{1}{2}(\theta^* - \theta_0)'\Pi_0(\theta^* - \theta_0). \tag{A.21}$$

where θ_0 and Π_0 are the mean and precision of the prior density, respectively. By inserting Eq. A.21 into Eq. A.19, the scheme proposed by (Friston et al., 2007) can be written as:

$$-F_{VB} = \ln p(y|\theta^*) + \frac{1}{2}\ln\frac{|\Pi_0|}{|\Pi|} - \frac{1}{2}(\theta^* - \theta_0)'\Pi_0(\theta^* - \theta_0). \quad \text{(A.22)}$$

Although VBL is typically orders of magnitude faster than MCMC sampling, it exhibits several limitations: it is susceptible to (i) local extrema, (ii) violations of the distributional assumptions imposed on the posterior, (iii) violations of the conditional independence assumptions of the mean field approximation (see Daunizeau et al., 2011) for discussion), and (iv) it is only defined when the Hessian in Eq. A.14 is not singular.

Returning to the connection between TI to VBL, one can write the variational negative free energy in terms of an approximate accuracy and complexity term (Eq. 58). One observes that the accuracy term can be computed as

$$-F_{VB} + KL(q(\theta)||p(\theta)) = A_{VB}. \quad \text{(A.23)}$$

Given a Gaussian prior and posterior, the KL divergence has the following analytical form:

$$\begin{aligned} KL(q(\theta)||p(\theta)) & \\ &= \frac{1}{2}\left[\ln\frac{|\Pi|}{|\Pi_0|} + tr(\Pi_0\Pi^{-1}) - d\right. \\ &\left. + (\theta^* - \theta_0)'\Pi_0(\theta^* - \theta_0)\right]. \end{aligned} \quad \text{(A.24)}$$

Replacing terms, we obtain

$$A = E[\ln p(y|\theta)]_{q(\theta)} \quad \text{(A.25)}$$

$$\approx \mathrm{A_{VB}} = \ln p(y|\theta^*) + \frac{tr(\Pi_0\Pi^{-1})}{2} - \frac{d}{2}. \quad \text{(A.26)}$$

A more familiar expression for the accuracy can be derived by noting that the posterior covariance can be written as the sum of the negative Hessian of the likelihood plus the prior covariance, such that

$$\mathrm{A_{VB}} = \ln p(y|\theta^*) + \frac{1}{2}tr\left(\frac{\Pi_0 + \Pi_L - \Pi_L}{\Pi_0 + \Pi_L}\right) - \frac{d}{2}, \quad \text{(A.27)}$$

$$= \ln p(y|\theta^*) - \frac{1}{2}tr\left(\frac{\Pi_L}{\Pi_0 + \Pi_L}\right), \quad \text{(A.28)}$$

$$\mathbb{p} = tr\left(\frac{\Pi_L}{\Pi_0 + \Pi_L}\right). \tag{A.29}$$

$\mathbb{p}$ is the effective number of parameters proposed by (Moody, 1992) Eq. 18 and see (Spiegelhalter et al., 2002) Eq. 15 and is commonly used for model selection. For example, the Deviance Information Criterion (Spiegelhalter et al., 2002) is

$$DIC = \ln p(y|\theta^*) - \mathbb{p}. \tag{A.30}$$

As a model selection criterion, DIC is motivated by the Akaike Information Criterion (ignoring multiplicative factors):

$$AIC = \ln p(y|\theta^*) - p, \tag{A.31}$$

Where p is the number of parameters of the model. As the largest eigenvalue of Π_0 goes to zero, $\mathbb{p} \to p$.

Appendix 3

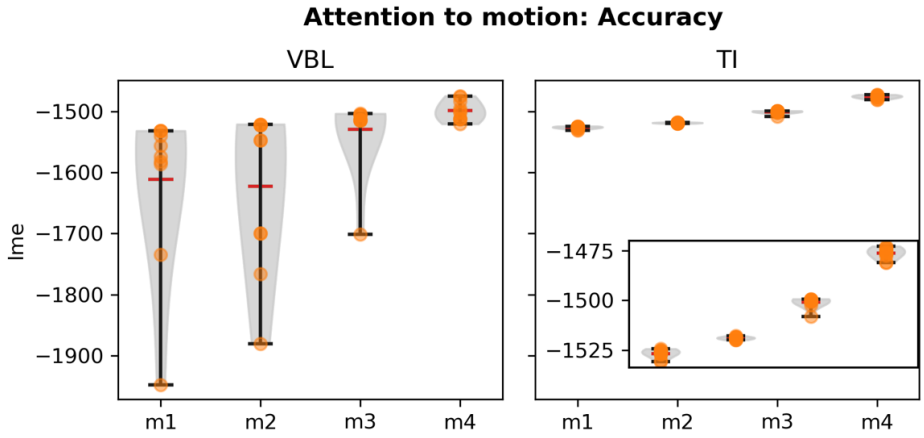


Figure 13: Estimates of the accuracy in the attention to motion dataset after initializing VBL and TI from 10 different expansion points drawn from the prior.

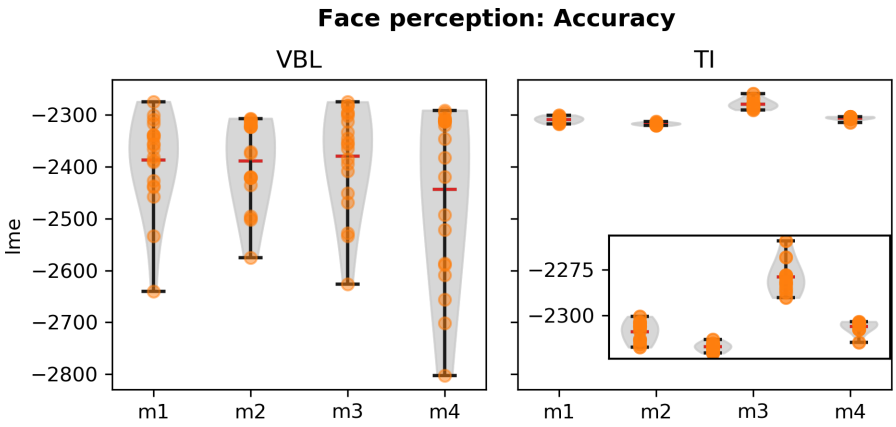


Figure 14: Estimates of the LME in the face perception dataset after initializing VBL and TI from 10 different expansion points drawn from the prior.

References

Aponte, E.A., Raman, S., Sengupta, B., Penny, W.D., Stephan, K.E., Heinzle, J., 2016. mpdcm: A toolbox for massively parallel dynamic causal modeling. J. Neurosci. Methods 257, 7–16.

Aponte, E.A., Schobi, D., Stephan, K.E., Heinzle, J., 2017. The Stochastic Early Reaction, Inhibition, and Late Action (SERIA) Model for Antisaccades SERIA - A model for errors and reaction times in the antisaccade task. bioRxiv. doi:10.1101/109090

Ballnus, B., Hug, S., Hatz, K., Gorlitz, L., Hasenauer, J., Theis, F.J., 2017. Comprehensive benchmarking of Markov chain Monte Carlo methods for dynamical systems. BMC Syst Biol 11, 63.

Blundell, S.J., Blundell, K.M., 2009. Concepts in thermal physics. OUP Oxford.

Buchel, C., Friston, K.J., 1997. Modulation of connectivity in visual pathways by attention: cortical interactions evaluated with structural equation modelling and fMRI. Cereb. Cortex 7, 768–778.

Buxton, R.B., Wong, E.C., Frank, L.R., 1998. Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. Magn Reson Med 39, 855–864.

Chumbley, J.R., Friston, K.J., Fearn, T., Kiebel, S.J., 2007. A Metropolis-Hastings algorithm for dynamic causal models. Neuroimage 38, 478–487.

Calderhead, Girolami, M.A., 2009. Estimating Bayes factors via thermodynamic integration and population MCMC. Computational Statistics \& Data Analysis 53, 4028–4045. doi:10.1016/j.csda.2009.07.025

Daunizeau, J., David, O., Stephan, K.E., 2011. Dynamic causal modelling: a critical review of the biophysical and statistical foundations. Neuroimage 58, 312–322.

Frassle, S., Krach, S., Paulus, F.M., Jansen, A., 2016a. Handedness is related to neural mechanisms underlying hemispheric lateralization of face processing. Sci Rep 6, 27153.

Frassle, S., Paulus, F.M., Krach, S., Jansen, A., 2016b. Test-retest reliability of effective connectivity in the face perception network. Hum Brain Mapp 37, 730–744.

Frassle, S., Paulus, F.M., Krach, S., Schweinberger, S.R., Stephan, K.E., Jansen, A., 2016c. Mechanisms of hemispheric lateralization: Asymmetric interhemispheric recruitment in the face perception network. Neuroimage 124, 977–988.

Frassle, S., Stephan, K.E., Friston, K.J., Steup, M., Krach, S., Paulus, F.M., Jansen, A., 2015. Test-retest reliability of dynamic causal modeling for fMRI. Neuroimage 117, 56–66.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., Penny, W., 2007. Variational free energy and the Laplace approximation. Neuroimage 34, 220–234.

Friston, K.J., Dolan, R.J., 2010. Computational and dynamic models in neuroimaging. Neuroimage 52, 752–765.

Friston, K.J., Harrison, L., Penny, W., 2003. Dynamic causal modelling. Neuroimage 19, 1273–1302.

Friston, K.J., Mechelli, A., Turner, R., Price, C.J., 2000. Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics. Neuroimage 12, 466–477.

Gelman, A., B, C.J., S, S.H., B, R.D., 2003. Bayesian Data Analysis. Chapman and Hall/CRC.

Gelman, A., Meng, X.L., 1998a. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. Statistical Science 13, 163–185.

Gelman, A., Meng, X.L., 1998b. Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling. Statistical Science 13, 163–185. doi:10.2307/2676756

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Statistical Science 457–472.

Haxby, J.V., Hoffman, E.A., Gobbini, M.I., 2000. The distributed human neural system for face perception. Trends Cogn. Sci. (Regul. Ed.) 4, 223–233.

Heinzle, J., Koopmans, P.J., Ouden, den, H.E.M., Raman, S., Stephan, K.E., 2016. A hemodynamic model for layered BOLD signals. Neuroimage 125, 556–570.

Henson, R.N., Mattout, J., Phillips, C., Friston, K.J., 2009. Selecting forward models for MEG source-reconstruction using model-evidence. Neuroimage 46, 168–176.

Jaynes, E.T., 1957. Information theory and statistical mechanics. Physical review 106, 620.

Kanwisher, N., McDermott, J., Chun, M.M., 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. 17, 4302–4311.

Kass, R.E., Raftery, A.E., 1995a. Bayes Factors. Journal of the american statistical association 90, 773–795. doi:10.2307/2291091

Kass, R.E., Raftery, A.E., 1995b. Bayes factors. Journal of the american statistical association 90, 773–795.

Kiebel, S.J., Garrido, M.I., Moran, R., Chen, C.C., Friston, K.J., 2009. Dynamic causal modeling for EEG and MEG. Hum Brain Mapp 30, 1866–1876.

Kirkwood, J.G., 1935. Statistical mechanics of fluid mixtures. The Journal of Chemical Physics 3, 300–313.

Koller, D., Friedman, N., 2009. Probabilistic graphical models: principles and techniques. MIT press.

Lartillot, N., Philippe, H., 2006. Computing Bayes factors using thermodynamic integration. Syst. Biol. 55, 195–207.

Lomakina, E.I., Paliwal, S., Diaconescu, A.O., Brodersen, K.H., Aponte, E.A., Buhmann, J.M., Stephan, K.E., 2015. Inversion of hierarchical Bayesian models using Gaussian processes. Neuroimage 118, 133–145.

MacKay, D.J.C., 2003. Information Theory, Inference, and Learning Algorithms. Cambridge University Press.

Marreiros, A.C., Kiebel, S.J., Friston, K.J., 2008. Dynamic causal modelling for fMRI: a two-state model. Neuroimage 39, 269–278.

Moody, J.E., 1992. The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems, in:.

Presented at the Advances in neural information processing systems, pp. 847–854.

Moran, R., Pinotsis, D.A., Friston, K., 2013a. Neural masses and fields in dynamic causal modeling. Front Comput Neurosci 7, 57.

Moran, R.J., Campo, P., Symmonds, M., Stephan, K.E., Dolan, R.J., Friston, K.J., 2013b. Free energy, precision and learning: the role of cholinergic neuromodulation. J. Neurosci. 33, 8227–8236.

Neal, R.M., Hinton, G.E., 1998. A view of the EM algorithm that justifies incremental, sparse, and other variants, in: Learning in Graphical Models. Springer, pp. 355–368.

Ortega, P.A., Braun, D.A., 2013. Thermodynamics as a theory of decision-making with information-processing costs, in:. Presented at the Proc. R. Soc. A, p. 20120683.

Penny, W., Sengupta, B., 2016. Annealed Importance Sampling for Neural Mass Models. PLoS Comput Biol 12, e1004797.

Penny, W.D., 2012. Comparing dynamic causal models using AIC, BIC and free energy. Neuroimage 59, 319–330.

Penny, W.D., Stephan, K.E., Daunizeau, J., Rosa, M.J., Friston, K.J., Schofield, T.M., Leff, A.P., 2010. Comparing families of dynamic causal models. PLoS Comput Biol 6, e1000709.

Penny, W.D., Stephan, K.E., Mechelli, A., Friston, K.J., 2004a. Modelling functional integration: a comparison of structural equation and dynamic causal models. Neuroimage 23 Suppl 1, S264–274.

Penny, W.D., Stephan, K.E., Mechelli, A., Friston, K.J., 2004b. Comparing dynamic causal models. Neuroimage 22, 1157–1172.

Puce, A., Allison, T., Asgari, M., Gore, J.C., McCarthy, G., 1996. Differential sensitivity of human visual cortex to faces, letterstrings, and textures: a functional magnetic resonance imaging study. J. Neurosci. 16, 5205–5215.

Raftery, A.E., Newton, M.A., Satagopan, J.M., Krivitsky, P.N., 2006. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity.

Raman, S., Deserno, L., Schlagenhauf, F., Stephan, K.E., 2016. A hierarchical model for integrating unsupervised generative embedding and empirical Bayes. J. Neurosci. Methods 269, 6–20.

Rigoux, L., Stephan, K.E., Friston, K.J., Daunizeau, J., 2014. Bayesian model selection for group studies - revisited. Neuroimage 84, 971–985.

Robert, C., Casella, G., 2013. Monte Carlo statistical methods. Springer Science \& Business Media.

Schwarz, G., others, 1978. Estimating the dimension of a model. The annals of statistics 6, 461–464.

Sengupta, B., Friston, K.J., Penny, W.D., 2016. Gradient-based MCMC samplers for dynamic causal modelling. Neuroimage 125, 1107–1118.

Sengupta, B., Friston, K.J., Penny, W.D., 2015. Gradient-free MCMC methods for dynamic causal modelling. Neuroimage 112, 375–381.

Shaby, B., Wells, M.T., 2010. Exploring an adaptative Metropolis algorithm. Department of statistical science. Duke Universitiy, Durham, NC, USA.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64, 583–639.

Stephan, K.E., Iglesias, S., Heinzle, J., Diaconescu, A.O., 2015. Translational Perspectives for Computational Neuroimaging. Neuron 87, 716–732.

Stephan, K.E., Kasper, L., Harrison, L.M., Daunizeau, J., Ouden, den, H.E., Breakspear, M., Friston, K.J., 2008. Nonlinear dynamic causal models for fMRI. Neuroimage 42, 649–662.

Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., Friston, K.J., 2009. Bayesian model selection for group studies. Neuroimage 46, 1004–1017.

Stephan, K.E., Schlagenhauf, F., Huys, Q.J., Raman, S., Aponte, E.A., Brodersen, K.H., Rigoux, L., Moran, R.J., Daunizeau, J., Dolan, R.J., Friston, K.J., Heinz, A., 2017. Computational neuroimaging strategies for single patient predictions. Neuroimage 145, 180–199.

Stephan, K.E., Weiskopf, N., Drysdale, P.M., Robinson, P.A., Friston, K.J., 2007. Comparing hemodynamic models with DCM. Neuroimage 38, 387–401.

Swendsen, R., Wang, J.-S., 1986. Replica Monte Carlo Simulation of Spin-Glasses. Phys. Rev. Lett. 57, 2607–2609.

Watanabe, S., 2013. A widely applicable Bayesian information criterion. Journal of Machine Learning Research 14, 867–897.

Welvaert, M., Rosseel, Y., 2013. On the definition of signal-to-noise ratio and contrast-to-noise ratio for FMRI data. PLoS ONE 8, e77089.

Wipf, D., Nagarajan, S., 2009. A unified Bayesian framework for MEG/EEG source imaging. Neuroimage 44, 947–966.

Wolpert, R.L., Schmidler, S.C., 2012. α-Stable limit laws for harmonic mean estimators of marginal likelihoods. Statistica Sinica 1233–1251.

Part II

Chapter 4

Eye movements are the outcome of the most common decision that humans are confronted with: 'Where should I look at?' Interestingly, several psychiatric and neurological diseases are accompanied by changes in eye movement behavior (Hutton and Ettinger, 2006, Bittencourt et al., 2013; Terao et al., 2013; Myles et al., 2017). This is particularly prominent in patients diagnosed with schizophrenia, who display deficits in pursuit (smooth eye movements that track a slowly moving target) as well as in the antisaccade task, a psychometric paradigm that tests both the inhibition of prepotent responses as well as the initiation of voluntary saccades.

Despite the large body of experimental research in eye movements in psychiatry, the biological or computational mechanisms that characterize deficits in eye movements are still unclear. This circumstance has limited our current understanding of oculomotor deficits to phenomenological characterizationss in terms of summary statistics. In this chapter, Heinzle, Aponte, and Stephan (2016) present a brief review of computational models of eye movements in schizophrenia and layout a research agenda for eye movement research in computational psychiatry. Two of these topics are pursued in the rest of this dissertation: saccadic adaptation, and eye movements in the antisaccade task.

This chapter was published as *Heinzle, J., Aponte, E. A., & Stephan, K. E. (2016). Computational models of eye movements and their application to schizophrenia. Current Opinion in Behavioral Sciences, 11, 21-29*. It is a verbatim copy of the document:

https://doi.org/10.1016/j.cobeha.2016.03.008.

.

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Behavioral Sciences

Computational models of eye movements and their application to schizophrenia

Jakob Heinzle[1], Eduardo A Aponte[1] and Klaas Enno Stephan[1,2,3]

Patients with neuropsychiatric disorders, in particular schizophrenia, show a variety of eye movement abnormalities that putatively reflect alterations of perceptual inference, learning and cognitive control. While these abnormalities are consistently found at the group level, a particularly difficult and important challenge is to translate these findings into clinically useful tests for single patients. In this paper, we argue that generative models of eye movement data, which allow for inferring individual computational and physiological mechanisms, could contribute to filling this gap. We present a selective overview of eye movement paradigms with clinical relevance for schizophrenia and review existing computational approaches that rest on (or could be turned into) generative models. We conclude by outlining desirable clinical applications at the individual subject level and discuss the necessary validation studies.

Addresses
[1] Translational Neuromodeling Unit, Inst. for Biomedical Engineering, University of Zurich & Swiss Federal Institute of Technology (ETH Zurich), Switzerland
[2] Wellcome Trust Centre for Neuroimaging, University College London, United Kingdom
[3] Max Planck Institute for Metabolism Research, Cologne, Germany

Corresponding author: Heinzle, Jakob (heinzle@biomed.ee.ethz.ch)

Current Opinion in Behavioral Sciences 2016, **11**:21–29

This review comes from a themed issue on **Computational modelling**

Edited by **Peter Dayan** and **Daniel Durstewitz**

http://dx.doi.org/10.1016/j.cobeha.2016.03.008

Introduction

Eye movements represent easily measurable behavioural responses which provide rich information about latent (hidden) cognitive processes — such as perceptual inference, learning and decision-making — which are of central interest for disease theories of psychiatric disorders. Additionally, many psychiatric disorders are accompanied by pronounced eye movement abnormalities (for reviews see [1–3]). Together with the practical ease of data acquisition, this makes eye movements of great interest for translational and clinical applications in psychiatry. However, the wealth of existing experimental findings has not yet been translated into diagnostic tools for clinical practice. For example, while a general deficit of smooth pursuit eye movements (SPEM) in patients with schizophrenia allows for a nearly perfect separation of patients from healthy controls [4], this does not constitute practically relevant progress: the diagnosis of schizophrenia is not a clinical problem; and the diagnostic label 'schizophrenia' does not allow for patient-specific predictions due to the heterogeneous nature of this disorder [5].

One strategy to address this is computational psychiatry [6,7] which strives for understanding the cognitive and physiological underpinnings of aberrant behaviour by using mathematical models and, ultimately, translate these findings into clinical practice. The ongoing application of this approach to neuroimaging data has highlighted the importance of so-called 'generative models' (Figure 1) for clinical applications [8]. This is due to three main features (for detailed discussion and review, see [9]): generative models enforce mechanistic thinking about how observed data could have been caused; they deal with uncertainty (about model structure and parameters) in a principled way and thus provide a natural fundament for formalizing differential diagnosis; and they can be combined with unsupervised approaches, such as clustering, for detecting mechanistically distinct patient subgroups in heterogeneous disorders (e.g. [10]). Here, we review emerging generative models for eye movement data and discuss their possible role for translational research in psychiatry, with a focus on schizophrenia.

While there is evidence for disturbed eye movements in schizophrenia in many different tasks, historically, SPEM [2] and voluntary control of eye movements in antisaccades [1] have been the most widely used eye movement paradigms in schizophrenia. More recently, theories highlighting failures of inference and predictions in schizophrenia [11–13] have triggered an additional line of research focusing on corollary discharge (CD) during eye movements. The following sections revisit these paradigms, describe a selection of existing computational models, and hint at possible developments towards generative models.

Generative models for eye movements

Generative models represent a probabilistic mapping from latent (unobservable or hidden) variables θ (e.g. the parameters of a system) to observed data. This

Figure 1

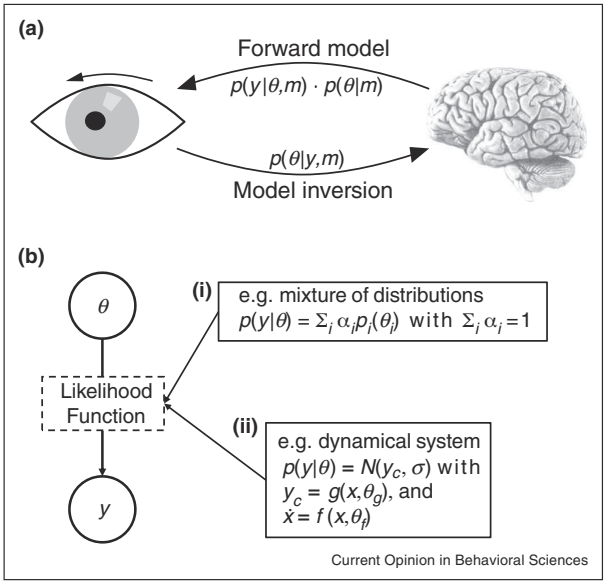


Illustration of generative modelling. **(a)** Schematic illustration of generative modelling. A generative model for eye movements describes how a latent (unobserved or hidden) neurophysiological process produces eye movements, for example, pursuit traces or reaction times (RTs) for saccades. The forward model m defines the joint probability of data (here: eye movement measurements) and model parameters; this results from the product of the a prior distribution $p(\theta|m)$ of model parameters and a likelihood function $p(y|\theta,m)$ that encodes the probability of the observed data y given the parameters θ. Model inversion corresponds to inferring the posterior probability of the parameters, given the data, $p(\theta|y,m)$. **(b)** Representation of the generative model in (a) as a graphical model. The likelihood specifies the mapping from parameters to data and thus encodes a particular proposal how the observed data were generated. A simple phenomenological approach is to assume that the data result from a weighted combination of distributions $p_i(\theta_i)$ with mixing weights a_i (panel **i**), Biologically more interpretable models can be constructed by choosing a hierarchically formulated likelihood, where hidden states x evolve according to a biophysically motivated dynamical system f (with parameters θ_f) and are linked to data through an observation function g (with parameters θ_g) and measurement noise ε. See panel **ii**. The data is assumed to be normally distributed around the predicted trace y_c with standard deviation σ. This formulation is known as dynamic causal modelling (DCM).

mapping is specified by two components (Figure 1; [14]): First, a prior distribution $p(\theta)$ defines the range of parameter values which are plausible a priori. Second, a likelihood function $p(y|\theta)$ specifies a mechanism by which measured data y are generated probabilistically, given the parameters. The product of prior and likelihood yields the joint probability of data and parameters. Models of this sort are called 'generative' because one can generate synthetic data, by feeding samples from the prior into the likelihood function.

The specific mechanism proposed by the likelihood function is one of the defining features of a particular generative model; for eye movements, this can take very different forms. A simple approach is to explain saccadic RTs phenomenologically, as a mixture of distributions (Figure 1, panel i). By contrast, biologically more interpretable models can be constructed by choosing a hierarchically structured likelihood function, where hidden (neuronal or computational) states evolve according to a biophysically motivated dynamical system f and are linked to data through a static observation function g with measurement noise σ (Figure 1, panel ii). This hierarchical formulation underlies a special class of generative models, so-called dynamic causal models (DCMs) [15]. It is possible, in principle, to extend existing dynamical models of eye movement control to full generative models. This requires rendering them fully probabilistic by introducing priors on the parameters and adding a probabilistic observation function.

For all generative models, statistical inference on the model parameters can be performed by computing the posterior probability of the parameters, given the data, using Bayes' rule (model inversion). The numerical feasibility of model inversion depends on the complexity of the model. Thus, restricting generative models to a limited number of unknown parameters is important for practical utility. Generative models also offer a principled approach for model comparison, based on the model evidence $p(y|m)$, which represents a principled measure for the trade-off between accuracy and complexity of a model.

This allows one to compare the relative plausibility of alternative dynamical system mechanisms [16] that might underlie observed eye movements.

Smooth pursuit eye movements

Among the different types of eye movements, studies of SPEM have the longest experimental tradition in schizophrenia research [2]. Patients with schizophrenia show a general deficit in SPEM which distinguishes them from healthy controls almost perfectly [4]. In addition, compared to controls, patients show reduced ability to predict a target's trajectory during occlusion [17]; at the same time, patients with schizophrenia are superior in tracking targets with unpredictable changes in their trajectory [18]. Both phenomena can be explained by the same putative mechanism, that is, reduced efficacy (precision) of predictions during perceptual inference [12,19]. This hypothesis is difficult to test with traditional mathematical models of SPEM, which have typically taken the form of dynamical systems with a focus on questions of gain control and less on prediction [20,21]. More recently, in order to account for predictions, Kalman filtering [22] and models based on the notion of 'predictive coding' (a hierarchical inference scheme where each level predicts the state of the next-lower level below and updates its predictions proportional to precision-weighted prediction errors; [23,24]) and 'active inference' (where actions are selected in order to fulfil sensory predictions) [19,25•] have been introduced to smooth pursuit. For example, the generative model introduced by Adams and colleagues [19,25•] is a dynamic causal model (DCM; Figure 1) that provides a physiological implementation of predictive coding principles during SPEM (Figure 2). This model has found application to empirical SPEM data from healthy volunteers [25•] and for simulating the empirically observed SPEM anomalies in patients with schizophrenia, including their superior performance in tracking target trajectories with unpredictable changes [19]. A recent combined SPEM and magnetoencephalography (MEG) application of this model demonstrated how a precision parameter of the pursuit model can be linked to recurrent connectivity in visual areas and inferred from MEG data [26••]. An important next step will be to apply this model to empirical data from patients, and to examine whether its parameter estimates allow for clinically relevant predictions in individual patients (see below).

Voluntary control of eye movements — the antisaccade task

In the antisaccade task, participants are required to withhold a reactive eye movement to a peripheral target and instead perform a saccade to the opposite location from the current fixation point. On this task, patients with schizophrenia show increased error rates (failures of withholding the reactive saccade) and increased latencies compared to healthy controls [1]. It is controversial whether this is due to a failure of inhibitory control or a failure of initiating the endogenous movement plan for the antisaccade [1,27]. This debate is mirrored by two models that have been applied to antisaccades in humans [28••,29•]; see Figure 3. In the LATER (Linear Approach to Threshold at Ergodic Rate) model [29•], the reactive prosaccade is stopped by a stop signal that races against the prosaccade; here, antisaccade errors are due to failure of inhibitory control. In the Cutsuridis model [28••], a competition between two alternative saccadic plans — prosaccade (error) and antisaccade (correct response) — determines the resulting saccade. The two models differ considerably in their implementation. The LATER model [29•] is a process model, which represents the evolution of a decision variable (essentially the log posterior odds ratio between two hypotheses) over time. It can be easily expanded into a full generative model of trial-wise reaction times (RTs), by formalizing its likelihood function [30] and specifying priors. The Cutsuridis model [28••], by contrast, specifies a detailed neuronal circuit within the superior colliculus whose activity determines RTs. So far, it has been used to simulate some of the deficits observed empirically in patients with schizophrenia [28••]. Using this model for inference from empirical data would require transforming it into a full generative model, possibly under appropriate simplifications.

Other models for antisaccades range from simple distributional models [31] to elaborate neurophysiological models of layered cortical units [32] or cortico-basal ganglia loops [33]. While offering direct links to physiology, the last two model types appear presently too complex to be transformed into generative models that could be inverted.

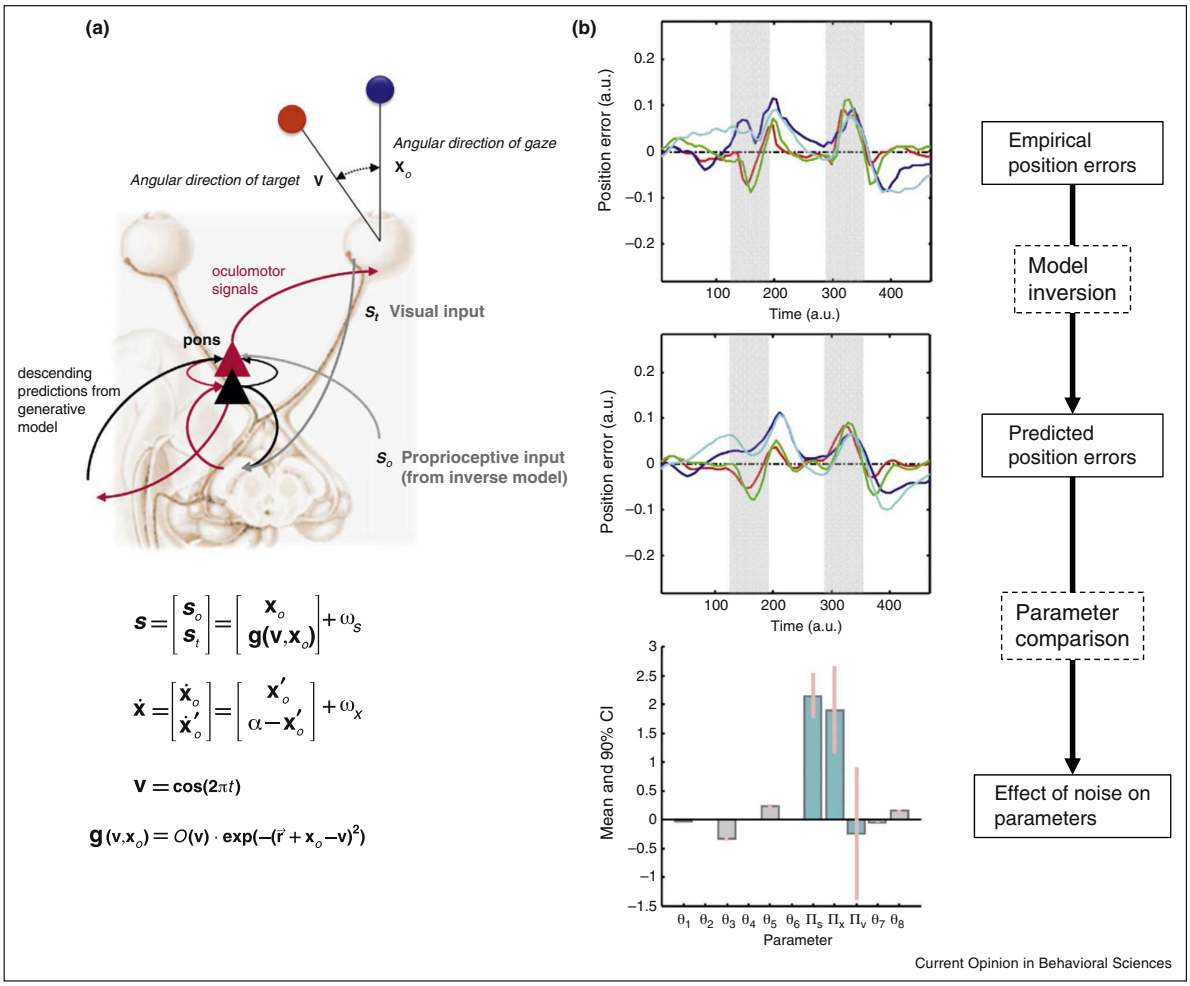
In summary, although fully generative models of antisaccades still need to be developed, a number of computational models exist which can be used as starting points.

Corollary discharge for saccadic eye movements

Corollary discharges (CD) are neuronal signals from executive (motor) areas that inform sensory areas about upcoming action [34] and thus enable a prediction about the changes in sensory inputs that result from one's own action. Influential hypotheses postulate that a failure of CD causes 'first rank' symptoms in schizophrenia: hallucinations and delusions of control (the sensation that an external force controls one's movements or thoughts) [11–13]. The neurophysiology of CD for saccadic eye movements has been extensively studied in primates (for a review see [34]) providing important constraints for models of CD. Figure 4 summarizes three eye movement tasks — double step saccades, perisaccadic change detection and saccadic adaptation — in which CD plays an essential role.

Several recent studies using these tasks have provided evidence for impaired CD during saccadic eye movements

Figure 2



A dynamic causal model of SPEM [25•], with state equations motivated by the notion of 'active inference' [19]. **(a)** Summary of the model's state equations for the generative process. Here, ***v*** is the angular direction of a target moving on a sinusoidal trajectory. Sensory input *s* includes proprioceptive (s_0) and retinal (s_t) input. Retinal input is modelled with Gaussian receptive fields (second term of observation function ***g***) and includes an occluder function $O(\boldsymbol{v})$ that turns retinal input on or off, depending on when the target is behind an occluder. ***x*** describes the hidden states, that is, angular position $\boldsymbol{x_0}$ and velocity $\boldsymbol{x'_0}$. Changes in position are driven by angular velocity; changes in velocity are driven by action (*a*). Both hidden states and sensory inputs are noisy, where the Gaussian noise is indicated by $\boldsymbol{\omega_s}$ and $\boldsymbol{\omega_x}$. For more details, see [25•]. **(b)** Top: Average empirical position errors (deviation of angular direction of gaze from target: $x_0 - v$) for four conditions (red: slow smooth target, blue: slow noisy target, green: fast smooth target, cyan: fast noisy target). Model inversion (parameter estimation) proceeds using these traces. Model fit is visible from predicted position errors (middle panel). Finally, comparison of posterior parameter estimates between noisy and non-noisy conditions allows for estimating an effect of sensory noise on model parameters (bottom).
Figure adapted from [25•] with permission (*Creative Commons Attribution License* (*CC BY*)).

in patients with schizophrenia. Using an error correction version of the double step paradigm, Thakkar *et al.* [35•] showed that CD for saccades is disrupted in patients with schizophrenia and is related to the severity of psychotic symptoms [36••]. Furthermore, both a stronger mislocalization in perisaccadic flash detection [37•] and reduced saccadic adaptation [38,39] have been reported in individuals with schizophrenia. While the latter was mainly interpreted as a cerebellar deficit by the authors, there is strong evidence that CD-based prediction errors play an important role in saccadic adaptation [40], with the superior colliculus as a crucial source of these error signals for adaptation [41].

To the best of our knowledge, no model of the double step paradigm exists so far. For perisaccadic change

Figure 3

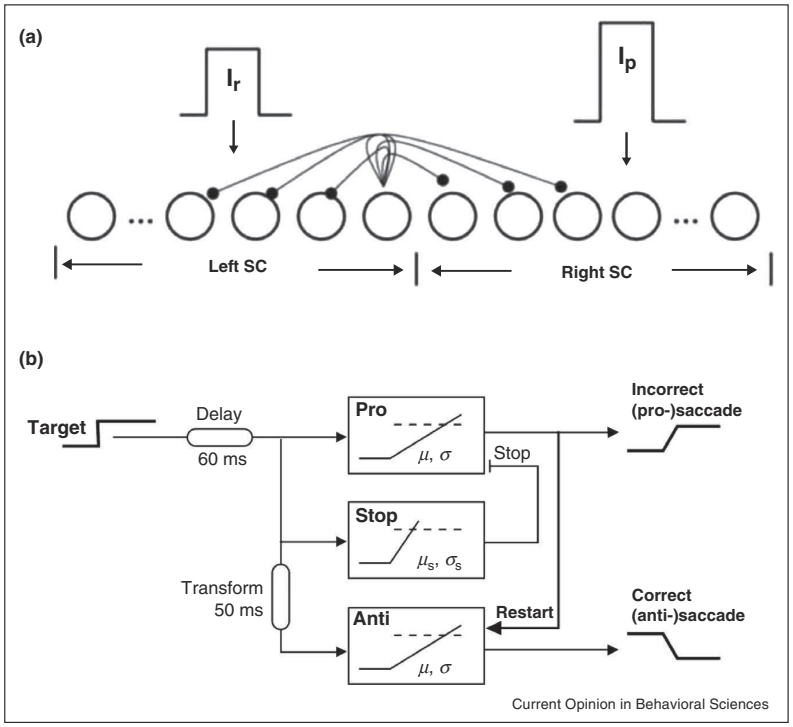


Illustration of two models of the antisaccade task. **(a)** Superior colliculus (SC) model by Cutsuridis *et al.* [28••]. A circuit of neuronal populations which code for different eccentricities along the horizontal axis and represent a competitive neural network within the SC. Two unspecified inputs, presumably of cortical origin, drive the prosaccade (reactive input I_r) and antisaccade (planned input I_p). Given these two inputs and some assumptions about differences in neuronal time constants between the two colliculi on the prosaccade and antisaccade side, the model reproduces a variety of findings from the antisaccade literature, including corrective saccades after errors. Figure reproduced from [28••] with permission *(Creative Commons Attribution License (CC BY))*. **(b)** The LATER model for antisaccades [29•] is based on three race-to-threshold units. On an antisaccade trial, prosaccade and stop units start a race, followed by the antisaccade unit with a delay. If the stop unit reaches threshold first, the prosaccade race is cancelled and the antisaccade unit defines the RT. If the prosaccade unit reaches threshold before the stop unit an error occurs, and the antisaccade unit is reset to trigger a corrective saccade.
Reproduced from [29•] with permission.

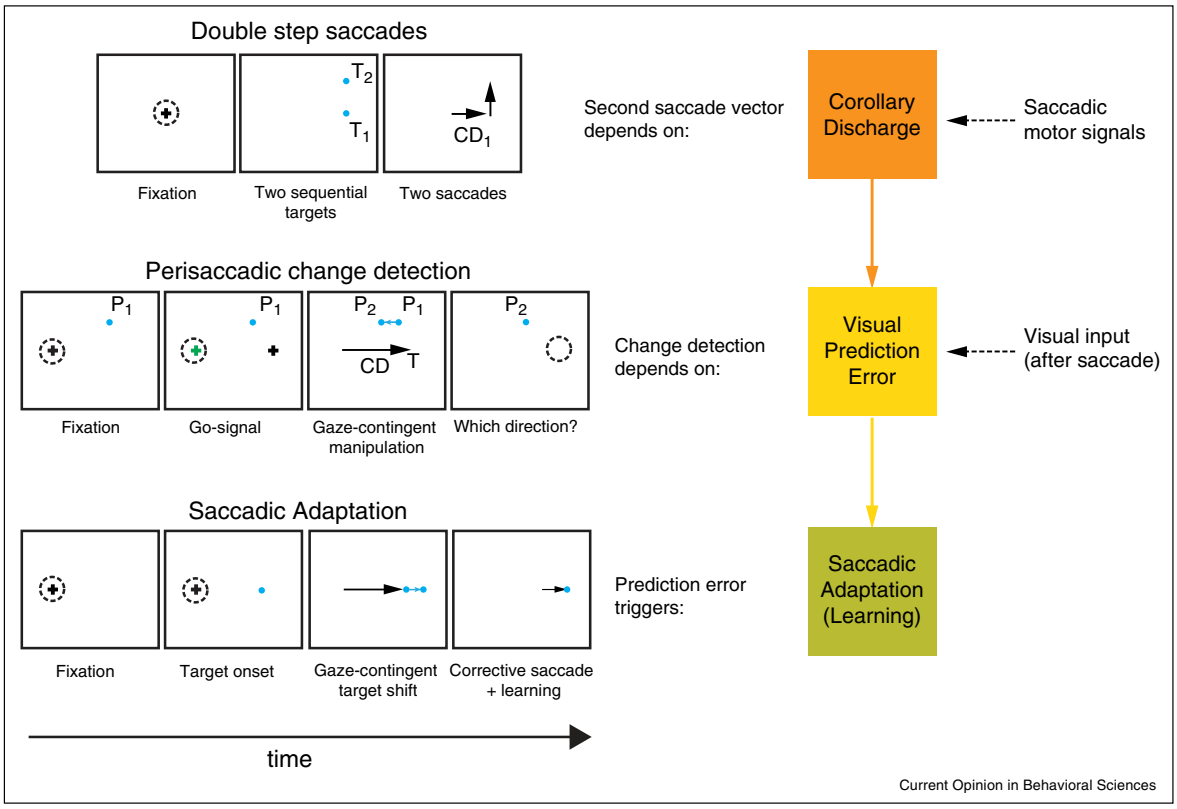
detection, Hamker and colleagues [42,43] have developed a detailed model of interactions between topographic cortical maps which could potentially be simplified to provide a generative model for perisaccadic change detection. Finally, saccadic adaption can be modelled by a simple learning rule [44] that is structurally equivalent to standard update equations in generative models of choice behaviour [45]. Models of this type can explain a range of observations for saccadic adaptation in monkeys [46]. Application of this or similar models to empirical data on saccadic adaptation from patients with schizophrenia [38,39] would be straightforward but is outstanding so far.

In summary, with the exception of saccadic adaption, generative models for CD in eye movements still need to be developed.

Additional directions

In this section, we briefly outline other eye movement paradigms with relevance for schizophrenia and generative modelling. Similar to SPEM, *visual scene scanning* almost perfectly distinguishes between patients and controls [4]. Generative models of scan paths can be derived from Bayesian models of attention [47] or active inference [48]. In addition to the scan paths, investigating fixational eye movements (small eye movements during fixations) would be of high interest. A recent study found that fixation stability during free viewing was the single most informative parameter for classification of schizophrenia patients [4]. Models of fixational eye movements are readily available [49,50] and have been fitted to data of healthy subjects using grid search [50]. It would be straight forward to extend these models to a fully generative framework. Second, *reading eye movements* are abnormal in patients [51]. Mathematical

Figure 4



Saccadic eye movement tasks that involve CD. Double step saccades require two consecutive saccades to briefly flashed targets T_1 and T_2. Both targets vanish before the first saccade is initiated from the fixation point F. In the absence of any visual target, the second saccade needs to be pre-computed as $FT_2 \rightarrow -CD_1 \rightarrow$, where $CD_1 \rightarrow$ is the vector represented by the corollary discharge for the first saccade. Hence, in this task, CD is only used for motor planning, not for predicting visual input. In the perisaccadic change detection task, a visual target at position P_1 is moved to position P_2 during the saccade. After landing, the expected retinal position of the target is $FP_1 \rightarrow -CD \rightarrow$, which has to be compared with the true retinal position $TP_2 \rightarrow$. Here, $CD \rightarrow$ is a vector representation of the corollary discharge and T the landing position of the saccade. In this setting, CD is used for the prediction of visual input after the saccade and thus enables computing a prediction error if target position changed. Finally, in saccadic adaption the visual target is moved consistently on every trial. The resulting prediction error is used to adapt saccade magnitude over trials. The right panel illustrates the relations between tasks.

models of cognitive and lexical processes [52,53] are able to reproduce a wide range of eye movement data in reading. These could be simplified to result in fully generative models. Finally, patients with schizophrenia show abnormal cue-guided spatial attention (*Posner paradigm*; [54]). Vossel *et al.* [55] have used a generative model, a hierarchical Gaussian filter [45], to infer the mechanisms which govern variation of saccadic RTs under volatility (changes in the predictive strength of the cue). This task and model have subsequently been combined with pharmacological (cholinergic) stimulation [56•] and fMRI [57].

Prospects for generative models of eye movements in schizophrenia research

In this final section, we briefly outline future translational and clinical opportunities for (generative) models of eye movements, with a focus on the three main paradigms described above.

Translation from animal to human research

The three eye movement paradigms described above are strongly dependent on cortical–subcortical loops that involve the frontal cortex [58,59] and are likely altered in the schizophrenia spectrum [60,61]. Studies of these circuits in primates [34,62,63] provide anatomical and physiological data which are essential for the development of biologically realistic models in humans [32,33]. An important next step is to simplify and recast these models as generative models in order to allow for inference on pathophysiological mechanisms in human patients.

Computational phenotyping, differential diagnosis and clinical predictions

Schizophrenia is a heterogeneous spectrum disease, where identical symptoms can arise from different mechanisms across patients. For example, while many symptoms in schizophrenia can be understood as arising from a general deficit in perceptual inference [12], this could be due to different causes. For example, from a computational perspective, hallucinations could plausibly arise from deficient CD, overly tight/inflexible high-level priors, or attenuated/misplaced low-level prediction errors (cf. [64]). A battery of simple eye movement tasks which allow cross-comparing models representing these competing explanations would introduce a valuable tool for differential diagnosis to clinical practice. This requires two things: prospective clinical studies which evaluate the predictive validity of model-based differential diagnosis against relevant clinical outcomes, and statistical model comparison techniques. The latter can require computationally demanding sampling techniques for complex models but will increasingly benefit from dedicated open source software [65].

Computational assays of neuromodulation

Similar to model-based EEG or MEG [66], generative models of eye movements could become useful as computational assays for neuromodulatory action, such as the availability of a particular neuromodulatory transmitter. While some model-based work has focused on neuromodulatory effects on pupil size [67,68], the pronounced sensitivity of saccadic eye movements to neuromodulatory alterations [69] has found remarkably little exploitation so far. If generative models of eye movements allowed for establishing sufficiently sensitive and specific assays of neuromodulatory abnormalities, this could provide valuable guidance for treatment decisions, for example, when deciding between antipsychotic drugs with differential emphasis on dopaminergic and cholinergic mechanisms [70]. Again, this eventually requires prospective clinical studies; initially, however, pharmacological validation studies need to be conducted that test whether generative models of eye movements can detect specific dopaminergic or cholinergic manipulations in single subjects.

Conclusion

Computational modelling of eye movement data is a promising way forward in schizophrenia research, but also for many other neuropsychiatric disorders where eye movement deficits are observed. In particular, analogous to similar developments in computational neuroimaging [8], generative models of eye movements might enable inference on pathophysiological and/or pathocomputational mechanisms which underlie eye movement abnormalities in single patients. Single subject parameter estimates or model comparison could then enable clinically relevant applications for differential diagnosis, to predict treatment outcome or aid treatment choices, and estimate risk of relapse or transition to disease. A key challenge for the future will be to finesse existing and develop novel generative models for eye movements; the neurophysiological interpretability and clinical utility of these models must then be evaluated in pharmacological validation studies and prospective patient studies.

Conflict of interest statement

Nothing declared.

Acknowledgements

We acknowledge support by the René and Susanne Braginsky Foundation and the University of Zurich.

References

1. Hutton SB, Ettinger U: **The antisaccade task as a research tool in psychopathology: a critical review**. *Psychophysiology* 2006, **43**:302-313.
2. O'Driscoll GA, Callahan BL: **Smooth pursuit in schizophrenia: a meta-analytic review of research since 1993**. *Brain Cogn* 2008, **68**:359-370.
3. Rommelse NN, Van der Stigchel S, Sergeant JA: **A review on eye movement studies in childhood and adolescent psychiatry**. *Brain Cogn* 2008, **68**:391-414.
4. Benson PJ *et al.*: **Simple viewing tests can detect eye movement abnormalities that distinguish schizophrenia cases from controls with exceptional accuracy**. *Biol Psychiatry* 2012, **72**:716-724.
5. Kapur S, Phillips AG, Insel TR: **Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it?** *Mol Psychiatry* 2012, **17**:1174-1179.
6. Stephan KE, Mathys C: **Computational approaches to psychiatry**. *Curr Opin Neurobiol* 2014, **25**:85-92.
7. Montague PR *et al.*: **Computational psychiatry**. *Trends Cogn Sci* 2012, **16**:72-80.
8. Stephan KE *et al.*: **Translational perspectives for computational neuroimaging**. *Neuron* 2015, **87**:716-732.
9. Stephan, K.E., et al., Computational neuroimaging and the single patient: a tutorial overview. Neuroimage, submitted for publication.
10. Brodersen KH *et al.*: **Dissecting psychiatric spectrum disorders by generative embedding**. *Neuroimage Clin* 2014, **4**:98-111.
11. Feinberg I: **Efference copy and corollary discharge: implications for thinking and its disorders**. *Schizophr Bull* 1978, **4**:636-640.
12. Fletcher PC, Frith CD: **Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia**. *Nat Rev Neurosci* 2009, **10**:48-58.
13. Pynn LK, DeSouza JF: **The function of efference copy signals: implications for symptoms of schizophrenia**. *Vision Res* 2013, **76**:124-133.
14. Bishop CM: *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York Inc.; 2007.
15. Friston KJ, Harrison L, Penny W: **Dynamic causal modelling**. *Neuroimage* 2003, **19**:1273-1302.
16. Penny WD *et al.*: **Comparing families of dynamic causal models**. *PLoS Comput Biol* 2010, **6**:e1000709.

17. Thaker GK *et al.*: **Smooth pursuit eye movements to extra-retinal motion signals: deficits in patients with schizophrenia**. *Psychiatry Res* 1999, **88**:209-219.

18. Hong LE *et al.*: **Is motion perception deficit in schizophrenia a consequence of eye-tracking abnormality?** *Biol Psychiatry* 2009, **65**:1079-1085.

19. Adams RA, Perrinet LU, Friston K: **Smooth pursuit and visual occlusion: active inference and oculomotor control in schizophrenia**. *PLoS ONE* 2012, **7**:e47502.

20. Barnes GR: **Cognitive processes involved in smooth pursuit eye movements**. *Brain Cogn* 2008, **68**:309-326.

21. Lisberger SG: **Visual guidance of smooth-pursuit eye movements: sensation, action, and what happens in between**. *Neuron* 2010, **66**:477-491.

22. Orban de Xivry JJ *et al.*: **Kalman filtering naturally accounts for visually guided and predictive smooth pursuit dynamics**. *J Neurosci* 2013, **33**:17301-17313.

23. Friston K: **A theory of cortical responses**. *Philos Trans R Soc Lond B Biol Sci* 2005, **360**:815-836.

24. Rao RP, Ballard DH: **Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects**. *Nat Neurosci* 1999, **2**:79-87.

25. • Adams RA *et al.*: **Active inference and oculomotor pursuit: the dynamic causal modelling of eye movements**. *J Neurosci Methods* 2015, **242**:1-14.
This paper describes the full inversion of a generative model for SPEM in humans.

26. •• Adams RA *et al.*: **Dynamic causal modelling of eye movements during pursuit: confirming precision-encoding in V1 using MEG**. *Neuroimage* 2016, **132**:175-189.
This paper provides a direct link between estimates from a generative model for SPEM to neural connectivity parameters estimated from simultaneously measured MEG data.

27. Reuter B, Kathmann N: **Using saccade tasks as a tool to analyze executive dysfunctions in schizophrenia**. *Acta Psychol (Amst)* 2004, **115**:255-269.

28. •• Cutsuridis V, Kumari V, Ettinger U: **Antisaccade performance in schizophrenia: a neural model of decision making in the superior colliculus**. *Front Neurosci* 2014, **8**:13.
Application of a biophysical model of the superior colliculus to antisaccade data from patients with schizophrenia. The authors demonstrate that their model can reproduce differences between patients and controls.

29. • Noorani I, Carpenter RH: **Re-starting a neural race: anti-saccade correction**. *Eur J Neurosci* 2014, **39**:159-164.
This paper proposes a distributional model for RTs in the antisaccade task. RTs are computed from a linear race-to-threshold model.

30. Brodersen KH *et al.*: **Integrated Bayesian models of learning and decision making for saccadic eye movements**. *Neural Netw* 2008, **21**:1247-1260.

31. Feng G: **Is there a common control mechanism for anti-saccades and reading eye movements? Evidence from distributional analyses**. *Vision Res* 2012, **57**:35-50.

32. Heinzle J, Hepp K, Martin KA: **A microcircuit model of the frontal eye fields**. *J Neurosci* 2007, **27**:9341-9353.

33. Wiecki TV, Frank MJ: **A computational model of inhibitory control in frontal cortex and basal ganglia**. *Psychol Rev* 2013, **120**:329-355.

34. Sommer MA, Wurtz RH: **Brain circuits for the internal monitoring of movements**. *Annu Rev Neurosci* 2008, **31**:317-338.

35. • Thakkar KN *et al.*: **Disrupted saccadic corollary discharge in schizophrenia**. *J Neurosci* 2015, **35**:9935-9945.
This paper shows that CD in an adapted double step paradigm is disrupted in patients with schizophrenia.

36. •• Rosler L *et al.*: **Failure to use corollary discharge to remap visual target locations is associated with psychotic symptom severity in schizophrenia**. *J Neurophysiol* 2015, **114**:1129-1136.
This paper goes beyond a simple correlation of CD with disease and shows that perisaccadic change detection is correlated with the severity of psychotic symptoms.

37. • Richard A *et al.*: **Perisaccadic perception of visual space in people with schizophrenia**. *J Neurosci* 2014, **34**:4760-4765.
This study applies a phenomenological model of CD during saccades and relates the estimated parameters to disease state and symptom severity.

38. Coesmans M *et al.*: **Cerebellar motor learning deficits in medicated and medication-free men with recent-onset schizophrenia**. *J Psychiatry Neurosci* 2014, **39**:E3-E11.

39. Picard H *et al.*: **Impaired saccadic adaptation in schizophrenic patients with high neurological soft sign scores**. *Psychiatry Res* 2012, **199**:12-18.

40. Collins T, Wallman J: **The relative importance of retinal error and prediction in saccadic adaptation**. *J Neurophysiol* 2012, **107**:3342-3348.

41. Kaku Y, Yoshida K, Iwamoto Y: **Learning signals from the superior colliculus for adaptation of saccadic eye movements in the monkey**. *J Neurosci* 2009, **29**:5266-5275.

42. Hamker FH *et al.*: **The peri-saccadic perception of objects and space**. *PLoS Comput Biol* 2008, **4**:e31.

43. Ziesche A, Hamker FH: **A computational model for the influence of corollary discharge and proprioception on the perisaccadic mislocalization of briefly presented stimuli in complete darkness**. *J Neurosci* 2011, **31**:17392-17405.

44. Srimal R *et al.*: **Obligatory adaptation of saccade gains**. *J Neurophysiol* 2008, **99**:1554-1558.

45. Mathys C *et al.*: **A bayesian foundation for individual learning under uncertainty**. *Front Hum Neurosci* 2011, **5**:39.

46. Kording KP, Tenenbaum JB, Shadmehr R: **The dynamics of memory as a consequence of optimal adaptation to a changing body**. *Nat Neurosci* 2007, **10**:779-786.

47. Baldi P, Itti L: **Of bits and wows: a Bayesian theory of surprise with applications to attention**. *Neural Netw* 2010, **23**:649-666.

48. Friston K *et al.*: **Perceptions as hypotheses: saccades as experiments**. *Front Psychol* 2012, **3**:151.

49. Engbert R *et al.*: **An integrated model of fixational eye movements and microsaccades**. *Proc Natl Acad Sci U S A* 2011, **108**:E765-E770.

50. Roberts JA, Wallis G, Breakspear M: **Fixational eye movements during viewing of dynamic natural scenes**. *Front Psychol* 2013, **4**:797.

51. Roberts EO *et al.*: **Reading in schizophrenic subjects and their nonsymptomatic first-degree relatives**. *Schizophr Bull* 2013, **39**:896-907.

52. Engbert R *et al.*: **SWIFT: a dynamical model of saccade generation during reading**. *Psychol Rev* 2005, **112**:777-813.

53. Reichle ED, Rayner K, Pollatsek A: **The E-Z Reader model of eye-movement control in reading: comparison to other models**. *Behav Brain Sci* 2003, **26**:445-526.

54. Gold JM *et al.*: **Visual orienting in schizophrenia**. *Schizophr Res* 1992, **7**:203-209.

55. Vossel S *et al.*: **Spatial attention, precision, and Bayesian inference: a study of saccadic response speed**. *Cereb Cortex* 2014, **24**:1436-1450.

56. • Vossel S *et al.*: **Cholinergic stimulation enhances Bayesian belief updating in the deployment of spatial attention**. *J Neurosci* 2014, **34**:15735-15742.
This paper demonstrates how a generative model for saccadic RTs can serve to estimate cholinergic drug effects on model parameters determining learning under uncertainty.

57. Vossel S *et al.*: **Cortical coupling reflects Bayesian belief updating in the deployment of spatial attention**. *J Neurosci* 2015, **35**:11532-11542.

58. McDowell JE *et al.*: **Neurophysiology and neuroanatomy of reflexive and volitional saccades: evidence from studies of humans**. *Brain Cogn* 2008, **68**:255-270.

59. Lencer R, Trillenberg P: **Neurophysiology and neuroanatomy of smooth pursuit in humans**. *Brain Cogn* 2008, **68**:219-228.

60. Adams RA *et al.*: **The computational anatomy of psychosis**. *Front Psychiatry* 2013, **4**:47.

61. Andreasen NC, Paradiso S, O'Leary DS: **"Cognitive dysmetria" as an integrative theory of schizophrenia: a dysfunction in cortical–subcortical–cerebellar circuitry?** *Schizophr Bull* 1998, **24**:203-218.

62. Schall JD: **Visuomotor functions in the frontal lobe**. *Annu Rev Vis Sci* 2015.

63. Krauzlis RJ: **Recasting the smooth pursuit eye movement system**. *J Neurophysiol* 2004, **91**:591-603.

64. Stephan KE *et al.*: *Charting the Landscape of Priority Problems in Psychiatry, Part 1: Classification and Diagnosis*. Lancet Psychiatry; 2015.

65. Aponte EA *et al.*: **mpdcm: a toolbox for massively parallel dynamic causal modeling**. *J Neurosci Methods* 2015, **257**: 7-16.

66. Moran RJ *et al.*: **An in vivo assay of synaptic function mediating human cognition**. *Curr Biol* 2011, **21**:1320-1325.

67. Nassar MR *et al.*: **Rational regulation of learning dynamics by pupil-linked arousal systems**. *Nat Neurosci* 2012, **15**: 1040-1046.

68. Preuschoff K, t Hart BM, Einhauser W: **Pupil dilation signals surprise: evidence for noradrenaline's role in decision making**. *Front Neurosci* 2011, **5**:115.

69. Michell AW *et al.*: **Saccadic latency distributions in Parkinson's disease and the effects of L-dopa**. *Exp Brain Res* 2006, **174**:7-18.

70. Schmechtig A *et al.*: **Effects of risperidone, amisulpride and nicotine on eye movement control and their modulation by schizotypy**. *Psychopharmacology (Berl)* 2013, **227**:331-345.

Chapter 5

In this chapter, we devote our attention to saccadic adaptation (SA) (McLaughlin, 1967), a form of oculomotor learning. Our interest in this phenomenon stems from two complementary theories of some of the symptoms that characterize schizophrenia: On the one hand, (Stephan et al., 2009a) proposed that hallucinations and perceptual disturbances in schizophrenia might be due to erroneous integration of predictions of the perceptual effects of self-generated actions. This type of predictions is thought to further depend on corollary discharge (CD), i.e., predictions of the motor effects of voluntarily generated actions. In the case of the oculomotor system, this theory postulates that if CD is compromised, visual percepts should appear fragmented, and self-generated eye movements could be perceived (at least partially) as generated by external causes. Thus, abnormal CD could partially explain some of the positive symptoms in schizophrenia: hallucinations and delusions of control.

Representatively, in a classical first-person report, a patient describes the following experience:

> "One will in any case not dispute that I must know myself whether my eyes are pulled towards an indifferent object or whether I look at something interesting around me of my own will. […] For instance, I notice almost daily that when I look for a book amongst my books or for certain scores or if I am searching for a small object (a needle or a pair of scissors, etc.), which I do not notice momentarily because it is so small, miracles direct my gaze (turn my eyes) to the desired object. This phenomenon, the reality of which cannot be doubted, is in my opinion of absolutely fundamental importance for the knowledge of divine qualities and forces. Two conclusions can be drawn from it: firstly that the rays (and I know this from many other reasons too) are able to read my thoughts (how could they otherwise know what I am

> looking for at the moment); secondly that they must be aware of where the looked for object is; in other words the place where such an object is is seen by God with the help of sunlight with much greater certainty and perfection than by human beings with their eyes" (Schreber, 1955 in Maes and Van Gool, 2008).

Beyond the possibility of impaired CD in schizophrenia, Stephan et al. (2006) proposed that dysfunction of the NMDA receptor mediated synaptic plasticity in cortical regions could lead to some of the cognitive deficits observed in this disease.

Based on these ideas, we proposed in Chapter 4 to investigate tasks that require both oculomotor CD as well as synaptic plasticity. Concretely, we suggested to use SA, a task in which the magnitude of saccades to a target changes when the target is systematically displaced. The reason to investigate SA is that shifting the target of a saccade triggers a prediction error between its expected location and the postsaccadic perceptual input.

Oculomotor CD has been well described in the macaque brain (Sommer and Wurtz, 2002), in which a re-entrant pathway connects the inferior layers of the superior colliculus to the frontal eye fields through the medio-dorsal thalamus. This thalamocortical connection is known to carry information about the magnitude of saccadic eye movements. Notoriously the discovery of this pathway was first reported in a human study (Gaymard et al., 1994) that showed that lesions of the thalamus lead to decreased accuracy in a task that requires motor predictions but not in a task that requires the generation of visually guided saccades. Thus, this pathway is likely to exist both in human and non-human primates.

Surprisingly, there is scant evidence connecting oculomotor CD and SA, despite the theoretical considerations above. Hence, the present study is concerned with the following question: Is there evidence of the involvement of the CD pathway during SA? This question is investigated with the help of a computational model of SA in combination with fMRI. Our results provide evidence that the hypothesized areas (superior colliculus, thalamus, and frontal eye fields), as well as regions located in the parietal cortex are activated during saccadic adaptation.

Cortical processing and corollary discharge during saccadic adaptation

Eduardo A. Aponte[1,*], *Klaas E. Stephan*[1,2,3], *Jakob Heinzle*[1]

[1] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich. Wilfriedstrasse 6, 8004, Zurich, Switzerland.

[2] Wellcome Trust Centre for Neuroimaging, University College London. 12 Queen Square London WC1N 3BG,.

[3] Max Planck Institute for Metabolism Research. Gleueler Strasse 50, 50931, Cologne, Germany.

[*]Corresponde: Eduardo A. Aponte; Jakob Heinzle

Abstract

Saccadic adaptation is a form of motor plasticity that can be triggered by systematically stepping the visual target of a saccade during its execution. While there is agreement on the important role of the cerebellum, a possible cortical involvement is still debated. Despite indirect evidence that corollary discharge and prediction errors are engaged in saccadic adaptation, the origin of the teaching signal for adaptation remains unclear. Here, we investigated the potential role of visual prediction errors, combining computational modelling and neuroimaging. We first employed Bayesian modelling to compare competing explanations of saccadic adaptation. We then performed functional magnetic resonance imaging (fMRI) analyses of putative prediction error activity, focusing on cortical and subcortical sources outside the cerebellum. Visual prediction errors across saccades caused activation of regions known to mediate corollary discharge in the monkey, including frontal eye fields, superior colliculus, and thalamus. Prediction error related activation in the right intraparietal sulcus correlated with the amount of adaptation as indexed by Bayesian model comparison, and activity in the left superior parietal lobule could be predicted by a model-based estimate of the saccadic length of the motor command. These imaging findings relate saccadic adaptation to the known pathway for corollary discharges of saccadic eye movements and highlight the possible involvement of parietal regions in saccadic adaptation.

Introduction

Saccadic adaptation (SA) is a form of motor plasticity in which saccadic gain is progressively modified by error signals (McLaughlin, 1967). It can be triggered by stepping the target of a saccade from its original position to a new location during the saccade (Hopp and Fuchs, 2004; Pelisson et al., 2010; Herman et al., 2013). This mismatch between pre- and post-saccadic visual targets changes saccadic lengths such that saccades increasingly approach the post-saccadic visual target. However, the exact nature of the teaching signal that triggers SA has not been fully elucidated.

There are at least two hypotheses about what error signal triggers SA: the *retinal error* model and the *prediction error* model. The first states that the error signal minimized during SA is the post-saccadic difference between fovea and target location. According to the second explanation, oculomotor plasticity is triggered by a *prediction error* (Wong and Shelhamer, 2011), the difference between internally generated predictions of post-saccadic target locations and visual feedback. In other words, this hypothesis postulates that SA reflects a form of learning that minimizes prediction errors (PE). Crucially, this requires the existence of predictive signals of the post-saccadic foveal location, so-called corollary discharges (CD; Sommer and Wurtz, 2008).

Corollary discharges (Sperry, 1950) – or efference copies (von Holst and Mittelstaedt, 1950) – are internal, re-entrant motor signals that affect sensory processing, motor planning, and learning. In the macaque brain, CD signals of saccadic eye movements are transmitted from the superior colliculus (SC) to the frontal eye fields (FEF) via the mediodorsal thalamus (Sommer and Wurtz, 2002, 2004b, a). However, to date there is no conclusive evidence that thalamo-cortical CDs are used for SA. In humans, CDs for saccadic eye movements are believed to guide behaviour when no visual information is available (Gaymard et al., 1994) and to be relevant for detection of perisaccadic changes of visual input (Collins et al., 2009). More recent studies suggest that CDs are also important during SA (Wong and Shelhamer, 2011; Collins and Wallman, 2012), because retinal errors alone (i.e., the post-saccadic difference between the foveal position and the target) cannot fully explain SA.

In summary, transsaccadic PEs could serve as the teaching signals that trigger SA. These require predictions of future sensory input that are compared to visual feedback. Since proprioception only has a minor role in SA (Lewis et al., 2001), these sensory predictions presumably require CD. Although there is conclusive evidence that predictive signals accompany saccadic eye movements, there is no direct evidence that these predictions are involved in SA.

Here, we used a formal statistical model, inspired by previous work (Kording et al., 2007; Srimal et al., 2008), to compare competing explanations of SA as a form of motor plasticity generated by PE. Second, we examined neural activations by PE during SA. In particular, we tested whether the FEF, the thalamus, and the SC are involved in SA and PE signalling. Finally, we used model-based estimates of internal motor commands (saccade length) and of the amount of adaptation in order to test for SA-related activity. Our findings suggest that the SC-Thalamus-FEF circuit is activated during SA in humans. Hence, this circuit is possibly involved in PE signalling. In addition, activations in parietal regions are correlated with model-based estimates and thus seem to be more directly linked to SA.

Materials and methods

Subjects

Thirty-two subjects participated in the experiment (mean age: 23, range: 19-28, female/male: 13/15). Eight subjects were excluded because of clear signs of somnolence (1), technical problems related to the eye tracking device (6), and incomplete data (1). All experimental procedures were approved by the ethics board of the canton of Zurich (Approval: KEK-ZH-Nr. 2010-0327). All subjects gave written informed consent prior to participating in the study.

Experimental procedure

The experiment was performed at the Laboratory for Social and Neural Systems at the University of Zurich. Subjects lay in supine position inside an MRI scanner in a dimly lit room. Participants viewed a horizontal screen ($58 \times 34.5\ cm$) at the rear of the scanner through a mirror outside the head coils (distance to screen approx. $125\ cm$). Horizontally, the screen subtended 25.1 degrees of visual angle (dva). Visual stimuli were

displayed using a Sanyo DPG-DWT50L projector, with a refresh rate of 60Hz and a resolution of 1280×1024 pixels.

Gaze location was measured using an infrared light eye tracker (Eyelink 1000, SR Research, Mississauga, Ontario, Canada). Saccades were detected online by a dedicated computer with a built-in algorithm (Stampe, 1993) provided by the manufacturer, which uses a saccade velocity threshold of $22\,dva/s$ and an acceleration threshold of $3800\ dva/s^2$ to detect saccades. Fixations were defined as constant gaze to a window of $2\ dva$ for at least $500\ ms$. Eye position was stored at a sampling rate of $250Hz$. Calibration was performed using a five-cue array (calibration points at centre, and $6\ dva$ to north, east, south and west). We aimed at a calibration error lower than $1\ dva$.

The visual gaze target was a red circle of $0.25\ dva$ on a gray background. Trials started with a fixation period with the fixation target displayed at $6\ dva$ to the left of the center of the screen. Once the target was fixated for at least $500\ ms$ and after an additional random delay of 0 to $1000\ ms$ (uniformly distributed), the target was stepped $12\ dva$ to the right, that is $6\ dva$ to the right of the centre of the screen. Subjects were instructed to fixate the circle and to saccade to the target as soon as it changed its position (Fig. 1A). No further instructions were provided. A trial was restarted if the subject failed to maintain fixation during the fixation interval and discarded if no saccade was performed within $800\ ms$ after the target was stepped.

Experimental design

Since we were interested in manipulating adaptation, we followed an experimental design similar to Gerardin et al. (2012). The 2x2 factorial design had one factor - the probability of stepping the target - controlling how often saccadic adaption would happen, and a second factor - the delay of the presentation of the post saccadic target - controlling the strength of adaptation (Bahcall and Kowler, 2000; Fujita et al., 2002). In other words, the two factors of the design aimed at modulating the frequency and the strength of adaptation. Each of the four blocks of the design consisted of 111 trials, starting with 16 *pre-adaptation* trials, followed by 50 *adaptation* trials. Sixteen *post-adaptation* trials were added to compare saccadic length before and after the experimental manipulation. In order to increase the efficiency of the fMRI design, *fixation* trials were interleaved randomly throughout each block.

In *pre-* and *post-adaptation* trials the target vanished once a saccade towards the main target was detected and crossed a 3 *dva* threshold. In these trials, the target did not reappear, and the screen remained blank for 500 *ms* until the start of the next trial.

Adaptation trials were randomly divided into *step* and *no step* trials. In all *adaptation* trials, the saccade target vanished at the time of saccade detection and reappeared either after 16 ms (*no delay* condition) or 500 *ms* (*delay* condition). In *step* trials (Fig. 1B, top) the saccadic target was shifted 3 *dva* to the left to induce backward adaptation. In *no step* trials (Fig. 1B, bottom) it reappeared at its original location.

We implemented the two factorial design in the following way. The first factor, the *proportion of step and no step trials,* was defined by the number of *step* trials as a percentage of the total number of *step* and *no step* trials in a block. Each block had either 44 or 6 *step* trials randomly interleaved with the 6 or 44 *no step* trials. The second factor, *the delay of the presentation of the post saccadic target,* was defined by the delay of the reappearance of the target (16 or 500 *ms*). For simplicity, we refer to the blocks with a short and long delay as the *delay* and *no delay* blocks, and to the 90% *step (~44/50)* and 10% *step (~6/50)* trial blocks as the *step90* and *step10* blocks, respectively. The order of presentation of the conditions (i.e. blocks) was pseudo-randomized and counterbalanced across subjects.

During *fixation* trials, the fixation target remained at the original location for either 2.5, 3.0, or 3.5 *s*, and then vanished, before the next trial started. The exact temporal pattern of the fixation trials was changed after 6 of 24 participants. Most (18) participants saw 29 fixation trials randomly distributed over the run. For the first six subjects we included 19 fixation trials and either 3 pauses of 10 seconds (2 subjects), or 20 fixations trials and 2 long pauses of 10 seconds (4 subjects). In these participants, the number of pre- and post-adaptation was each 15. Crucially, the number and structure of adaptation trials, which are the only trials we analysed here, was identical in all participants.

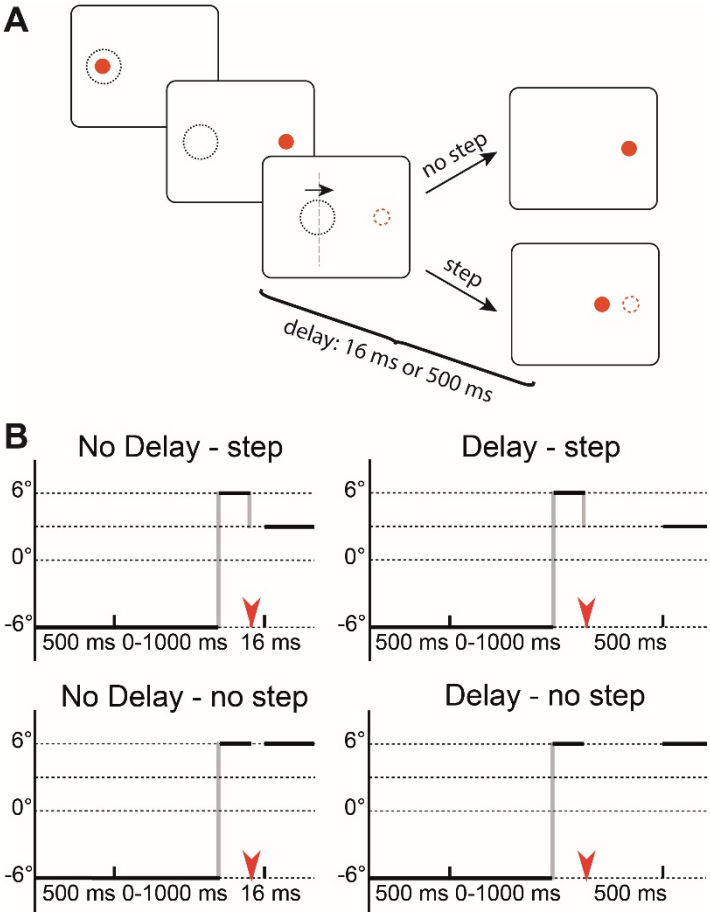


Figure 1: Illustration of the task showing the sequence of stimuli. A. and their timing **B**. At the beginning of each trial, participants fixated the red dot indicated by the dashed circle at the left side of the screen (-6 dva). The fixation target was displayed for 500 to 1500 ms after start of the trial. If subjects failed to fixate during this interval, the trial was restarted. Following the fixation interval, the target was stepped 12 dva to the right. Once a saccade was detected and gaze crossed a 3 dva threshold from the fixation target (timepoint depicted by the red arrow head in B), the saccadic target vanished. The post-saccadic target was displayed again with either a 16 ms or 500 ms delay. In step trials, the target was stepped 3 dva to the left. In no step trials, the target reappeared at its original location. Note that in B, time lengths are not proportional to their true duration.

In order to minimize adaptation effects between blocks, a de-adaptation procedure was introduced between scanning blocks. Subjects were instructed to follow a target that was stepped 50 times between −6 and 6 *dva* to the centre of the screen.

Statistical analysis

Analysis of behavioural data

In this section, we first describe the classical analysis of the behavioural data based on pre- and post-adaptation trials. We then introduce a generative model of SA that allows for a more detailed analysis of the behaviour. Only trials with a saccade length between 8 *dva* and 14.5 *dva* were used for analysis. Trials were rejected if a blink was detected after the fixation target was stepped or if there was a saccade before the target was stepped.

Classical analysis

In order to assess the effect of the experimental manipulation on saccadic length, we entered the average saccadic length difference between *pre-* and *post-adaptation* into a 2-by-2 within subject ANOVA test where the independent variables were *subject, delay* and *step* probability. This analysis was conducted in MATLAB 2014a (8.3.0), with the Statistics Toolbox (9.0). We tested for both main effects and their interaction.

Model based analysis of saccadic adaptation

A model-based analysis was used to refine the behavioural analysis. Here, we first provide a concise mathematical description, while in the next section we will give an interpretation of the update equation in terms of visual errors and CD. The SA model was inspired by previous work (Kording et al., 2007; Srimal et al., 2008).

We refer to the pre- and post-saccadic locations of the target in trial t as T_t^0 and T_t^1, respectively. SA was modelled under the assumption that subjects maintain an internal representation of the motor command $\hat{s}$ associated with a particular retinotopic presaccadic eccentricity T^0. The model describes the horizontal magnitude of the motor command $\hat{s}$. Thus, we assumed a mapping $\hat{s}(T_t^0)$ between target eccentricity and the magnitude of saccades towards that target. Because our design had a single pre-saccadic target eccentricity, we drop any dependency of $\hat{s}$ on T_t^0 and refer to $\hat{s}(T_t^0)$ simply as $\hat{s}_t$. Furthermore, we assumed that saccades are the result of both a motor command $\hat{s}_t$ and motor error ε_t^m. The magnitude of a saccade s generated at trial t is equal to

$$s_t = \hat{s}_t + \varepsilon_t^m. \quad (1)$$

Errors ε_t^m were assumed to be i.i.d. Gaussian distributed with mean 0 and variance σ^2. Thus, the likelihood of a saccade of magnitude s at trial t was assumed to be Gaussian, with mean $\hat{s}_t$ and variance σ^2. In order to account for plasticity, motor commands $\hat{s}_t$ were assumed to change across trials t. During *pre-* and *post-adaptation* phases we assumed no adaptation took place (Rolfs et al., 2010) due to lack of post-saccadic perceptual feedback and thus:

$$\hat{s}_t = \hat{s}_{t-1}. \tag{2}$$

During adaptation, the trialwise update equation with adaptation rate $\alpha > 0$ was given by:

$$\hat{s}_{t+1} = \hat{s}_t + \alpha(T_t^1 - \hat{s}_t). \tag{3}$$

Here, the teaching signal that drives SA is the difference between the motor command $\hat{s}_t$ and the postsaccadic location of the target T_t^1. We refer to this quantity as prediction error throughout this paper.

Finally, we finessed the model in order to account for two experimental findings. First, adaptation often plateaus before reaching the stepped target (Collins et al., 2009; Picard et al., 2012). This can be accounted for by introducing a scaling parameter γ_b for the stepped location T_t^1. Second, forward adaptation can differ in several aspects from backward adaptation (Hopp and Fuchs, 2004), which we took into account by introducing a separate adaptation rate and scaling factor for forward and backward adaptation, respectively

$$\hat{s}_{t+1} = \hat{s}_t + f(\hat{s}_t, T), \tag{4}$$

$$f(\hat{s}, T) = \begin{cases} \alpha_b(\gamma_b T - \hat{s}) \;\; if\; T = T_t^1, \\ \alpha_f\left(\gamma_f T - \hat{s}\right) \;\; if\; T = T_t^0. \end{cases} \tag{5}$$

The full model is thus controlled by six parameters: backward and forward adaptation rates α_b and α_f, backward and forward scaling γ_b and γ_f, the initial saccadic command $\hat{s}_0$, and the error variance σ^2. The scaling parameter γ_f is the gain of the saccade towards a visual target, while the backward scaling γ_b is related to the adaptation gain (the relative amount of adaptation) by $\gamma_b^* = (\gamma_b T^1 - T^0)/(T^1 - T^0)$. Fig. 2 shows a graphical representation of the model and illustrates the effect of the parameters associated to backward adaptation in simulations of the model.

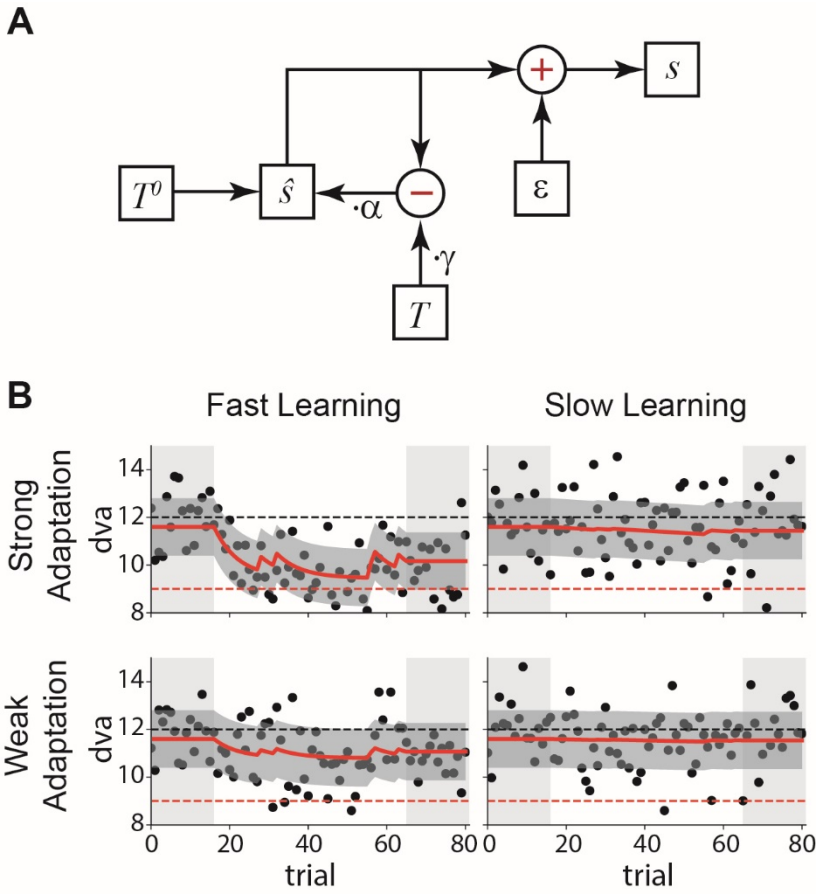


Figure 2: **A. Graphical illustration of the SA model.** The internal motor command is directly compared to the target location and updated accordingly. **B. Illustration of effect of model parameters on adaptation**. Simulations of the SA model with 4 different parametrizations (top left: $\alpha_b = 0.15, \gamma_b^* = 0.85$; top right: $\alpha_b = 0.005, \gamma_b^* = 0.85$; bottom left: $\alpha_b = 0.15, \gamma_b^* = 0.4$; bottom right: $\alpha_b = 0.005, \gamma_b^* = 0.4$) with 90% step trial probability. As in the experiment, T_0 was assumed to be 12 dva (black dashed line). In the step trials, T_1 was equal to 9 dva (red dashed line). The time course of the motor command $\hat{s}$ is depicted by the solid red line. Simulated saccades are depicted as black circles. The dark grey area displays one standard deviation around $\hat{s}$. Pre- and post-adaptation phases are highlighted by light grey. During the pre- and post-adaptation phase, it is assumed that no adaptation takes place.

Visual errors and corollary discharge

In this section, we briefly discuss the assumption about the learning model and illustrate how a CD about the saccadic command can be combined with the post-saccadic visual error to yield the learning rule implemented in the model. For simplicity we assume in this derivation that the two scaling constants γ_a, γ_d are equal to one. After each saccade

the post-saccadic *visual error* ε_t^v is given by the difference between the foveal location s_t and the target T_t^1:

$$\varepsilon_t^v = T_t^1 - s_{t,} \quad (6)$$

$$= T_t^1 - \hat{s}_t - \varepsilon_t^m. \quad (7)$$

Fundamentally, we assumed that it is not the visual error that is used as teaching signal to update the motor command $\hat{s}_t$, but the trans-saccadic visual PE, i.e. the mismatch between the expected location of the visual target after the saccade and its true location. Mathematically, this corresponds to the difference between the visual error and the predicted motor error $(\varepsilon_v - \varepsilon_m)$. The predicted motor error is the difference between the actual saccade s_t and the associated internal motor command $\hat{s}_t$:

$$s_t - \hat{s}_t = \varepsilon_t^m. \quad (8)$$

Note that the motor error can be easily computed if the brain receives a CD of the actual saccade s_t which can be compared to the internal motor command $\hat{s}_t$. In other words, the visual system requires an accurate prediction of the magnitude of a saccade s_t generated in a trial given by a CD signal.

Under this model, the trial wise update equation with adaptation rate $\alpha > 0$ is given by:

$$\hat{s}_{t+1} = \hat{s}_t + \alpha(\varepsilon_t^v - \varepsilon_t^m), \quad (9)$$

$$= \hat{s}_t + \alpha(T_t^1 - \hat{s}_t). \quad (10)$$

This formulation of the learning rule shows that both the post-saccadic visual error as well as a precise CD (and thus knowledge of the motor error) are required for SA in accordance with the model.

Model space

In order to test which set of parameters provided the most parsimonious explanation of the data, we defined a series of nested models. Starting from the full model with six free parameters, eight models were defined by fixing a subset of the parameters to particular values. A list of all models is presented in Table 1.

In the full model m_1, the adaptation rate and scaling were allowed to differ between the two directions of adaptation. In model m_2, we assumed that the adaptation rate was equal in both directions $(\alpha_b = \alpha_f)$, while in models m_3 and m_4 the adaptation scale was fixed to 1. In

addition, model m_4 (as model m_2) assumed that the adaptation rate was equal during the *step* and *no step* trials. In models m_5 and m_6, we assumed that in *no step* trials subjects adapted back to their baseline $\hat{s}_0$, or equivalently that $\gamma_f = \hat{s}_0/T_t^1$. In model m_6, the adaptation rates α_b and α_f were equal. In model m_7, no adaptation took place in the *no-step* trials, which corresponds to $\alpha_f = 0$. Finally, in the null model m_8, no SA was allowed by fixing $\alpha_b = \alpha_f = 0$. m_8 represents the null hypothesis that no adaptation took place in a block.

	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8
Back. update	T^1	T^1	T^1	T^1	$\hat{s}_0$	$\hat{s}_0$	N.A.	N.A.
α_b	✓	✓	✓	✓	✓	✓	✓	0
α_f	✓	α_b	✓	α_b	✓	α_b	0	0
γ_b	✓	✓	1	1	✓	✓	✓	N.A.
γ_f	✓	✓	1	1	$\hat{s}_0/T_t^1$	$\hat{s}_0/T_t^1$	N.A.	N.A.
$\hat{s}_0$	✓	✓	✓	✓	✓	✓	✓	✓
σ^2	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Behavioral model space. Free parameters are designated by "✓". All models shared the same prior distributions. N.A. stands for "does not apply".

Specification of a full generative model requires setting a prior distribution on the parameters. Table 2 summarizes the priors used here.

Parameter	Probability density function
σ	$\Gamma(\sigma; 1.5, 0.5)$
s_0	$N(s_0; 12.0, 2.0)$
α_b, α_f	$U_{[0,1]}(\alpha)$
γ_a, γ_d	$\Gamma(\gamma; 12.8, 16.0)$

Table 2: Prior probability density functions. $N(y; \mu, \sigma^2)$ corresponds to a Gaussian distribution with mean μ and variance σ^2. $U_{[0,1]}(\alpha)$ corresponds to the uniform distribution in the interval [0,1]. $\Gamma(t; \alpha, \beta)$ corresponds to the gamma distribution $(\beta^{-\alpha}/\Gamma(\alpha))\, t^{\alpha-1} \exp -t/\beta$, where $\Gamma(\alpha)$ is the gamma function.

Inference

To quantitatively compare models with and without adaptation, we relied on Bayesian inference (Gelman et al., 2013), a family of statistical methods based on Bayes' formula:

$$\underbrace{p(\theta|y,u,m)}_{posterior} = \frac{\overbrace{p(y|\theta,u,m)}^{likelihood}\overbrace{p(\theta|m)}^{prior}}{\underbrace{p(y|u,m)}_{model\ evidence}} \tag{11}$$

where θ is the vector of model parameters, y represents the observed experimental data, u encodes the experimental design, m is the model used to describe the data, and p denotes the appropriate conditional probability density functions.

Two common goals of Bayes inference are to (i) compute the *posterior* distribution $p(\theta|y,u,m)$, which describes the probability of the parameters conditioned on a given model, an experimental design or condition u, and observed data y, and (ii) the model evidence or marginal likelihood $p(y|u,m)$, which describes the probability of the data given a model. Two models *i* and *j* can then be compared using the Bayes factor:

$$B_{i,j} = \frac{p(y|u,m_{i,})}{p(y|u,m_j)}. \tag{12}$$

Bayes factors can be used to compare competing models on the same data set, and it is a complementary method to for example model cross validation (Kass and Raftery, 1995; MacKay, 2003; Gelman et al., 2013). Throughout this paper we will us log Bayes Factors which are equal to

the difference in log model evidence (LME): $log\, B_{i,j} = log \frac{p(y|m_i)}{p(y|m_j)} = \Delta LME_{i,j}$. For statistics on the group level, we use a fixed effect analysis, that is, we summed the LMEs across subjects. This tests the evidence that our experimental manipulation was successful. Following recommended interpretations, we consider $\Delta LME > 3$ strong evidence (Kass and Raftery, 1995). This roughly corresponds to a Bayes factor of 20. We only report results for which we obtained strong evidence.

A sampling algorithm was used to approximate the posterior distribution of each model. It was implemented in *python* 2.7, using the libraries *numpy* (1.8) and *scipy (0.13)*. 3×10^4 samples from the target distribution were drawn using the Metropolis-Hastings algorithm and a Gaussian proposal distribution. Prior to this, 10^4 samples were drawn and discarded in the so-called burn-in phase. Samples of the learning rate were drawn from $\alpha^* = \tan \pi(\alpha - 0.5)$ and then transformed to the open interval]0, 1[. The sampling algorithm was further refined by adapting the proposal distribution during the burn-in phase, following Shaby and Wells (2010). In order to compute the marginal likelihood of each model, we relied on thermodynamic integration (Gelman and Meng, 1998; Gelman et al., 2013), a technique that provides an accurate estimate of the marginal likelihood of a model. This is achieved by first constructing a smooth path between the prior and posterior distributions and by then integrating along this path. In order to increase the statistical efficiency of this technique, an exchange operator between chains was implemented (Swendsen and Wang, 1986; Aponte et al., 2016).

Imaging

The experiment was performed on a 3T Philips Achieva MR Scanner with an eight-channel head coil. T_2^* weighted images were acquired using a gradient-echo EPI sequence with the following parameters: slice thickness: $2.5\, mm$; in-plane resolution: $2.5 \times 2.5\, mm$; interslice gap: $0.5\, mm$; ascending continuous in-plane acquisition; $TR = 2000\, ms$; $TE = 36\, ms$; flip angle $= 90^o$; field of view $= 192 \times 192 \times 98.5\, mm$; SENSE factor = 2; EPI factor = 41. T_1 anatomical images were acquired for structural preprocessing with the following parameters: resolution $1 \times 1 \times 1\, mm$, $TR = 8.3\, ms$, $TE = 3.9\, ms$, field of view $= 256 \times 256 \times 181\, mm$, flip angle $= 8°$. The main goal of the present study was to investigate cortical correlates of SA. Thus, the scan volume was set to

cover the entire cortex. Since we aimed for a relatively high temporal (TR = 2s) and spatial (2.5 mm isotropic) resolution it was not possible to increase the scan volume and reliably include the cerebellum for all head sizes. Pulse and respiration were concurrently measured using electrocardiography and a breathing belt. These physiological measures served as confound regressors as explained below.

fMRI Analysis

Data analysis was performed with SPM8 (v5236). Preprocessing of functional images included realignment, coregistration of the EPI images to the T1 image and normalization and segmentation of the T1 image using SPMs New Segment function with subsequent normalization of all EPI images. Normalization included resampling of the images to 1.5mm isotropic resolution. Finally, the images were smoothed with an isotropic Gaussian kernel of FWHM=6mm.

Subject specific data were modelled at the first level using a general linear model (GLM). Four main regressors were defined for each experimental run, one for each trial type: (i) *step* and (ii) *no step* trials, and (iii) *pre-* and (iv) *post-adaptation* trials. Trials were modelled as delta functions aligned to the end of the main saccade of each trial and then convolved with the hemodynamic response function and its time derivative. A nuisance regressor that modelled the onset of all saccade trials that were excluded from the analysis (see below) was added. We also included head movements (as encoded by realignment parameters), and respiration and heartbeat regressors according to RETROCIOR (Glover et al., 2000), as constructed by the PHYSIO toolbox (Kasper et al., 2017; version 1.5).

Because we manipulated the probability of stepping, the number of step and no step trials was different within each run (90/10 or 10/90 ratio). In order to reduce potential differences between beta values due to large differences in trial number, the group level analysis was restricted to the most common trial type in each block. For example, for the *step90* conditions, only the β images of the *step* trials were included in the group level analysis. Similarly, only the β images of the *no step* trials were used for the *step10* conditions. This yielded a 2 × 2 factorial design with levels *delay* vs. *no delay* and *step90* vs. *step10*. In the statistical analysis, we assessed the two main effects (step and delay) as well as their interaction. We were particularly interested in the effects of stepping the

target because precisely in this condition the predictions conveyed by CD do not match the visual outcome, which according to our hypothesis should trigger PEs. Thus, we assessed PE responses by means of a categorical contrast on the conditions described above, comparing step trials during the *step90* condition (when a PE occurs) to non-step trials during the step10 condition (when there is no PE). Moreover, we were interested in the interaction between the *step* and *delay* conditions, because the delay condition is reported to abolish or diminish the amount of adaptation even under the presence of a visual prediction error. Thus, this contrast could potentially disentangle prediction errors and saccadic adaptation.

Model based fMRI

We used the above generative model of saccades in a model-based trial-by-trial analysis of fMRI data in two ways. First, we included the (mean-corrected) predicted trace of the internal estimate of saccadic length $\hat{s}_t$ as a parametric modulation in the first level GLM. We computed a contrast representing the mean β value for the parametric regressor over the two step90 conditions (*no-delay* and *delay*) and then assessed on the second level whether this parameter was significantly larger than zero.

Furthermore, we use the model to investigate whether the difference in neural activation between the two stepping conditions ((*step90+no-delay*) – (*step90+delay*)) was correlated with difference in adaptation between the two conditions. The model directly provides a measure of adaptation in terms of the log Bayes Factor between the best adaptation model and the no-adaptation model (m_8). We thus correlated the difference of this measure and the difference of activation between the two runs.

fMRI statistics

All second level tests were performed using cluster level FWE correction at $p<0.05$ with a cluster defining threshold of $p=0.001$ (Flandin and Friston, 2016). In addition to the whole brain analysis we also assessed small volume corrected (S.V.C.) statistics within regions of interest in the CD pathway. In particular, we used a bilateral ROI of the superior colliculus as in (Limbrick-Oldfield et al., 2012) and masks for the left and right thalamus connected to prefrontal cortex (Behrens et al., 2003). Within ROIs we applied FWE correction at the voxel level. In order to take into account multiple testing for the three ROIs we adjusted the

significance level to $\alpha<0.0166$. Finally, in the model-based analysis, we restricted the analysis of activity related to the internal estimate of saccadic length to those regions that showed a significant effect of step using the result of the factorial analysis as a mask. Please note that we orthogonalised the model based regressor to the main regressor of step. This assures that any significant finding is purely driven by fluctuations in the model based regressor and cannot be explained by the mask defining contrast. We used the Anatomy Toolbox (version 2.2c) for SPM (Eickhoff et al., 2007) to confirm anatomical labelling.

Results

In this section, we first report the behavioural and modelling results and then proceed to the imaging findings.

Behavioral results

A total of 7803 primary saccades were analysed, from which roughly 20% (1593) were discarded, because of blinking, aborted trial, no saccade before 800ms, or length (below 8.5 dva or above14.0 dva). Careful inspection of saccade timing revealed that 416 trials were incorrectly detected, triggering the presentation of the post-saccadic target not in time with the saccade. These trials were also excluded from the analysis. The mean time between the start of a saccade and the command for vanishing the target was 25.5ms (std. 3.0 ms).

Classical analysis

We submitted the mean saccadic difference between the pre- and post-adaptation phases to a within subject ANOVA. While the *step* factor had a significant effect ($F_{24,1} = 7.49,\ p < 0.05$), on the difference between phases, neither the *delay* factor ($F_{1,24} = 1.0, p = 0.32$) nor the interaction ($F_{1,24} = 0.71, p = 0.40$) was significant.

Modelling saccadic adaptation – simulations

As a first step, we verified the reliability of our inference scheme by generating data from each model through random sampling from the prior distributions and then computing the evidence of each of the models for each simulation. Twenty data sets were generated from each model under both the *step10* and *step90* conditions, with the resulting LME being added across all the data sets. This yielded an 8×8 matrix whose cells were the marginal likelihood of each model, given data

generated by one of the eight possible models. The results are summarized in Table 3. Please note that we have considered a complicated case here, for two reasons. First, we included both conditions, *step10* and *step90*, in this simulation. In the *step10* condition, it is more difficult to disambiguate models, because only few trials could induce adaptation. Second, we sampled the parameters from the prior, which leads to many instances where the parameter of a more complicated model is chosen to be the fixed value of a simpler model which will then exceed in explaining the data. Nevertheless, we found that all models but m_3 were clearly identifiable. When data were generated with model m_3, Bayesian model comparison could not distinguish between models m_1, m_2 and m_3.

		Model used for inversion							
		m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8
Data generating model	m_1	440.4	436.3	380.0	314.6	386.6	340.9	200.0	0.0
	m_2	702.1	719.4	447.3	450.6	419.0	347.6	277.1	0.0
	m_3	530.4	527.7	530.3	498.7	373.2	334.5	223.7	0.0
	m_4	588.7	599.1	585.3	607.0	334.9	306.9	214.8	0.0
	m_5	218.9	200.0	158.2	135.0	237.1	223.2	108.0	0.0
	m_6	250.6	257.3	185.1	175.6	273.1	284.5	119.6	0.0
	m_7	612.9	539.2	470.0	187.8	607.0	258.6	628.6	0.0
	m_8	2.2	2.4	0.00	3.7	21.9	8.5	9.3	40.1

Table 3: Matrix displaying the LME given a data generating model (rows) and a model used for inversion (columns). Except for model m_3, the data generating model had always the highest LME with $\Delta LME > 3$ compared to the second-best model. Data generated by model m_3 did not result in a distinctly superior model.

Modelling saccadic adaptation – empirical data

To test the effect of the experimental design we first compared all models within each of the four conditions independently. The results for each condition and the sum across factors are displayed in Fig. 3, i.e. a fixed effect analysis. In the *delay*, *step90* condition, m_8 was favored over all other models, with model m_5 ($LME_{m8} - LME_{m5} = 3.14$) having the second highest log marginal likelihood. In the *no delay*, *step90* condition m_5 and m_6 did not differ strongly in their model evidence ($LME_{m5} -$

$LME_{m6} = 2.90$), but model m_5 was strongly favoured with respect to all other models, e.g. ($LME_{m5} - LME_{m2} = 3.44$, $LME_{m5} - LME_{m4} = 3.24$). Importantly, all adaptation models were strongly favoured compared to the no adaption model m_8. In the two *step10* conditions, there was strong evidence for model m_8 as compared to all other models, irrespective of the delay.

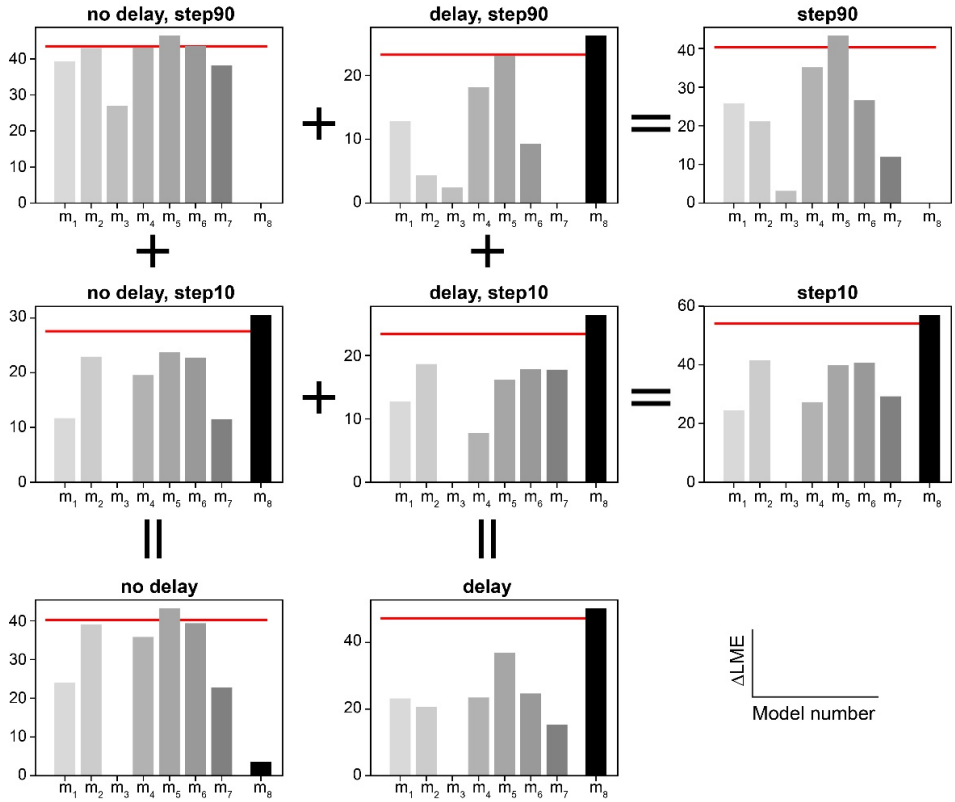


Figure 3: Log model evidence across the four conditions and pooled over conditions (right and bottom). Bars depict summed model evidence. Grey bars are different versions of models with adaptation (m_1 to m_7), with decreasing number of degrees of freedom from left to right. The black bar shows the LME for the no adaptation model m_8. LME values are shown relative to the model with the lowest *LME* in each of the panels. Red horizontal lines show a difference of 3 log units from the model with the highest *LME*.

We then considered the question whether subjects adapted during the *delay* and *no delay* conditions. In order to test this hypothesis, we compared the *LME* of each model by pooling over the *step10* and *step90* conditions. The results are shown in Fig. 3 bottom. Model m_5, according to which backward adaptation aimed at the base line, was better than any other model in the *no delay* condition. Moreover, we found strong

evidence (> 30 log units) in favour of m_5 with respect to the no adaptation model m_8. In the *delay* condition, there was strong evidence in favour of m_8 compared to any other model: Its evidence was over 10 log units higher compared to model m_5. Finally, when considering only the effect of step (Fig. 3 right), model m_5 obtained the highest evidence when the target was stepped 90% of times. In the *step10* condition, model m_8 obtained the largest evidence. In brief, this means that the adaptation model m_5 was superior when prediction errors frequently occurred; by contrast, the simpler model m_8 (that did not account for adaptation) was sufficient when prediction errors were rare.

We also used the model to explore individual differences in saccadic adaptation and compare them to brain activity. For this, we first computed for every subject and every condition the log Bayes factor ($=\Delta LME$) between models m_5 and m_8. This difference provides a measure of the strength of adaptation by quantifying how likely it is that the data was generated by the adapting model m_5 compared to the non-adapting model m_8. In general, we found that the spread of LME differences was larger in the two *step90* conditions. This reflects the fact that these conditions contained much more *step* trials that could inform the model. In contrast, in the *step10* condition, it was not possible to distinguish between models on a single subject basis, most likely because there were too few adaptation trials to inform the inference. Note that in the *no delay, step90* condition, there were several subjects (n=8) for whom model comparison clearly indicates adaptation happening. Some participants (n=3) also showed adaptation in the *delay, step90* condition. These findings are remarkable given the relatively small number of trials in a single block. See below (Fig. 7) for an illustration of these model comparisons.

fMRI results

Unless otherwise stated, all results are reported with familywise error, whole brain correction at the cluster level $p_{FWE} < \alpha = 0.05$ at a cluster defining threshold of $p < 0.001$. Reported coordinates correspond to peak activation within clusters.

We first investigated the effect of stepping the target vs. it reappearing at the pre-saccadic location. Under the assumption that stepping generated an error signal, this contrast reveals regions whose activation correlates with prediction errors. See Table 4 for a list of activations. The

comparison of *step90* condition with the *step10* condition mainly increased BOLD signals in bilateral FEF (bilateral precentral gyrus; PreCG) as well as bilateral parietal areas (cf. Fig. 4A). In addition, there were activations in left inferior frontal gyrus (extending into anterior insula), right inferior frontal gyrus (Area 44), bilateral medial occipital gyrus, and the right supramarginal gyrus/postcentral gyrus. The peak activation in the right FEF was the only activation that also survived a correction for multiple comparisons at the voxel level ($t_{92} = 5.39$, $p_{FWE} = 0.030$).

Main effect of stepping probability						
Region	p_{FWE}	c.s.	t_{92}	x	y	z
Left SPL (Area 7A) and Precuneus	<0.001	910	4.65	-23	-61	57
Right PreCG (FEF), Right SFG and MFG	<0.001	712	5.39	22	-1	55
Left PreCG (FEF), Right SFG and MFG	<0.001	393	4.66	-23	-4	57
Right SPL (Area 7A and 7P) and Precuneus	0.002	352	4.65	20	-69	54
Right MOG and SOG	0.003	316	4.59	28	-72	33
Left IPL	0.005	295	4.61	-31	-46	51
Left IFG (anterior insula)	0.012	248	4.55	-24	21	-2
Right SMG and PoCG	0.016	232	4.37	34	-37	43
Right IFG (Area 44)	0.030	202	4.19	46	7	13
Left MOG	0.043	185	4.30	-25	-72	24

Table 4: Significant clusters for main effect of step. Reported is the whole brain corrected family wise error cluster-level probability (p_{FWE}), cluster size (c.s.), *t* value and coordinates of the cluster peak in the MNI space. SPL: superior parietal lobule, PreCG: precentral gyrus, SFG: superior frontal gyrus, MFG: middle frontal gyrus, MOG: middle occipital gyrus, SOG: superior occipital gyrus, IPL: inferior parietal lobule, SMG: supramarginal gyrus, PoCG: postcentral gyrus, IFG: inferior frontal gyrus. All p values larger than 0.001 are rounded to the next higher value with 3 decimals after the comma (hence actual p-values are slightly lower). Note that the activation in the left IFG (anterior insula) extended through the white matter to the left putamen.

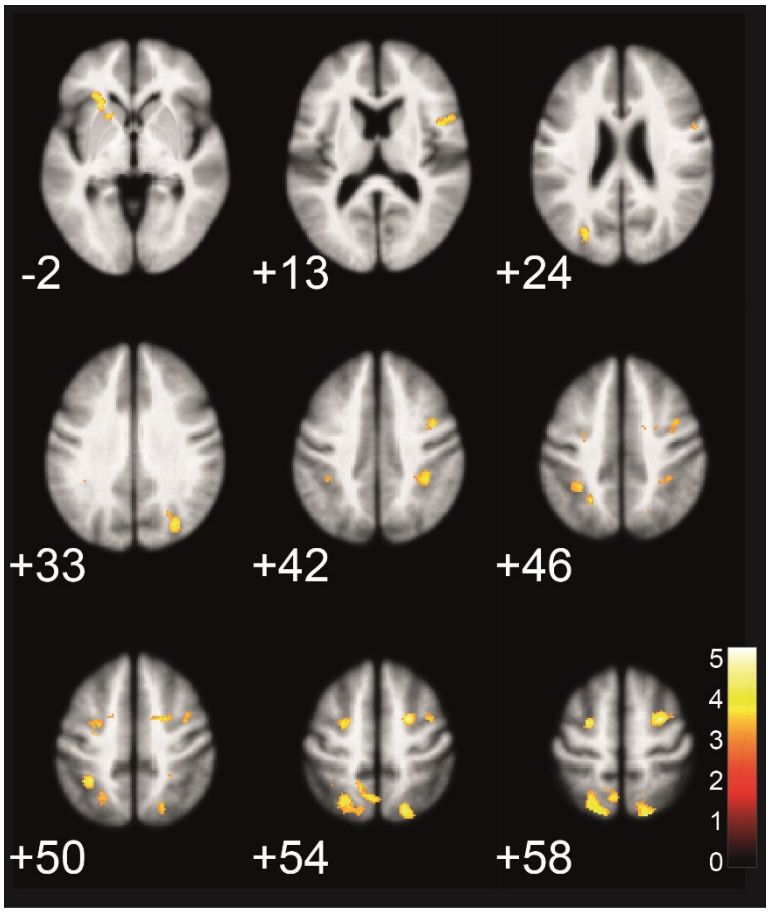


Figure 4: Main effect of stepping probability. Illustration of significant activations for *step90* condition compared to *step10* condition (cf. Table 4). Activation was thresholded at $p<0.001$ uncorrected with a cluster size $k>170$, showing significant clusters only. The colorbar indicates t-values.

Next, we tested whether two other key regions for CD, the SC and the thalamus, (Sommer and Wurtz, 2008) were activated by stepping the target during the saccade. We employed small volume correction for the SC (Limbrick-Oldfield et al., 2012) and, separately, for the right and left prefrontal thalamus (Behrens et al., 2003) using anatomically defined masks. Within each of these regions we assessed significance at the voxel level using SPM's small volume correction for multiple comparison and taking into account multiple comparison (3 ROIs) by setting the significance level to $0.0166=0.05/3$. In all three ROIs there was a significant activation below this threshold. In the SC, the peak was at coordinate [9 -30 -3] with $p_{FWE} = 0.011$ ($t_{92} = 3.65$). Within the right and left prefrontal thalamus the strongest activations were found at [6 -

22 0] ($p_{FWE} = 0.010$, $t_{92} = 4.20$) and [-9 -7 0] ($p_{FWE} < 0.001$, $t_{92} = 5.21$), respectively. Figure 5 illustrates the location of these activations within the SC and the prefrontal thalamus.

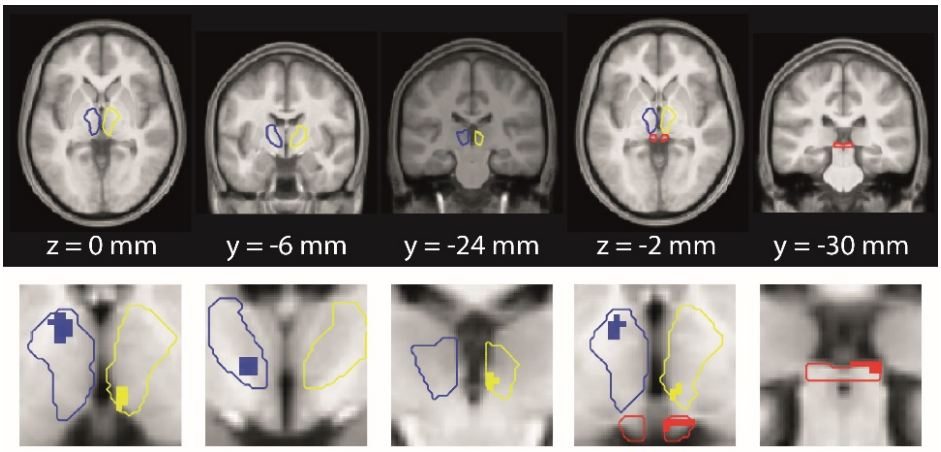


Figure 5: Main effect of stepping probability within thalamus and SC. Top: Illustration of left (blue) and right (yellow) thalamic ROIs (Behrens et al., 2003) and of the SC mask (red; Limbrick-Oldfield et al., 2012). **Bottom:** Activations within the three masks shown in colours corresponding to the top row. Please note that activations are plotted at a FWE peak level corrected threshold of $p<0.05$ (small volume corrected for each mask individually), for illustration. Peak activation in all three regions survived the Bonferroni corrected alpha level of $0.0166=0.05/3$.

The second analysis investigated the main effect of the *delay* factor. Delayed presentation of the post-saccadic target resulted in increased activity in three clusters (cf. Table 5 and Figure 6). In addition to these clusters, there was a strong activation at the anterior tip of the right insula which survived FWE correction at the voxel level ($p_{FWE} < 0.05$, peak at [29, 24, -8] with $p_{FWE}=0.002$). There was no significant voxel for this contrast in any of the three ROI. Finally, we did not find any region that showed an interaction effect between stepping probability and delay.

Positive effect of delay						
Region	p_{FWE}	c.s.	t_{92}	x	y	z
Right PreCG and MFG	<0.001	924	5.21	47	13	40
Right STG and MTG	<0.001	763	4.67	53	-45	15
Right Precuneus	0.002	354	4.63	8	-64	43

Table 5: Summary of the main effect of delay. We report the whole brain corrected family wise error cluster-level probability (p_{FWE}) at a cluster defining threshold of $p<0.001$, cluster size (c.s.), t value and coordinates in the MNI space of the cluster peak. PreCG: precentral gyrus, MFG: middle frontal gyrus, MTG: middle temporal gyrus, STG: superior temporal gyrus.

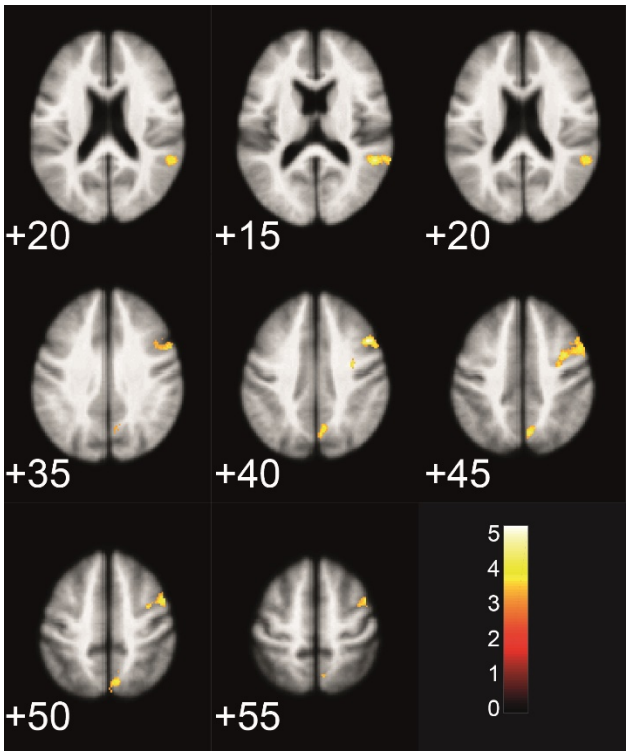


Figure 6: Main effect of delay. Regions that are more activated during the delay condition compared to the no-delay condition are displayed. For a detailed list of regions see Table 5. Note that the activation in the anterior insula is not shown, because it was not significant at the cluster level. The colour bar indicates t-values.

Effect of delay on adaptation and BOLD signal during step trials

The difference between the *delay* and *no-delay* conditions is of a particular interest when the target was stepped, it has been shown that the delay reduces adaptation (Bahcall and Kowler, 2000; Fujita et al., 2002). To better understand the effect of delay on SA and visual PE, we computed a first level contrast between the *no-delay, step90* and *delay, step90* conditions and assessed group effects at the second level with a t-test. Since both conditions included the same number of trials, any difference would point to a different processing of prediction errors in these two conditions. There was no significant cluster showing higher activity in the *no-delay* condition compared to the *delay* condition. For the contrast between *delay – no-delay* conditions, two regions survived the family wise error correction at the threshold level: One cluster in the precentral gyrus and middle frontal gyrus ($p_{cluster} = 0.002$, maximum at MNI: [48, 1, 48]) and one cluster in the anterior insula ($p_{cluster} = 0.02$,

maximum at MNI: [27, 24, -7]). While this analysis is not orthogonal and, thus, not completely independent from the main effect of delay shown above, it provides additional information for the specific effect of a delay during adaptation trials.

Relation of cortical activations to saccadic adaptation

In this analysis, we investigated whether there was any brain signal related to how much participants adapted the length of their saccades. The difference in log model evidence or log Bayes factor between the best adaptation (m_5) and the non-adapting model (m_8) provides a measure of the strength of SA. See Figure 7A for an illustration of subject specific *ΔLME* in all 4 conditions. We correlated the difference in log Bayes factors (m_5 vs. m_8) between the two conditions (*no-delay¸step90* vs. *delay¸step90*) with the fMRI contrast that models the difference between the same two conditions. This resulted in a cluster in the intraparietal sulcus and superior parietal lobule (Fig. 7) in which activation difference between *no-delay* and *delay* was correlated with the difference in log model evidence between the two conditions ($p_{cluster}<0.001$, Cluster size: 470, Peak at MNI: [30, -54 36]). Hence, activation within this region was related to how much a participant adapted.

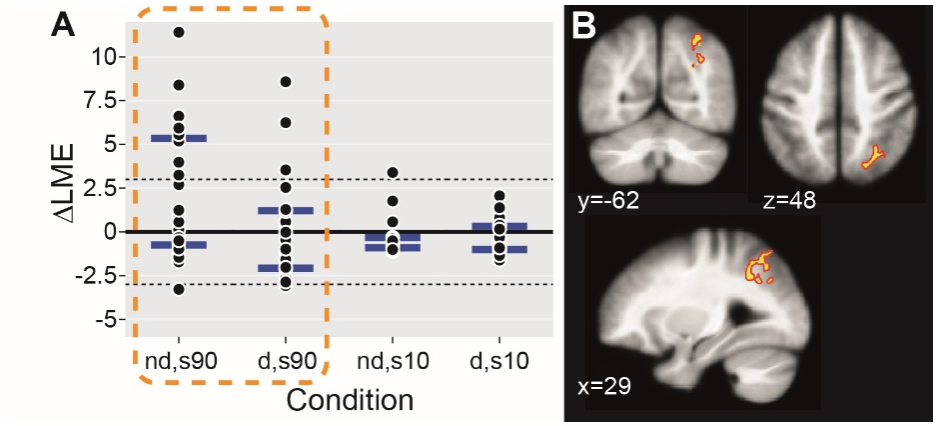


Figure 7: Correlation of adaptation with brain activation. A. Difference in LME between the best adaptation model (m_5) and the no adaptation model (m_8). Horizontal black dashed lines indicate the threshold of ΔLME=3. The orange dashed box highlights the data that went into the second level analysis. **B.** Illustration of the parietal cluster in which the difference in fMRI activation between *delay* and *no-delay* (*step90* condition) correlated with the difference in LME between the same two conditions.

Neural correlates of the internal motor command

Finally, we investigated whether there was any cortical or subcortical activation correlated with the internal motor command. For this, we used the estimated saccadic length $\hat{s}_t$ from the best model (m_5; see Eq. 1 and Fig 8A for example traces) and correlated it with the brain activity during the two *step90* conditions. Within a mask of all regions with significant activation for step vs. no-step (Fig. 4), we found a significant cluster ($p_{cluster}<0.01$ at a cluster defining threshold of $p<0.001$, cluster size: 100, peak at MNI: [-14 -66 58], FWE-corrected for the mask volume) in the left superior parietal lobule (area 7A). Hence, within this region, the activation during stepping trials was modulated by the internal motor command for saccadic length. There was no significant modulation in any of the subcortical ROIs.

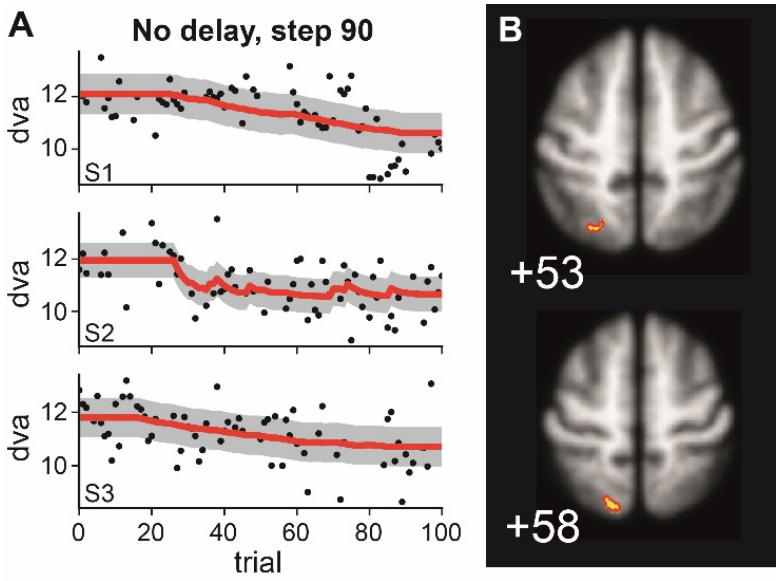


Figure 8: Modulation of brain activation by saccadic length of internal motor command. A. Illustration of internal saccadic length $\hat{s}$ for three sample subjects (S1-S3) in the no delay, step 90 condition. **B.** Brain activity correlated with this internal saccadic length estimate in a parietal cluster ($p_{cluster}<0.01$, Cluster size: 100, Peak at MNI: [-14 -66 58], statistic S.V.C. for the mask).

Discussion

In this paper, we show that cortical and subcortical regions are activated by PE during SA. We used a factorial design to disentangle the effects of the delay of the post-saccadic target and stepping probability, both of which showed an impact on learning. Bayesian model comparison of adapting vs. non-adapting models demonstrated that our design had the expected behavioural effect. Focusing on blocks with a high probability of stepping, we observed clusters in parietal regions whose activation was significantly correlated with how much participants adapted according to log Bayes factors and with an internal estimate of saccadic length, respectively.

Behaviour

We hypothesized that SA in the double step task (McLaughlin, 1967) is driven by a prediction error signal, computed from sensory feedback and a prediction of the retinal post-saccadic location of the target. The rate of adaptation was modelled with the parameters α_a and α_d, and the total adaptation gain with the bias terms γ_a and γ_d. In previous models, SA

was described by a similar form of learning (Kording et al., 2007; Srimal et al., 2008). Here, we have extended these earlier accounts by including motor biases and differences in learning rate depending on the direction of adaptation.

Bayesian model comparison was used to test whether the experimental conditions had an effect on the participants' behaviour. This approach selects within a series of nested models the most parsimonious one, i.e. the least complex model which adequately explains the experimental observations. In the *no-delay* blocks, model comparison favoured a model that included forward and backward adaptation, whereas in sessions with a delay, a simpler model without adaptation was favoured. In general, the model without adaptation was favoured in sessions with only 10% step trials, suggesting that there was not enough evidence for adaption because of the low number of step trials.

Comparison to previous imaging studies in humans

Our experimental paradigm was similar to the one proposed by Gerardin et al. (2012), but included *step* and *no step* trials in all conditions. As in the above study, delayed presentation of the post-saccadic target reduced or even abolished SA as demonstrated by the model comparison results (Fig. 3). However, irrespective of the delay, the shift of the saccadic target should trigger a visual prediction error. Therefore, the difference between the *step* and *no step* conditions should reveal areas involved in post-saccadic visual prediction error signalling. The delay should reduce SA, but not visual prediction errors. Hence, regions with activation differences between no-delay and delay condition might also influences adaptation.

There is little doubt about the pivotal role of the cerebellum during SA (for reviews see Hopp and Fuchs, 2004; Pelisson et al., 2010; Herman et al., 2013). Because we did not measure the entire cerebellum, we cannot directly compare our results to previous fMRI (van Broekhoven et al., 2009; Liem et al., 2013) and positron emission tomography (PET) (Desmurget et al., 1998; Desmurget et al., 2000) studies of SA focused on the cerebellum. More recently two studies have investigated cortical activation during SA. For reactive saccades, as the ones investigated in our experiment, Gerardin et al. (2012) reported a difference between *delay* and *no-delay* in the right TPJ and hMT+, as well as bilaterally in the inferior precentral sulcus (iPrCS). These results overlap with our

finding that the contrast *delay* vs. *no delay* showed increase activation of the right STG, MTG and angular gyrus as well as the iPrCS. However, we did not find an activation in hMT+. Blurton et al. (2012) also relied on a similar paradigm, but the saccadic target was not removed during the delay period. It is therefore difficult to directly compare our results to their findings.

A novel finding by our study is the cortical activation by visual prediction errors across saccades (*step90* vs. *step10*) in the bilateral FEF as well as prefrontal thalamus and the right SC. Parietal areas were activated in both main contrasts: *step90* vs. *step10* as well as *no-delay* vs. *delay*.

There are at least two potential reasons for the discrepancies with previous studies. First, we randomly interleaved *step* and *no step* trials, as opposed to having a block design. Second, our study included more subjects (n=24) compared to n=12 (Blurton et al., 2012) and n=6x2 (Gerardin et al., 2012).

Potential involvement of parietal cortex in saccadic adaptation

In line with recent findings, our study provides evidence for activation of parietal regions during SA. For example, a TMS study (Panouilleres et al., 2014) showed that stimulation of the right posterior intraparietal sulcus (iPS) increased the backward adaptation gain of reactive saccades to the right. Interestingly, the parietal activation of effect of step in our study overlapped with the stimulation site (Tal: x=13±5, y=-63±6, y=52±5; MNI: x=13±5, y=-68±6, y=59±5) of Panouilleres et al. (2014). The parietal activation for the effect of delay was more medial, mainly located in the precuneus. A third parietal region was the only cluster which exhibited a significant correlation between brain activity and the amount of adaptation as indexed by Bayesian model comparison. This region did not overlap with the TMS stimulation site of Panouilleres et al. (2014). Neurons in the lateral intraparietal area have recently been implicated with the monitoring of saccadic error in the macaque (Zhou et al., 2016). Thus, one could speculate that SA might be related to the amount of saccadic error signalling in parietal cortex. Future studies will have to test how these regions interact with the cerebellar circuits known to be crucial for SA (Hopp and Fuchs, 2004), and whether they might indeed play a causal role in SA as indicated by TMS (Panouilleres et al., 2014). Finally, a left parietal region exhibited activation correlated to the saccadic length of the

internal motor command derived from the model. The left hemispheric location is in line with saccades made towards the right. Hence, this activation is more likely to represent an adapted saccadic signal, instead of an error signal.

Involvement of the saccadic corollary discharge pathway in saccadic adaptation

We found increased activation of the FEF, SC and thalamus when the saccadic target was shifted from its original location. These regions form the core circuit of internal monitoring of saccadic eye movements (Sommer and Wurtz, 2002, 2004b, a). Consistent with our finding, neurons in the FEF monitor intra-saccadic changes of visual input (Crapse and Sommer, 2012). The activation in the SC is of particular interest, because electrical stimulation in the SC induces SA (Kaku et al., 2009). In accordance with these experiments, the right SC (contralateral to the direction of adaptation) was activated for backward adaptation of rightward saccades (Kaku et al., 2009). Finally, thalamic lesions in humans result in asymmetric SA with the same lateralization (Gaymard et al., 2001; Zimmermann et al., 2015) and cause deficits in the perception of perisaccadic shifts (Ostendorf et al., 2013). Interestingly, the lesioned regions reported by this group partially overlap with our activation findings (see their Figure 6). In summary, the agreement of our results with the known literature supports the notion that the visual PE based on CD as incorporated in our model and signalled in the SC-Thalamus-FEF pathway could be involved in SA.

While the pattern of fMRI activation together with previous evidence points to the involvement of the SC-Thalamus-FEF pathway in generating error signals, our experiment cannot prove a causal relationship. However, analysis of identified FEF cells projecting to the SC suggests that complete remapping of visual input happens at the level of output cells in layer 5 (Sommer and Wurtz, 2006; Shin and Sommer, 2012) and is sent to the SC. Hence, the error signal in the SC at least partly depends on cortical input. Accordingly, subthreshold electrical stimulation of the SC probably elicited SA by triggering a signal similar to PE from the FEF (Kaku et al., 2009). To our knowledge the FEF-SC-Thalamus pathway has not been studied in the context of SA.

Summary

In this study, we have shown that SA and, in particular the hypothesized trans-saccadic visual PE, elicit a wide range of cortical activations. Parietal areas and the FEF were sensitive to intrasaccadic steps of the target. In addition, the SC as well as the thalamus were activated in this condition pointing towards a potential involvement of the SC-Thalamus-FEF pathway. In parietal regions, both the delay as well as the intra-saccadic step changed cortical activation. Finally, two model-based findings hint at the importance of parietal circuits in SA in humans: A correlation between brain activity and log Bayes' factors (encoding the evidence for adaptation processes) in the right parietal cortex, and a correlation of BOLD activity with an estimate of saccadic length of the internal motor command in the left hemisphere. In summary, we have shown that SA is associated with activation in a set of cortical and subcortical areas, which might act in coordination with the well-established role of the cerebellum.

References

Aponte EA, Raman S, Sengupta B, Penny WD, Stephan KE, Heinzle J (2016) mpdcm: A toolbox for massively parallel dynamic causal modeling. Journal of neuroscience methods 257:7-16.

Bahcall DO, Kowler E (2000) The control of saccadic adaptation: implications for the scanning of natural visual scenes. Vision research 40:2779-2796.

Behrens TE, Johansen-Berg H, Woolrich MW, Smith SM, Wheeler-Kingshott CA, Boulby PA, Barker GJ, Sillery EL, Sheehan K, Ciccarelli O, Thompson AJ, Brady JM, Matthews PM (2003) Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. Nature neuroscience 6:750-757.

Blurton SP, Raabe M, Greenlee MW (2012) Differential cortical activation during saccadic adaptation. Journal of neurophysiology 107:1738-1747.

Collins T, Wallman J (2012) The relative importance of retinal error and prediction in saccadic adaptation. Journal of neurophysiology 107:3342-3348.

Collins T, Rolfs M, Deubel H, Cavanagh P (2009) Post-saccadic location judgments reveal remapping of saccade targets to non-foveal locations. Journal of vision 9:29 21-29.

Crapse TB, Sommer MA (2012) Frontal eye field neurons assess visual stability across saccades. The Journal of neuroscience : the official journal of the Society for Neuroscience 32:2835-2845.

Desmurget M, Pelisson D, Urquizar C, Prablanc C, Alexander GE, Grafton ST (1998) Functional anatomy of saccadic adaptation in humans. Nature neuroscience 1:524-528.

Desmurget M, Pelisson D, Grethe JS, Alexander GE, Urquizar C, Prablanc C, Grafton ST (2000) Functional adaptation of reactive saccades in humans: a PET study. Experimental brain research Experimentelle Hirnforschung Experimentation cerebrale 132:243-259.

Eickhoff SB, Paus T, Caspers S, Grosbras MH, Evans AC, Zilles K, Amunts K (2007) Assignment of functional activations to probabilistic cytoarchitectonic areas revisited. NeuroImage 36:511-521.

Flandin G, Friston KJ (2016) Analysis of family-wise error rates in statistical parametric mapping using random field theory. arXiv preprint arXiv:160608199.

Fujita M, Amagai A, Minakawa F, Aoki M (2002) Selective and delay adaptation of human saccades. Cognitive Brain Research 13:41-52.

Gaymard B, Rivaud S, Pierrot-Deseilligny C (1994) Impairment of extraretinal eye position signals after central thalamic lesions in humans. Experimental Brain Research 102:1-9.

Gaymard B, Rivaud-Pechoux S, Yelnik J, Pidoux B, Ploner CJ (2001) Involvement of the cerebellar thalamus in human saccade adaptation. The European journal of neuroscience 14:554-560.

Gelman A, Meng X-L (1998) Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling. Statistical Science 13:163-185.

Gelman A, Carlin JD, Stern HS, Dunson DB, Vehtari A, B. RD (2013) Bayesian Data Analysis - Third Edition. Boca Raton, FL: CRC Press, Taylor & Francis Group.

Gerardin P, Miquee A, Urquizar C, Pelisson D (2012) Functional activation of the cerebral cortex related to sensorimotor adaptation of reactive and voluntary saccades. NeuroImage 61:1100-1112.

Glover GH, Li TQ, Ress D (2000) Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine 44:162-167.

Herman JP, Blangero A, Madelain L, Khan A, Harwood MR (2013) Saccade adaptation as a model of flexible and general motor learning. Exp Eye Res 114:6-15.

Hopp JJ, Fuchs AF (2004) The characteristics and neuronal substrate of saccadic eye movement plasticity. Progress in neurobiology 72:27-53.

Kaku Y, Yoshida K, Iwamoto Y (2009) Learning signals from the superior colliculus for adaptation of saccadic eye movements in the monkey. The Journal of neuroscience : the official journal of the Society for Neuroscience 29:5266-5275.

Kasper L, Bollmann S, Diaconescu AO, Hutton C, Heinzle J, Iglesias S, Hauser TU, Sebold M, Manjaly ZM, Pruessmann KP, Stephan KE (2017) The PhysIO Toolbox for Modeling Physiological Noise in fMRI Data. Journal of neuroscience methods 276:56-72.

Kass RE, Raftery AE (1995) Bayes Factors. J Am Stat Assoc 90:773-795.

Kording KP, Tenenbaum JB, Shadmehr R (2007) The dynamics of memory as a consequence of optimal adaptation to a changing body. Nature neuroscience 10:779-786.

Lewis R, Zee D, Hayman M, Tamargo R (2001) Oculomotor function in the rhesus monkey after deafferentation of the extraocular muscles. Experimental Brain Research 141:349-358.

Liem EI, Frens MA, Smits M, van der Geest JN (2013) Cerebellar activation related to saccadic inaccuracies. Cerebellum 12:224-235.

Limbrick-Oldfield EH, Brooks JCW, Wise RJS, Padormo F, Hajnal JV, Beckmann CF, Ungless MA (2012) Identification and characterisation of midbrain nuclei using optimised functional magnetic resonance imaging. NeuroImage 59:1230-1238.

MacKay DJC (2003) Information Theory, Inference, and Learning Algorithms: Cambridge University Press.

McLaughlin SC (1967) Parametric adjustment in saccadic eye movements. Perception & Psychophysics 2:359-362.

Ostendorf F, Liebermann D, Ploner CJ (2013) A role of the human thalamus in predicting the perceptual consequences of eye movements. Frontiers in systems neuroscience 7:10.

Panouilleres M, Habchi O, Gerardin P, Salemme R, Urquizar C, Farne A, Pelisson D (2014) A role for the parietal cortex in sensorimotor adaptation of saccades. Cereb Cortex 24:304-314.

Pelisson D, Alahyane N, Panouilleres M, Tilikete C (2010) Sensorimotor adaptation of saccadic eye movements. Neuroscience and biobehavioral reviews 34:1103-1120.

Picard H, Le Seac'h A, Amado I, Gaillard R, Krebs MO, Beauvillain C (2012) Impaired saccadic adaptation in schizophrenic patients with high neurological soft sign scores. Psychiatry research 199:12-18.

Rolfs M, Knapen T, Cavanagh P (2010) Global saccadic adaptation. Vision research 50:1882-1890.

Shaby B, Wells MT (2010) Exploring an adaptive Metropolis algorithm. Technical Report.

Shin S, Sommer MA (2012) Division of labor in frontal eye field neurons during presaccadic remapping of visual receptive fields. Journal of neurophysiology 108:2144-2159.

Sommer MA, Wurtz RH (2002) A pathway in primate brain for internal monitoring of movements. Science 296:1480-1482.

Sommer MA, Wurtz RH (2004a) What the brain stem tells the frontal cortex. II. Role of the SC-MD-FEF pathway in corollary discharge. Journal of neurophysiology 91:1403-1423.

Sommer MA, Wurtz RH (2004b) 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 91:1381-1402.

Sommer MA, Wurtz RH (2006) Influence of the thalamus on spatial visual processing in frontal cortex. Nature 444:374-377.

Sommer MA, Wurtz RH (2008) Brain circuits for the internal monitoring of movements. Annual review of neuroscience 31:317-338.

Sperry RW (1950) Neural basis of the spontaneous optokinetic response produced by visual inversion. Journal of comparative and physiological psychology 43:482-489.

Srimal R, Diedrichsen J, Ryklin EB, Curtis CE (2008) Obligatory adaptation of saccade gains. Journal of neurophysiology 99:1554-1558.

Stampe DM (1993) Heuristic Filtering and Reliable Calibration Methods for Video-Based Pupil-Tracking Systems. Behav Res Meth Instr 25:137-142.

Swendsen RH, Wang J-S (1986) Replica Monte Carlo Simulation of Spin-Glasses. Physical review letters 57:2607-2609.

van Broekhoven PC, Schraa-Tam CK, van der Lugt A, Smits M, Frens MA, van der Geest JN (2009) Cerebellar contributions to the processing of saccadic errors. Cerebellum 8:403-415.

von Holst E, Mittelstaedt H (1950) Das Reafferenzprinzip. (Wechselwirkungen zwischen Zentralnervensystem und Peripherie.). Naturwissenschaften 37:464-376.

Wong AL, Shelhamer M (2011) Sensorimotor adaptation error signals are derived from realistic predictions of movement outcomes. Journal of neurophysiology 105:1130-1140.

Zhou Y, Liu Y, Lu H, Wu S, Zhang M (2016) Neuronal representation of saccadic error in macaque posterior parietal cortex (PPC). Elife 5.

Zimmermann E, Ostendorf F, Ploner CJ, Lappe M (2015) Impairment of saccade adaptation in a patient with a focal thalamic lesion. Journal of neurophysiology 113:2351-2359.

Chapter 6

In Chapter 4, we argued that the antisaccade task is an interesting research topic for computational psychiatry. There are at least two reasons for this: First, recent studies have shown that deficits in this task are likely endophenotypes of schizophrenia (e.g. Radant et al., 2010; Reilly et al., 2014). Second, the biological underpinnings of the antisaccade task have been widely investigated in human and non-human primates (Munoz and Everling, 2004), making it possible, at least in principle, to relate behavioral findings to physiological processes. Unfortunately, only few computational models of this task exist (reviewed in Cutsuridis, 2017), and none of them has offered a formal probabilistic approach to the two main variables of interest measured in this paradigm: reaction times and error rates. This precludes the comparison of these models, and a truly quantitative analysis of experimental data.

This chapter presents a novel modeling approach to the antisaccade task, which we refer to as the *Stochastic Early Reaction, Inhibition, and late Action* (SERIA) model. SERIA is formal in the sense that it provides a well-defined likelihood function of reaction times and actions, in a trial-by-trial fashion.

Our results reveal the complexity of the decision processes that occur in the tenths of a second that precede a saccade. Furthermore, the development of a formal model such as SERIA opens the door to quantitative analysis of experimental data. This idea is followed in the rest of this dissertation.

This chapter was published as *Aponte, E. A., Schobi, D., Stephan, K. E., & Heinzle, J. (2017). The Stochastic Early Reaction, Inhibition, and late Action (SERIA) model for antisaccades. PLoS Computational Biology, 13(8), e1005692*. It is verbatim copy of the document https://doi.org/10.1371/journal.pcbi.1005692.

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

The Stochastic Early Reaction, Inhibition, and late Action (SERIA) model for antisaccades

Eduardo A. Aponte[1]*, Dario Schöbi[1], Klaas E. Stephan[1,2], Jakob Heinzle[1]*

1 Translational Neuromodeling Unit, Institute for Biomedical Engineering, University of Zurich & Swiss Institute of Technology Zurich, Zurich, Switzerland, **2** Wellcome Trust Centre for Neuroimaging, University College London, London, United Kingdom

* aponte@biomed.ee.ethz.ch (EAA); heinzle@biomed.ee.ethz.ch (JH)

OPEN ACCESS

Citation: Aponte EA, Schöbi D, Stephan KE, Heinzle J (2017) The Stochastic Early Reaction, Inhibition, and late Action (SERIA) model for antisaccades. PLoS Comput Biol 13(8): e1005692. https://doi.org/10.1371/journal.pcbi.1005692

Editor: Adrian M. Haith, Johns Hopkins University, UNITED STATES

Received: February 5, 2017

Accepted: July 20, 2017

Published: August 2, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the René and Susanne Braginsky Foundation (KES) and the University of Zurich. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

The antisaccade task is a classic paradigm used to study the voluntary control of eye movements. It requires participants to suppress a reactive eye movement to a visual target and to concurrently initiate a saccade in the opposite direction. Although several models have been proposed to explain error rates and reaction times in this task, no formal model comparison has yet been performed. Here, we describe a Bayesian modeling approach to the antisaccade task that allows us to formally compare different models on the basis of their evidence. First, we provide a formal likelihood function of actions (pro- and antisaccades) and reaction times based on previously published models. Second, we introduce the *Stochastic Early Reaction*, *Inhibition*, *and late Action model* (SERIA), a novel model postulating two different mechanisms that interact in the antisaccade task: an early GO/NO-GO race decision process and a late GO/GO decision process. Third, we apply these models to a data set from an experiment with three mixed blocks of pro- and antisaccade trials. Bayesian model comparison demonstrates that the SERIA model explains the data better than competing models that do not incorporate a late decision process. Moreover, we show that the early decision process postulated by the SERIA model is, to a large extent, insensitive to the cue presented in a single trial. Finally, we use parameter estimates to demonstrate that changes in reaction time and error rate due to the probability of a trial type (pro- or antisaccade) are best explained by faster or slower inhibition and the probability of generating late voluntary prosaccades.

Author summary

One widely replicated finding in schizophrenia research is that patients tend to make more errors than healthy controls in the antisaccade task, a psychometric paradigm in which participants are required to look in the opposite direction of a visual cue. This deficit has been suggested to be an endophenotype of schizophrenia, as first order relatives of patients tend to show similar but milder deficits. Currently, most models applied to experimental findings in this task are limited to fit average reaction times and error rates. Here, we propose a novel statistical model that fits experimental data from the antisaccade task,

beyond summary statistics. The model is inspired by the hypothesis that antisaccades are the result of several competing decision processes that interact nonlinearly with each other. In applying this model to a relatively large experimental data set, we show that mean reaction times and error rates do not fully reflect the complexity of the processes that are likely to underlie experimental findings. In the future, our model could help to understand the nature of the deficits observed in schizophrenia by providing a statistical tool to study their biological underpinnings.

Introduction

In the antisaccade task ([1]; for reviews, see [2,3]), participants are required to saccade in the contralateral direction of a visual cue. This behavior is thought to require both the inhibition of a reflexive saccadic response towards the cue and the initiation of a voluntary eye movement in the opposite direction. A failure to inhibit the reflexive response leads to an erroneous saccade towards the cue (i.e., a prosaccade), which is often followed by a corrective eye movement in the opposite direction (i.e., an antisaccade). As a probe of inhibitory capacity, the antisaccade task has been widely used to study psychiatric and neurological diseases [3]. Notably, since the initial report [4], studies have consistently found an increased number of errors in patients with schizophrenia when compared to healthy controls, independent of medication and clinical status [5–8]. Moreover, there is evidence that an increased error rate constitutes an endophenotype of schizophrenia, as antisaccade deficits are also present in non-affected, first-degree relatives of diagnosed individuals (for example [5,7]; but for negative findings see for example [9,10]).

Unfortunately, the exact nature of the antisaccade deficits and their biological origin in schizophrenia remain unclear. One path to improve our understanding of these experimental findings is to develop generative models of their putative computational and/or neurophysiological causes [11]. Generative models that capture the entire distribution of responses can reveal features of the data that are not apparent when only considering summary statistics such as mean error rate (ER) and reaction time (RT) [12–15]. Additionally, this type of model can potentially relate behavioral findings in humans to their biological substrate.

Here, we apply a generative modeling approach to the antisaccade task. First, we introduce a novel model of this paradigm based on previous proposals [16–20]. For this, we formalize the ideas introduced by Noorani and Carpenter [17] and extend them into what we refer to as the *Stochastic Early Reaction, Inhibition and late Action* (SERIA) model. Second, we apply both models to an experimental data set of three mixed blocks of pro- and antisaccades trials with different trial type probability. More specifically, we compare several models using Bayesian model comparison. Third, we use the parameter estimates from the best model to investigate the effects of our experimental manipulation. We found that there was positive evidence in favor of the SERIA model when compared to our formalization of the model proposed in [17]. Moreover, the parameters estimated through model inversion revealed the complexity of the decision processes underlying the antisaccade task that is not obvious from mean RT and ER.

This paper is organized as follows. First, we formalize the model developed in [17] and introduce the SERIA model. Second, we describe our experimental setup. Third, we present our behavioral findings in terms of summary statistics (mean RT and ER), the comparison between different models, and the parameter estimates. Finally, we review our findings, discuss other recent models, potential future developments, and translational applications.

Materials and methods

Ethics statement

All participants gave written informed consent before the study. All experimental procedures were approved by the local ethics board (Kantonale Ethikkomission Zürich, KEK-ZH-Nr.2014-0246).

Race models for antisaccades

In this section, we derive a formal description of the models evaluated in this paper. We start with a formalized version of the model proposed by Noorani and Carpenter in [17] and proceed to extend it. Their approach resembles the model originally proposed by Camalier and colleagues [21] to explain RT and ER in the double step and search step tasks, in which participants are either asked to saccade to successively presented targets or to saccade to a target after a distractor was shown. Common to all these tasks is that subjects are required to inhibit a prepotent reaction to an initial stimulus and then to generate an action towards a secondary goal. Briefly, Camalier and colleagues [21] extended the original 'horse-race' model [16] by including a secondary action in countermanding tasks. In [17], Noorani and Carpenter used a similar model in combination with the LATER model [22] in the context of the antisaccade task by postulating an endogenously generated inhibitory signal. Note that this model, or variants of it, have been used in several experimental paradigms (reviewed in [20]). Here, we limit our discussion to the antisaccade task.

The pro, stop, and antisaccade model (PROSA)

Following [17], we assume that the RT and the type of saccade generated in a given trial are caused by the interaction of three competing processes or units. The first unit u_p represents a command to perform a prosaccade, the second unit u_s represents an inhibitory command to stop a prosaccade, and the third unit u_a represents a command to perform an antisaccade. The time t required for unit u_i to arrive at threshold s_i is given by:

$$s_i = r_i t, \tag{1}$$

$$\frac{s_i}{r_i} = t, \tag{2}$$

where r_i represents the slope or increase rate of unit u_i, s_i represents the height of the threshold, and t represents time. We assume that, on each trial, the increase rates are stochastic and independent of each other.

The time and order in which the units reach their thresholds s_i determines the action and RT in a trial. If the prosaccade unit u_p reaches threshold before any other unit at time t, a prosaccade is elicited at t. If the antisaccade unit arrives first, an antisaccade is elicited at t. Finally, if the stop unit arrives before the prosaccade unit, an antisaccade is elicited at the time when the antisaccade unit reaches threshold. It is worth mentioning that, although this model is motivated as a race-to-threshold model, actions and RTs depend only on the arrival times of each of the units and ultimately no explicit model of increase rates or thresholds is required. Thus, for the sake of clarity, we refer to this approach as a 'race' model, in contrast to 'race-to-threshold' models that explicitly describe increase rates and thresholds.

Formally (but in a slight abuse of language), the two random variables of interest, the reaction time $T \in [0,\infty[$ and the type of action performed $A \in \{pro,anti\}$, depend only on three further random variables: the arrival times U_p, U_s, $U_a \in [0,\infty[$ of each of the units. The

probability of performing a prosaccade at time t is given by the probability of the prosaccade unit arriving at time t, and the stop and antisaccade unit arriving afterwards:

$$p(A = pro, T = t) = p(U_p = t)p(U_a > t)p(U_s > t). \quad (3)$$

The probability of performing an antisaccade at time t is given by

$$p(A = anti, T = t) = p(U_a = t)p(U_p > t)p(U_s > t) + p(U_a = t)\int_0^t p(U_s = \tau)p(U_p > \tau)d\tau. \quad (4)$$

The first term on the right side of Eq 4 corresponds to the unlikely case that the antisaccade unit arrives before the prosaccade and the stop units. The second term describes trials in which the stop unit arrives before the prosaccade unit. It can be decomposed into two terms:

$$p(U_a = t)\int_0^t p(U_s = \tau)p(U_p > \tau)d\tau = p(U_a = t)\left(p(U_s < t)p(U_p > t) + \int_0^t p(U_s = \tau)p(\tau < U_p < t)d\tau\right) \quad (5)$$

$$= p(U_a = t)\left(p(U_s < t)p(U_p > t) + \int_0^t p(U_s < \tau)p(U_p = \tau)d\tau\right) \quad (6)$$

The term $p(U_a = t)\int_0^t p(U_s < \tau)p(U_p = \tau)d\tau$ describes the condition in which the prosaccade unit is inhibited by the stop unit allowing for an antisaccade. Note that if the prosaccade unit arrives later than the antisaccade unit, the arrival time of the stop unit is irrelevant. That means that we can simplify Eq 4 to

$$p(A = anti, T = t) = p(U_a = t)\left(p(U_p > t) + \int_0^t p(U_s < \tau)p(U_p = \tau)d\tau\right). \quad (7)$$

Eqs 3 and 7 constitute the likelihood function of a single trial, and define the joint probability of an action and the corresponding RT. We refer to this likelihood function as the PRO-Stop-Antisaccade (PROSA) model. It shares the central assumptions of [17], namely: (i) the time to reach threshold of each of the units is assumed to depend linearly on the rate r, (ii) it includes a stop unit whose function is to inhibit prosaccades and (iii) there is no lateral inhibition between the different units. Finally, (iv) RTs are assumed to be equal to the arrive-at-threshold times. Note that the RT distributions are different from the arrival time distributions because of the interactions between the units described above. The main difference of this model compared to [17] is that we do not exclude *a priori* the possibility of the antisaccade unit arriving earlier than the other units. Aside from this, both models are conceptually equivalent.

The Stochastic Early Reaction, Inhibition, and Late Action Model (SERIA)

The PROSA model is characterized by a strict association between units and action types. In other words, the unit u_p leads unequivocally to a prosaccade, whereas the unit u_a always triggers an antisaccade. This implies that if the distribution of the arrival times of the units is unimodal and strictly positive, the PROSA model cannot predict voluntary slow prosaccades with a late peak, or in simple words, the PROSA model cannot account for slow, voluntary prosaccades that have been postulated in the antisaccade task [23]. Similarly, it has been argued that prosaccade RT can be described by the mixture of two distributions (for example [2,22]). To account for this, we introduce the Stochastic Early Reaction, Inhibition and Late Action (SERIA) model.

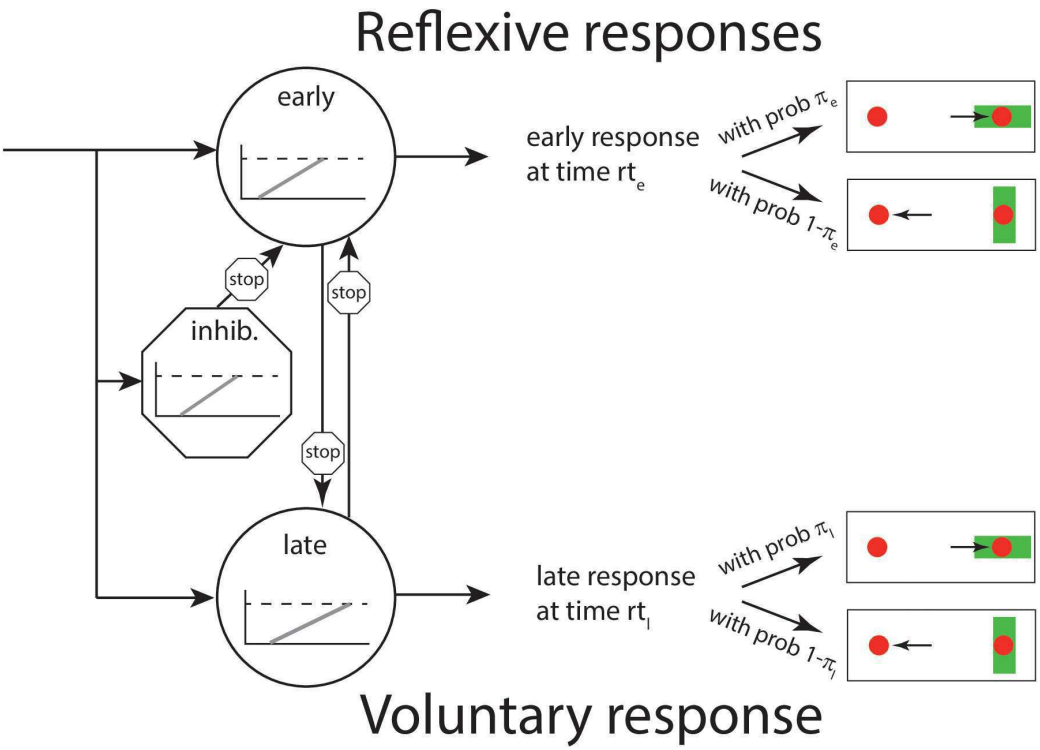


Fig 1. Layout of the SERIA model. The presentation of a visual cue (a green bar) triggers the race of three independent units. The inhibitory unit can stop an early response. Importantly, both early and late responses can trigger pro- and antisaccades. Note that the PROSA model is a special case of the SERIA model in which $\pi_e = 1$ and $\pi_l = 0$, i.e. all early responses are prosaccades, whereas all late responses are antisaccades.

https://doi.org/10.1371/journal.pcbi.1005692.g001

According to this model, and in analogy to the PROSA model, an early reaction takes place at time t if the early unit u_e arrives before the late and inhibitory units, u_l and u_i, respectively. If the inhibitory or late unit arrives before the early unit, a late response is triggered at the time the late unit reaches threshold. Crucially, both early and late responses can trigger pro- and antisaccades with a certain probability. Thus, in parallel to the race processes which determine RTs, an independent, secondary decision process is responsible for which reaction is generated. Fig 1 shows the structure of the SERIA model.

To formalize the concept of early and late responses, we introduce a new unobservable random variable that represents the type of response $R \in \{early, late\}$. The distribution of the RTs is analogous to the PROSA-model, such that, for instance, the probability of an early response at time t is given by

$$p(R = early, T = t) = p(U_e = t)p(U_i > t)p(U_l > t) \quad (8)$$

where U_e, U_i, and U_l represent the arrival times of the early, inhibitory, and late units, respectively. The fundamental assumption of the SERIA model is that a secondary decision process, beyond the race between early, inhibitory, and late units, decides the action generated in a single trial. An initial approach to model this secondary decision process is to assume that the action type (pro- or antisaccade) is conditionally independent of the RT given the response

type (early or late). Hence, the distribution of RTs is not *a priori* coupled to the saccade type anymore; RT distributions for both pro- and antisaccades could in principle be bimodal, consisting of both fast reactive and slow voluntary saccades.

Formally, the conditional independency assumption can be written down as

$$p(A, T|R) = p(A|R)p(T|R), \tag{9}$$

$$p(A, T|R)p(R) = p(A|R)p(T|R)p(R), \tag{10}$$

$$p(A, T, R) = p(A|R)p(T, R). \tag{11}$$

The term $p(A|R)$ is simply the probability of an action, given a response type. We denote it as

$$p(A = pro|R = early) = \pi_e \in [0, 1], \tag{12}$$

$$p(A = anti|R = early) = 1 - \pi_e, \tag{13}$$

$$p(A = pro|R = late) = \pi_l \in [0, 1], \tag{14}$$

$$p(A = anti|R = late) = 1 - \pi_l. \tag{15}$$

Since the type of response R is not observable, it is necessary to marginalize it out in Eq 11 to obtain the likelihood of the SERIA model:

$$p(A, T) = p(A, T, R = early) + p(A, T, R = late). \tag{16}$$

The complete likelihood of the model is given by substituting the terms in Eq 16

$$p(A = pro, T = t) = \pi_e p(U_e = t)p(U_i > t)p(U_l > t) +$$

$$\pi_l p(U_l = t)\left(p(U_e > t) + \int_0^t p(U_e = \tau)p(U_i < \tau)d\tau\right), \tag{17}$$

$$p(A = anti, T = t) = (1 - \pi_e)p(U_e = t)p(U_i > t)p(U_l > t) +$$

$$(1 - \pi_l)p(U_l = t)\left(p(U_e > t) + \int_0^t (U_e = \tau)p(U_i < \tau)d\tau\right). \tag{18}$$

It is worth noting here that the PROSA model is a special case of the SERIA model, namely, it corresponds to the assumption that $\pi_e = 1$ and $\pi_l = 0$. The SERIA model allows for bimodal distributions, as both early and late responses can be pro- and antisaccades. Importantly, one prediction of the model is that late prosaccades have the same distribution as late antisaccades.

A late race competition model for saccade type

Until now, we have assumed that the competition that leads to late pro- and antisaccades does not depend on time in the sense that late actions are conditionally independent of RT. This assumption can be weakened by postulating a secondary race between late responses; this leads us to a modified version of the SERIA model, that we refer to as the late race SERIA model ($SERIA_{lr}$). The derivation proceeds similarly to the SERIA model, except that we postulate a fourth unit that generates late prosaccades instead of assuming that the late decision process is time insensitive.

This version of the SERIA model includes an early unit u_e that, for simplicity, we assume produces only prosaccades, an inhibitory unit that stops early responses u_i, a late unit that triggers antisaccades u_a, and a further unit that triggers late prosaccades u_p. As before, if the early unit reaches threshold before any other unit, a prosaccade is generated with probability

$$p(U_e = t)p(U_i > t)p(U_a > t)p(U_p > t). \tag{19}$$

If any of the late units arrive first, the respective action is generated with probability:

$$\text{Antisaccade}: \; p(U_a = t)p(U_p > t)p(U_e > t)p(U_i > t). \tag{20}$$

$$\text{Prosaccade}: \; p(U_p = t)p(U_a > t)p(U_e > t)p(U_i > t). \tag{21}$$

Finally, if the inhibitory unit arrives first, either a late pro- or antisaccade is generated with probability

$$\text{Antisaccades}: \; p(U_a = t)p(U_p > t)\left(\int_0^t p(U_i = \tau)p(U_e > \tau)d\tau\right), \tag{22}$$

$$\text{Prosaccades}: \; p(U_p = t)p(U_a > t)\left(\int_0^t p(U_i = \tau)p(U_e > \tau)d\tau\right). \tag{23}$$

Implicit in the last two terms is the competition between the late units, which are assumed again to be independent of each other. Formally, this competition is expressed as the probability of, for example, the late antisaccade unit arriving before a late prosaccade $p(U_a = t)p(U_p > t)$. A schematic representation of the model is shown in Fig 2. This late race is similar to the Linear Ballistic Accumulation model proposed by [24], although in that model decisions are seen as the result of a race of ballistic accumulation processes with fixed threshold but stochastic base line and increase rate. Here we only assume that the late decision process is a GO-GO race [21].

The likelihood of an action is given by summing over all possible outcomes that lead to that action:

$$p(A = pro, T = t) = p(U_e = t)p(U_i > t)p(U_a > t)p(U_p > t) +$$

$$p(U_p = t)p(U_a > t)p(U_i > t)p(U_e > t) + p(U_p = t)p(U_a > t)\left(\int_0^t p(U_i = \tau)p(U_e > \tau)d\tau\right), \tag{24}$$

$$p(A = anti, T = t) = p(U_a = t)p(U_p > t)p(U_i > t)p(U_e > t) +$$

$$p(U_a = t)p(U_p > t)\left(\int_0^t p(U_i = \tau)p(U_e > \tau)d\tau\right). \tag{25}$$

We have left out some possible simplifications in Eqs 24 and 25 for the sake of clarity.

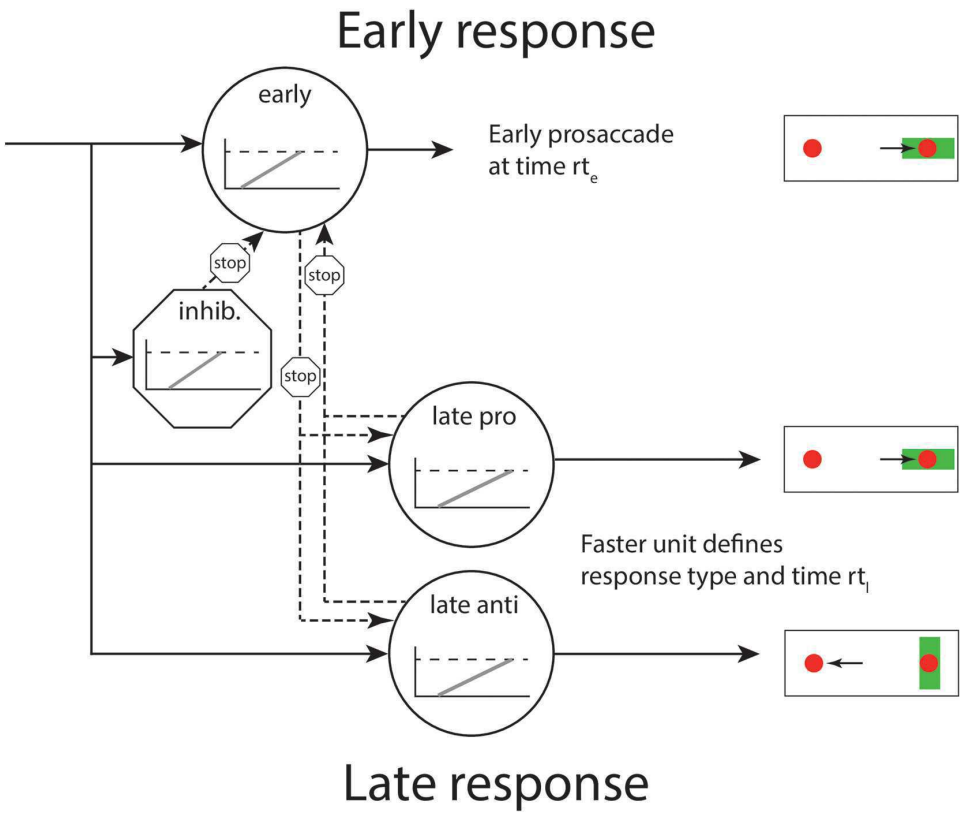


Fig 2. Layout of the SERIA$_{lr}$ model. The presentation of a visual cue (a green bar) triggers the race of four independent units. The inhibitory unit can stop an early response. The late decision process is triggered by the competition between two further units.

https://doi.org/10.1371/journal.pcbi.1005692.g002

The conditional probability of a late antisaccade is given by the interaction between the late units, such that

$$p(U_a < U_p) = \int_0^\infty p(U_a = t)p(U_p > t)dt = 1 - p(U_p < U_a), \tag{26}$$

is analogous to the probability of a late antisaccade $1-\pi_l$ in the SERIA model. This observation shows that the main difference between the SERIA and SERIA$_{lr}$ model is that the former postulates that the distribution of late pro- and antisaccades are equal and conditionally independent of the action performed, whereas the latter constrains the probability of a late antisaccade to be a function of the arrival times of the late units.

The expected *response time* of late pro- and antisaccade actions is given by

$$\frac{1}{p(U_p < U_a)} \int_0^\infty t\, p(U_p = t)p(U_a > t)dt, \tag{27}$$

$$\frac{1}{p(U_a < U_p)} \int_0^\infty t\, p(U_a = t) p(U_p > t) dt. \tag{28}$$

We will refer to these terms as the mean *response time* of pro- and antisaccade actions, in contrast to the mean arrival times, which are the expected value of any single unit.

Non-decision time

The models above can be further finessed to account for non-decision times δ by transforming the RT t to $t_\delta = t - \delta$. The delay δ might be caused by, for example, conductance delays from the retina to the cortex. In addition, the antisaccade or late units might include a constant delay δ_a, which is often referred to as the antisaccade cost [1]. Note that the model is highly sensitive to δ because any RT below it has zero probability. In order to relax this condition and to account for early outliers, we assumed that saccades could be generated before δ at a rate $\eta \in [0,1]$ such that the marginal likelihood of an outlier is

$$p(T < \delta) = p(T_\delta < 0) = \eta. \tag{29}$$

For simplicity, we assume that outliers are generated with uniform probability in the interval $[0,\delta]$:

$$p(T = t) = \frac{\eta}{\delta} \text{ if } t < \delta. \tag{30}$$

Furthermore, we assume that the probability of an early outlier being a prosaccade was approximately 100 times higher than being an antisaccade. Because of the new parameter η, the distribution of saccades with a RT larger than δ needs to be renormalized by the factor $1-\eta$. In the case of the PROSA model, for example, this means that the joint distribution of action and RT is given by the conditional probability

$$p(A = pro, T = t_\delta | t_\delta > 0) = p(U_p = t_\delta) p(U_a > t_\delta - \delta_a) p(U_s > t_\delta), \tag{31}$$

$$p(U_a < 0) = 0, \tag{32}$$

$$p(A = anti, T = t_\delta | t_\delta > 0) = p(U_a = t_\delta - \delta_a) \left(p(U_p > t_\delta) + \int_0^{t_\delta} p(U_p = \tau) p(U_s < \tau) d\tau \right). \tag{33}$$

A similar expression holds for the SERIA models. However, in the PROSA model a unit-specific delay is equal to an action-specific delay. By contrast, in the SERIA model both early and late responses can generate pro- and antisaccades. Thus, δ_a represents a delay in the late actions that affects both late pro- and antisaccades.

Parametric distributions of the increase rate

The models discussed in the previous sections can be defined independently of the distribution of the rate of each of the units. In order to fit experimental data, we considered four parametric distributions with positive support for the rates: gamma [13], inverse gamma, lognormal [25] and the truncated normal distribution (similarly to [22] and [24]). Table 1 and Fig 3 summarize these distributions, their parameters, and the corresponding arrival time densities. We considered five different configurations: 1) all units were assigned *inverse gamma* distributed rates, 2) all units were assigned *gamma* distributed rates, 3) the increase rate of the prosaccade

Table 1. Parametric density functions of the increase rates.

Name	Parameters	Rate p.d.f.	Arrival time p.d.f.
Gamma	k,θ	$\frac{\theta^{-k}}{\Gamma(k)} e^{-r/\theta} r^{k-1}$	$\frac{\theta^{k}}{\Gamma(k)} e^{-\theta/t} t^{-k-1}$
Inv. gamma	k,θ	$\frac{\theta^{k}}{\Gamma(k)} e^{-\theta/r} r^{-k-1}$	$\frac{\theta^{-k}}{\Gamma(k)} e^{-t/\theta} t^{k-1}$
Log normal	μ,σ^2	$\frac{1}{\sqrt{2\pi}\sigma r} e^{-\frac{1}{2}\left(\frac{\ln r-\mu}{\sigma}\right)^2}$	$\frac{1}{\sqrt{2\pi}\sigma t} e^{-\frac{1}{2}\left(\frac{\ln t+\mu}{\sigma}\right)^2}$
T. normal	μ,σ^2	$\frac{1}{Z} e^{-\frac{1}{2}\left(\frac{r-\mu}{\sigma}\right)^2}$	$\frac{1}{Z t^2} e^{-\frac{1}{2}\left(\frac{t^{-1}-\mu}{\sigma}\right)^2}$

Z is the normalization constant $Z = \int_0^\infty exp\left(-\frac{(r-\mu)^2}{2\sigma^2}\right) dr$.

https://doi.org/10.1371/journal.pcbi.1005692.t001

and stop units (or early and the inhibitory units) was *gamma distributed* but the antisaccade (late) unit's increase rate was *inverse gamma* distributed, 4) all the units were assigned *lognormal* distributed rates or 5) all units were assigned *truncated normal* distributed rates.

All the parametric distributions considered here can be fully characterized by two parameters which we generically refer to as k and θ. Hence, the PROSA model is characterized by the parameters for each unit $k_p,k_a,k_s,\theta_p,\theta_a,\theta_s$. The SERIA model can be characterized by analogous parameters $k_e,k_l,k_i,\theta_e,\theta_l,\theta_i$ and the probabilities of early and late prosaccades π_e and π_l. In the case of the SERIA_{lr} model, the probability of a late prosaccade is replaced by the parameters of a late prosaccade unit k_p, θ_p. In addition to the unit parameters, all models included the non-decision time δ, the antisaccade (or late unit) cost δ_a, and the marginal rate of early outliers η.

Experimental procedures

In this section, we describe the experimental procedures, statistical methods, and inference scheme used to invert the models above. The data is from the placebo condition of a larger pharmacological study that will be reported elsewhere.

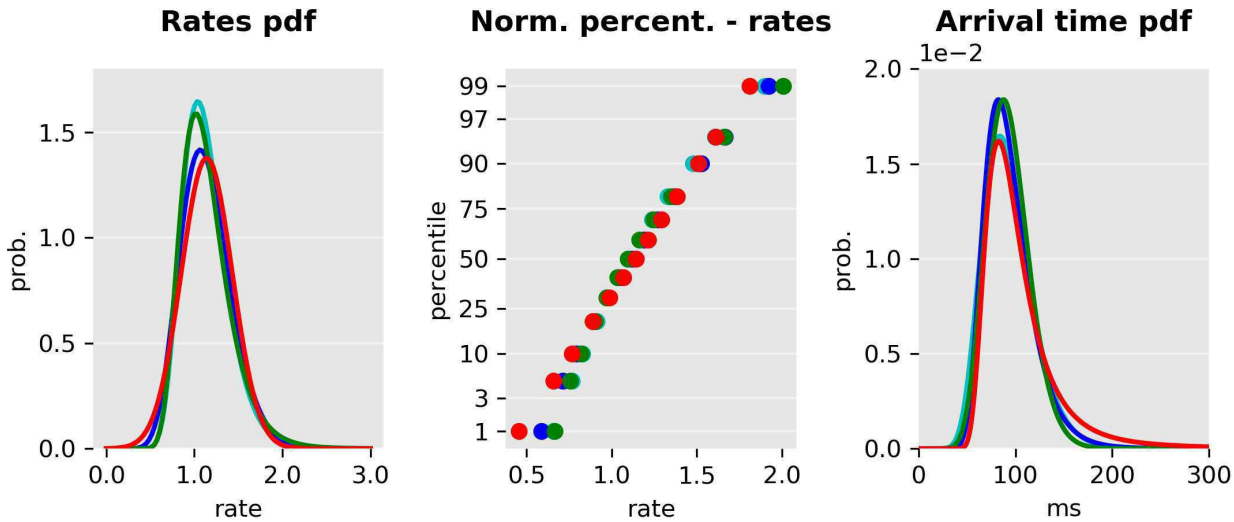


Fig 3. Illustration of probability distributions used to model increase rates. Left: Distribution of the rates based on different probability density functions: Normal (red), gamma (blue), inverse gamma (green), and log-normal (cyan). All distributions were matched to have equal mean and variance. Center: Probit plots of the same distributions. While the gamma and lognormal distributions are very close to the straight line induced by the normal distribution, the inverse gamma distribution diverges slightly more from linearity. Right: Arrival time distribution (scaled to ms).

https://doi.org/10.1371/journal.pcbi.1005692.g003

Participants. Fifty-two healthy adult males naïve to the antisaccade task were invited to a screening session through the recruitment system of the Laboratory of Social and Neural Systems Research of the University of Zurich. During screening, and after being debriefed about the experiment, subjects underwent an electrocardiogram, a health survey, a visual acuity test, and a color blindness test. Subjects were excluded if any of the following criteria were met: age below 18 or above 40 years, regular smoking, alcohol consumption the day before the experiment, any possible interaction between current medication and levodopa or benserazide, pulse outside the range 55–100bpm, recreational drug intake in the past 6 months, history of serious mental or neurological illness, or if the medical doctor supervising the experiment deemed the participant not apt. All subjects gave their written informed consent to participate in the study and received monetary compensation.

Procedure. Each subject was invited to two sessions. During both visits, the same experimental protocol was followed. After arrival, placebo or levodopa (Madopar DR 250, 200mg of levopa + 50 mg benserazide) was orally administered in the form of shape- and color-matched capsules. The present study is restricted to data from the session in which subjects received placebo. Participants and experimenters were not informed about the identity of the substance. Immediately afterwards subjects were introduced to the experimental setup and to the task through a written document. This was followed by a short training block (see below).

The experiment started 70 minutes after substance administration. Subjects participated in three blocks of 192 randomly interleaved pro- and antisaccade trials. The percentages of prosaccade trials in the three blocks were 20%, 50%, or 80%. This yielded three *prosaccade probability* (PP) conditions: PP20, PP50, and PP80. Thus, in the PP20 block, subjects were presented a prosaccade cue in 38 trials, while in all other 154 trials an antisaccade cue was shown. The order of the trials was randomized in each block, but the same order was used in all subjects and sessions. The order of the conditions was counterbalanced across subjects.

Stimulus and apparatus. During the experiment, subjects sat in front of a CRT monitor (Philipps 20B40, distance eye-screen: $\simeq 60cm$, refresh rate: 75Hz). The screen subtended a horizontal visual angle of 38 degrees of visual angle (dva). Eye movements were recorded using a remote infrared camera (Eyelink II, SR-Research, Canada). Participants' head was stabilized with a chin rest. Data were stored at a sampling rate of 500 Hz.

During the task, two red dots (0.25dva) that constituted the saccadic targets were constantly displayed at an eccentricity of ±12dva. Displaying the saccadic target before the execution of an antisaccade has been reported to affect saccadic velocity and accuracy, but not RTs [26], and arguably decreases the need for sensorimotor transformations [27]. At the beginning of each trial, a gray fixation cross (0.6×0.6 *dva*) was displayed at the center of the screen. After a random fixation interval (500 to 1000 ms), the cross disappeared, and the cue instructing either a pro- or an antisaccade trial (see below) was shown centered on either of the red dots. As mentioned above, in each block, subjects were presented with a prosaccade cue in either 20, 50, or 80 percent of the trials. The order of the presentation of the cues was randomized. The cue was a green rectangle ($3.48 \times 0.8 dva$) displayed for 500ms in either horizontal (prosaccade) or vertical orientation (antisaccade). Once the cue was removed and after 1000ms, the next trial started.

Subjects were instructed to saccade in the direction of the cue when a horizontal bar was presented (prosaccade trial) and to saccade in the opposite direction when a vertical bar was displayed (antisaccade trial, see Fig 4). See [28,29] for similar task designs.

Prior to the main experiment, participants were trained on the task in a block of 50 prosaccade trials, immediately followed by 50 antisaccade trials. During the training, subjects were automatically informed after each trial whether their response was correct or not (see below), or whether they had failed to produce a saccade within 500ms after cue presentation (CP). No feedback was given during the main experimental blocks.

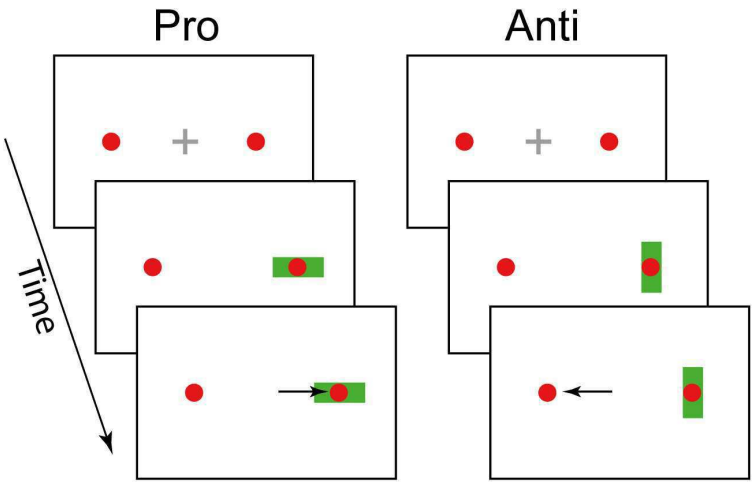


Fig 4. Task design. After a variable fixation period of 500–1000ms (top) the cue (green rectangle) appeared on the screen for 500 ms. The orientation of the cue (horizontal or vertical) indicated the required response (prosaccade or antisaccade).

https://doi.org/10.1371/journal.pcbi.1005692.g004

Data preparation. Data were parsed and preprocessed using the Python programming language (2.7). Saccades were detected using the algorithm provided by the eyetracker manufacturer (SR Research), which uses a velocity and acceleration threshold of 22*dva*/*s* and 3800dva/s^2 [30]. We only considered saccades with a magnitude larger than 2*dva*. RT was defined as the time between CP and the first saccade larger than 2*dva*. A prosaccade trial was considered correct if the end position of the saccade was ipsilateral to the cue and, conversely, an antisaccade trial was considered correct if the end position of the saccade was contralateral to the cue.

Trials were excluded from further analysis if a) data were missing, b) a blink occurred between CP and the main saccade, c) the trial was aborted by the experimenter, d) subjects failed to fixate in the interval between fixation detection and CP, e) if a saccade was detected only later than 800ms after CP, f) if the RT was below 50ms, and in the case of an antisaccade if it was below 110ms. Corrective antisaccades were defined as saccades that a) followed a prosaccade error, b) occurred no later than 900*ms* after CP, and c) had less than 3*dva* horizontal error from the red circle contralateral to the cue.

Besides the fitted non-decision time δ we assumed a fixed non-decision time of 50*ms* for all participants [17]. This was implemented by subtracting 50*ms* of all saccades before being entered into the model. In order to avoid numerical instabilities, RT were rescaled from millisecond to tenths of a second during all numerical analysis. All results are presented in ms.

Modeling

We aimed to answer three questions with the models considered here. First, we investigated which model (i.e. PROSA, SERIA or $SERIA_{lr}$) explained the experimental data best, and whether all important qualitative features of the data were captured by this model. We did not have a strong hypothesis regarding the parametric distribution of the data and hence, comparisons of parametric distributions were only of secondary interest in our analysis. Second, we investigated whether reduced models that kept certain parameters fixed across trial types were

Table 2. Model families with the respective increase-rate distributions.

	PROSA		
Model	**Prosaccade/ stop units**	**Anti. unit**	**# Param. full/const.**
m_1/m_1^c	Inv. gamma	Inv. gamma	15/13
m_2/m_2^c	Gamma	Gamma	15/13
m_3/m_3^c	Gamma	Inv. gamma	15/13
m_4/m_4^c	Lognorm.	Lognorm.	15/13
m_5/m_5^c	T. norm.	T. norm.	15/13
	SERIA		
	Early/stop units	**Late unit**	
m_6/m_6^c	Inv. gamma	Inv. gamma	19/13
m_7/m_7^c	Gamma	Gamma	19/13
m_8/m_8^c	Gamma	Inv. gamma	19/13
m_9/m_9^c	Lognorm.	Lognorm.	19/13
m_{10}/m_{10}^c	T. norm.	T. norm.	19/13
	$SERIA_{lr}$		
	Early/stop units	**Late pro./anti. units**	
m_{11}/m_{11}^c	Inv. gamma	Inv. gamma	19/15
m_{12}/m_{12}^c	Gamma	Gamma	19/15
m_{13}/m_{13}^c	Gamma	Inv. gamma	19/15
m_{14}/m_{14}^c	Lognorm.	Lognorm.	19/15
m_{15}/m_{15}^c	T. norm.	T. norm.	19/15

Models with parameters constrained to be equal across trial types are referred through the superscript *c*.

https://doi.org/10.1371/journal.pcbi.1005692.t002

sufficient to model the data. Third, we investigated how the probability of a trial type in a block affected the parameters of the model.

Model space. Initially, we considered 15 different models as shown in Table 2. Each model was fitted independently for each subject and condition. Since our experimental design included mixed blocks, we allowed for different parameters in pro- and antisaccade trials, i.e., different increase-rate distributions depending on the *trial type* (TT). Under this hypothesis, the PROSA model had 12 free parameters (6 for each TT), whereas the SERIA model required 4 further parameters (π_e and π_l in each TT). The late race $SERIA_{lr}$ model included 16 parameters for the units (8 for each TT). We did not investigate the case that early reactions could trigger antisaccades but rather fixed the probability of an early antisaccade $1-\pi_e$ to 10^{-3}. The rationale behind this was that if early reactions are a priori assumed to never trigger antisaccades, rare but possible early antisaccades might cause large biases when fitting a model.

Regarding the non-decision time δ, antisaccade cost δ_a, and rate of outliers η, we assumed equal parameters in both TT. Consequently, the full PROSA model had 15 free parameters whereas the full SERIA and $SERIA_{lr}$ models had both 19 free parameters.

In addition to the full models, we evaluated restricted versions of each of them by constraining some parameters to be equal across TT. In the case of the SERIA model, we hypothesized that the parameters of all units were equal, irrespective of TT (i.e., that the rate of the units was not affected by the cue presented in a trial). However, we assumed that the probability that an early or late response was a prosaccade was different in pro- and antisaccade trials. Therefore, in the case of the SERIA model, instead of 12 unit parameters (6 per TT), the restricted model had only 6 parameters for the units' rates. The parameters π_e and π_l were allowed to differ in

pro- and antisaccade trials. In the case of the restricted $SERIA_{lr}$ model, the units that underlie the late decision process were allowed to vary across TT, yielding a restricted model with 4 parameters for the early and inhibitory units, and 8 for the late decision process, half of them for each trial type. In the case of the PROSA model, similarly to [17], it is possible to assume that the parameters of the prosaccade unit remain constant across TT, and that the parameters of the stop and antisaccade units depend on TT, yielding 10 parameters for the units.

Prior distributions for model parameters. To complete the definition of the models, the prior distribution of the parameters was specified. This distribution reflects beliefs that are independent of the data and provides a form of regularization when inverting a model. In order to avoid any undesired bias regarding the parametric distributions considered here, we reparametrized all but the truncated normal distribution in terms of their mean and variance. We then assumed that the log of the mean and variance of the rate of the units were equally normally distributed (see Table 3). Therefore, the parametric distributions had the same prior in terms of their first two central moments. In the case of the truncated normal distribution, instead of an analytical transformation between its first two moments and its natural parameters μ and σ^2, we defined the prior distribution as a density of μ and $\ln \sigma^2$. To ensure that μ was positive with high probability (96%) we assumed that $\mu \sim N(0.55,0.09)$. The variance term was distributed as displayed in Table 3. As a further constraint, we restricted the parameter space to enforce that the first two moments of the distributions of rates and RTs existed. We relaxed this constraint for the late units of the $SERIA_{lr}$ in order to allow for 'flat' distributions with possibly infinite mean and variance. This can describe a case in which the increase rate of one of the late units is extremely low.

For the non-decision time δ and the antisaccade cost δ_a, the prior of their log transform was a normal distribution, consistent across all models. Note that the scale of the parameters δ and δ_a in Table 3 is tenths of a second. The fraction of early outliers η, and early and late prosaccades π_e and π_l were assumed to be Beta distributed, with parameters 0.5 and 0.5. Thus, for example, the prior probability of an early outlier is given by

$$p(\eta) \propto \eta^{0.5}(1-\eta)^{0.5}. \tag{34}$$

This parametrization constitutes the minimally informative prior distribution, as it is the Jeffrey's prior of η, π_e and π_l. Table 3 displays the parameters used for the prior distributions.

Bayesian inference

Inference on the model parameters was performed using the Metropolis-Hastings algorithm [31]. To increase the efficiency of our sampling scheme, we iteratively modified the proposal distribution during an initial 'burn-in' phase as proposed by [32]. Moreover, we extended this method by drawing from a set of chains at different temperatures and swapping samples across chains. This method, called population MCMC or parallel tempering, increases the statistical

Table 3. Prior probability density functions.

Parameter	Probability density function	Expected value	Variance
μ_r	$\mathcal{N}(\ln\mu_r; -1.08, 0.97)$	0.55	0.5
σ_r^2	$\mathcal{N}(\ln\sigma_r^2; -2.64, 0.69)$	0.1	0.01
δ	$\mathcal{N}(\ln\delta; -1.58, 1.79)$	0.5	1.25
δ_a	$\mathcal{N}(\ln\delta_a; -0.87, 1.17)$	0.75	1.25
π_e	*Beta*(π_e;0.5,0.5)	0.5	0.145
π_l	*Beta*(π_e;0.5,0.5)	0.5	0.145
η	*Beta*(η;0.5,0.5)	0.5	0.145

https://doi.org/10.1371/journal.pcbi.1005692.t003

efficiency of the Metropolis-Hasting algorithm [33] and has been used in similar contexts before [34]. We simulated 16 chains with a 5-th order temperature schedule [35]. For all but the models including a truncated normal distribution, we drew 4.1×10^4 samples per chain, from which the first 1.6×10^4 samples were discarded as part of the burn-in phase. When a truncated normal distribution was included (models m_5, m_{10}, and m_{15}), the total number of samples was increased to 6×10^4, from which 2×10^4 were discarded. The convergence of the algorithm was assessed using the Gelman-Rubin criterion [33,36] such that the $\tilde{R}$ statistic of the parameters of the model was aimed to be below 1.1. When a simulation did not satisfy this criterion, it was repeated until 99.5 percent of all simulations satisfied it.

Models were scored using their log marginal likelihood or log model evidence (LME). This is defined as the log probability of the data given a model after marginalizing out all its parameters. When comparing different models, the LME corresponds to the log posterior probability of a model under a uniform prior on model identity. Thus, for a single subject with data y, the posterior probability of model k, given models 1 to n is

$$p(m_k|y) = \frac{p(y|m_k)p(m_k)}{\sum_{i=1}^{n} p(y|m_i)p(m_i)} = \frac{p(y|m_k)}{\sum_{i=1}^{n} p(y|m_i)}. \quad (35)$$

Importantly, this method takes into account not only the accuracy of the model but also its complexity, such that overparameterized models are penalized [37]. A widely used approximation to the LME is the Bayesian Information Criterion (BIC) which, although easy to compute, has limitations (for discussion, see [38]). Here, we computed the LME through thermodynamic integration [33,39]. This method provides robust estimates and can be easily computed using samples obtained through population MCMC.

One important observation here is that the LME is sensitive to the prior distribution, and thus can be strongly influenced by it [40]. We addressed this issue in two ways: On one hand and as mentioned above, we defined the prior distribution of the increase rates of all models in terms of the same mean and variance. This implies that the priors were equal up to their first two moments, and hence all models were similarly calibrated. On the other hand, we complemented our quantitative analysis with qualitative posterior checks [33] as shown in the results section.

Besides comparing the evidence of each model, we also performed a hierarchical or random effects analysis described in [38,41]. This method can be understood as a form of soft clustering in which each subject is assigned to a model using the LME as assignment criterion. Here, we report the expected probability of the model r_i, which represents the percentage of subjects that is assigned to the cluster representing model i. This hierarchical approach is robust to population heterogeneity and outliers, and complements reporting the group-level LME. Finally, we compared families of models [42] based on the evidence of each model for each subject summed across conditions.

Classical statistics

In addition to a Bayesian analysis of the data, we used classical statistics to investigate the effect of our experimental manipulation on behavioral variables (mean RT and ER) and the parameters of the models. We have suggested previously [11,43,44] that generative models can be used to extract hidden features from experimental data that might not be directly captured by, for example, standard linear methods or purely data driven machine learning techniques. In this sense, classical statistical inference can be boosted by extracting interpretable data features through Bayesian techniques.

Frequentist analyses of RT, ER, and parameter estimates were performed using a mixed effects generalized linear model with independent variables *subject* (SUBJECT), *prosaccade*

probability (PP) with levels PP20, PP50 and PP80, and when pro- and antisaccade trials were analyzed together, *trial type* (TT). The factor SUBJECT was always entered as a random effect, whereas PP and TT were treated as categorical fixed effects. In the case of ER, we used the probit function as link function.

Analyses were conducted with the function *fitglme.m* in MATLAB 9.0. The significance threshold α was set to 0.05.

Implementation

All likelihood functions were implemented in the *C* programming language using the GSL numerical package (v.1.16). Integrals without an analytical form or well-known approximations were computed through numerical integration using the Gauss-Kronrod-Patterson algorithm [45] implemented in the function *gsl_integration_qng*. The sampling routine was implemented in MATLAB (v. 8.1) and is available as a module of the open source software package TAPAS (www.translationalneuromodeling.org/tapas).

Results

Behavior

Forty-seven subjects (age: 23.8 ± 2.9) completed all blocks and were included in further analyses. A total of 27072 trials were recorded, from which 569 trials (2%) were excluded (see Table 4).

Both ER and RT showed a strong dependence on PP (Fig 5 and Table 5). Individual data is included in the S1 Dataset and is displayed in S1 Fig. The mean RT of correct pro- and antisaccade trials was analyzed independently with two ANOVA tests with factors SUBJECT and PP. We found that in both pro- ($F_{2,138} = 46.9$, $p < 10^{-5}$) and antisaccade trials ($F_{2,138} = 37.3$, $p < 10^{-5}$) the effect of PP was significant; with higher PP, prosaccade RT decreased, whereas the RT of correct antisaccades increased. On a subject-by-subject basis, we found that between the PP20 and PP80 conditions, 91% of the participants showed increased RT in correct antisaccade trials, while 81% demonstrated the opposite effect (a decrease in RT) in correct prosaccade trials. Similarly, there was a significant effect of PP on ER in both prosaccade ($F_{2,138} = 376.1$, $p < 10^{-5}$) as well as in antisaccade ($F_{2,138} = 347.0$, $p < 10^{-5}$) trials. This effect was present in all but one participant in antisaccade trials and in all subjects in prosaccade trials. Exemplary RT data of one subject in the PP50 condition is displayed in Fig 6.

Modeling

Model comparison results. Initially, we considered the models outlined in Table 2. The LME over all participants (fixed effects analysis) and the posterior probability of all models

Table 4. Summary of trials per subject.

	Valid	Blink	Missing	Aborted	FE	Late S.	Early S.	Total
Total	26503	188	60	42	249	0	30	27072
Mean	563.9	4.0	1.3	0.9	5.3	0.0	0.6	576
Std.	9.9	5.1	2.5	1.5	5.0	0.0	1.3	-
Min.	536	0	0	0	0	0	0	-
Max.	576	22	15	6	19	0	8	-

FE: Fixation errors. Late saccades are saccades elicited after 800ms. Early saccades are prosaccades elicited before 50ms after CP or antisaccades elicited before 110ms after CP.

https://doi.org/10.1371/journal.pcbi.1005692.t004

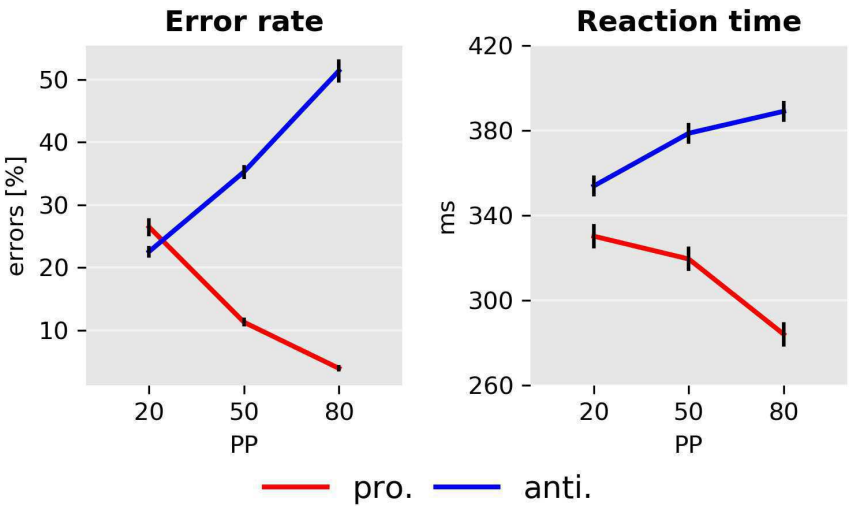


Fig 5. Error rate and mean reaction time as a function of prosaccade trial probability (PP). Left panel: Mean error rates for pro- and antisaccade trials. Right panel: Mean reaction time in *ms*. Error bars indicate standard errors of the mean. Only correct responses are displayed.

https://doi.org/10.1371/journal.pcbi.1005692.g005

and all subjects are presented in Fig 7. Independently of the particular parametric distribution of the units, the SERIA_{lr} models had higher evidence compared to the PROSA and SERIA models. A random effects, family-wise model comparison [42] resulted in an expected frequency of $r = 87\%$ for the SERIA_{lr} family, $r = 11\%$ for the SERIA family, and $r = 2\%$ for the PROSA family. In addition, constraining the parameters to be equal across trial types increased the model evidence irrespective of the parametric distribution assigned to the units (Fig 7). Here, the family-wise model comparison showed that models with constrained parameters had an expected frequency of $r = 98\%$. Over all 30 models, m^{c}_{13} (SERIA_{lr} with constrained parameters, early and inhibitory increase rates gamma distributed, and late units' rate inverse gamma distributed) showed the highest LME with $\Delta LME > 200$ compared to all other models. Following [40], a difference in LME larger than 3 corresponds to strong evidence.

To verify that the SERIA_{lr} family was not preferred simply because the probability of early prosaccades was fixed, we considered models in the SERIA family with the same property (not displayed). We found that although fixing this value increased the LME of the SERIA family,

Table 5. Summary of mean RTs and ERs.

Trial type	Action	Reaction times [ms]		
		PP 20	PP 50	PP80
Pro.	Pro.	330(72)	319(67)	284(59)
Pro.	Anti.	326(68)	329(46)	336(57)
Anti.	Anti	354(60)	378(57)	389(61)
Anti.	Pro	234(50)	231(47)	225(31)
		Error rates [%]		
Pro.		26(15)	11(8)	4(4)
Anti.		23(17)	35(21)	51(20)

Standard deviations are shown in brackets.

https://doi.org/10.1371/journal.pcbi.1005692.t005

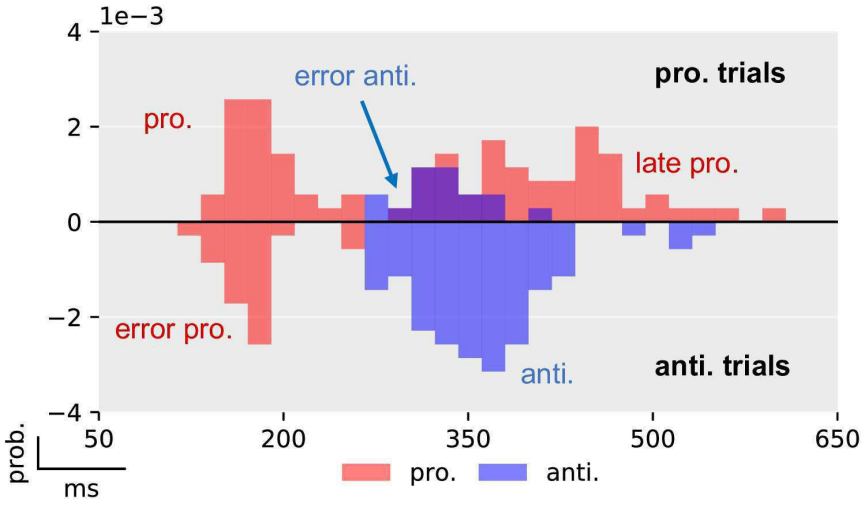


Fig 6. Exemplary histogram of the reaction times of one subject in the PP50 condition. Prosaccade trials are displayed in the upper half plane and antisaccade trials in the lower (negative) half plane. Prosaccade actions are depicted in red color, whereas antisaccade actions are shown in blue. Errors in prosaccade trials are antisaccades that for this subject occurred after the first peak of early prosaccades. Errors in antisaccade trials (lower half plane) occurred at a similar latency as early prosaccades in prosaccade trials. The histograms have been normalized to have unit probability mass, i.e., the sum of the area of all bars is one.

https://doi.org/10.1371/journal.pcbi.1005692.g006

there was still a difference of $\Delta LME > 90$ when comparing the best model of the SERIA_{lr} family and the best model of the SERIA family with a fix probability of early prosaccades.

Fits of four subjects using the posterior samples of the best PROSA (m_1), SERIA (m_8^c), and SERIA_{lr} (m_{13}^c) models are depicted in Fig 8. Although model m_1 was the best model in the PROSA family, it clearly did not explain the apparent bimodality of the prosaccade RT distributions. Instead, RTs were explained through wider distributions. No obvious difference could be observed between the SERIA and SERIA_{lr} models. We further examined the model fits in Fig 9 and Fig 10 by plotting the weighted fits and cumulative density functions of the reciprocal RT in the probit scale (reciprobit plot [22]) collapsed across subjects for the best model of each family. The histograms of RTs clearly show a large number of late prosaccades whose distribution is similar to the distribution of antisaccade RTs. The most pronounced–but still small—difference between the SERIA and SERIA_{lr} models was visible in prosaccade trials in the PP20 condition (left panel, upper half plane), in which antisaccade errors displayed lower RT than correct late prosaccades.

Corrective antisaccades. The RTs of antisaccades that follow an error prosaccade were not directly modeled. However, we hypothesized that corrective antisaccades are delayed late antisaccade actions, whose distribution is given by the *response time* distribution of late antisaccades

$$\frac{1}{p(U_a < U_p)} p(U_a = t) p(U_p > t) \tag{36}$$

A total of 2989 corrective antisaccades were included in the analysis. The mean (±std) end time of the erroneous prosaccades was 268(±63)ms. The mean RT of corrective antisaccades was 447(±103)ms, and the weighted mean arrival time of the late antisaccade unit was 367ms. Fig 11 displays the histogram of the end time of all prosaccade errors, the RT of all corrective

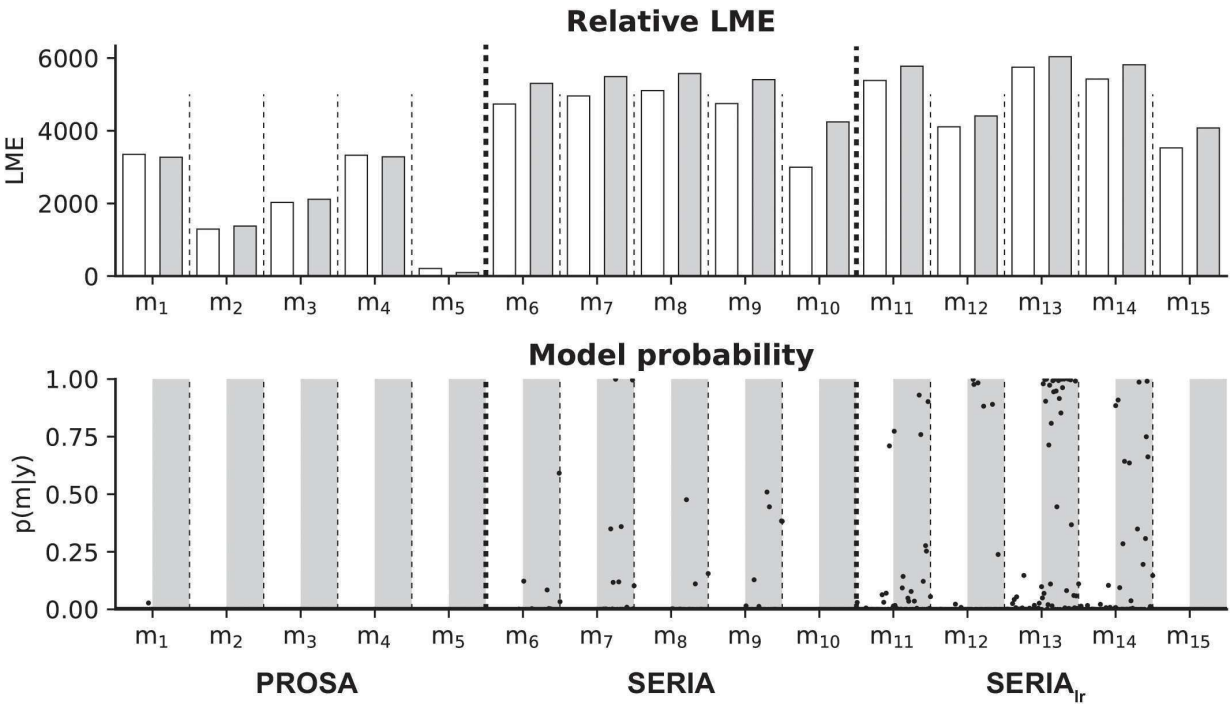


Fig 7. Summary of model comparison. Top: Summed LME of all subjects for all 30 models. White bars show models with all parameters free, grey bars models with constrained parameters. LMEs are normalized by subtracting the lowest LME (m_5). Model m^c_{13} (constrained $SERIA_{lr}$) exceeded all other models ($\Delta LME > 200$). Bottom: Illustration of model probability for all subjects. The posterior model probabilities for all subjects are shown as black dots. In white shading are models with all parameters free, grey bars represent models with restricted parameters. Note that in nearly all subjects, the $SERIA_{lr}$ models with restricted parameters showed high model probabilities.

https://doi.org/10.1371/journal.pcbi.1005692.g007

antisaccades and the time shifted (+80ms) predicted response time of late antisaccades. Since we did not have a strong hypothesis regarding the magnitude of the delay of the corrective antisaccades, we selected the time shift to be the difference between the mean corrective antisaccade RT and the mean predicted response time of late antisaccades. Visual inspection strongly suggests that the distribution of corrective antisaccade RTs is well approximated by the distribution of the late responses. The short difference between corrective antisaccades' RT and the expected response time of the late antisaccade unit (80ms) favors the hypothesis that the plan for a corrective antisaccade is initiated before the incorrect prosaccade had finished.

Effects of prosaccade probability on model parameters. The effect of PP on the parameters of the model was investigated by examining the expected value of the parameters of the best scoring model (m^c_{13}). Initially, we considered the question of whether the mean arrival or response time of each of the units changed as a function of PP. For arrival times, this corresponds to

$$\frac{1}{N}\sum_{j=1}^{N}(E[U_i|k^i_j, \theta^i_j] + \delta^i_j) \tag{37}$$

where i is an index over the units, j is an index over N samples collected using MCMC, and δ^i_j is the estimated delay. In the case of the late units, we considered only the response time of correct actions. Fig 12 left displays the mean arrival and response times. These were submitted to

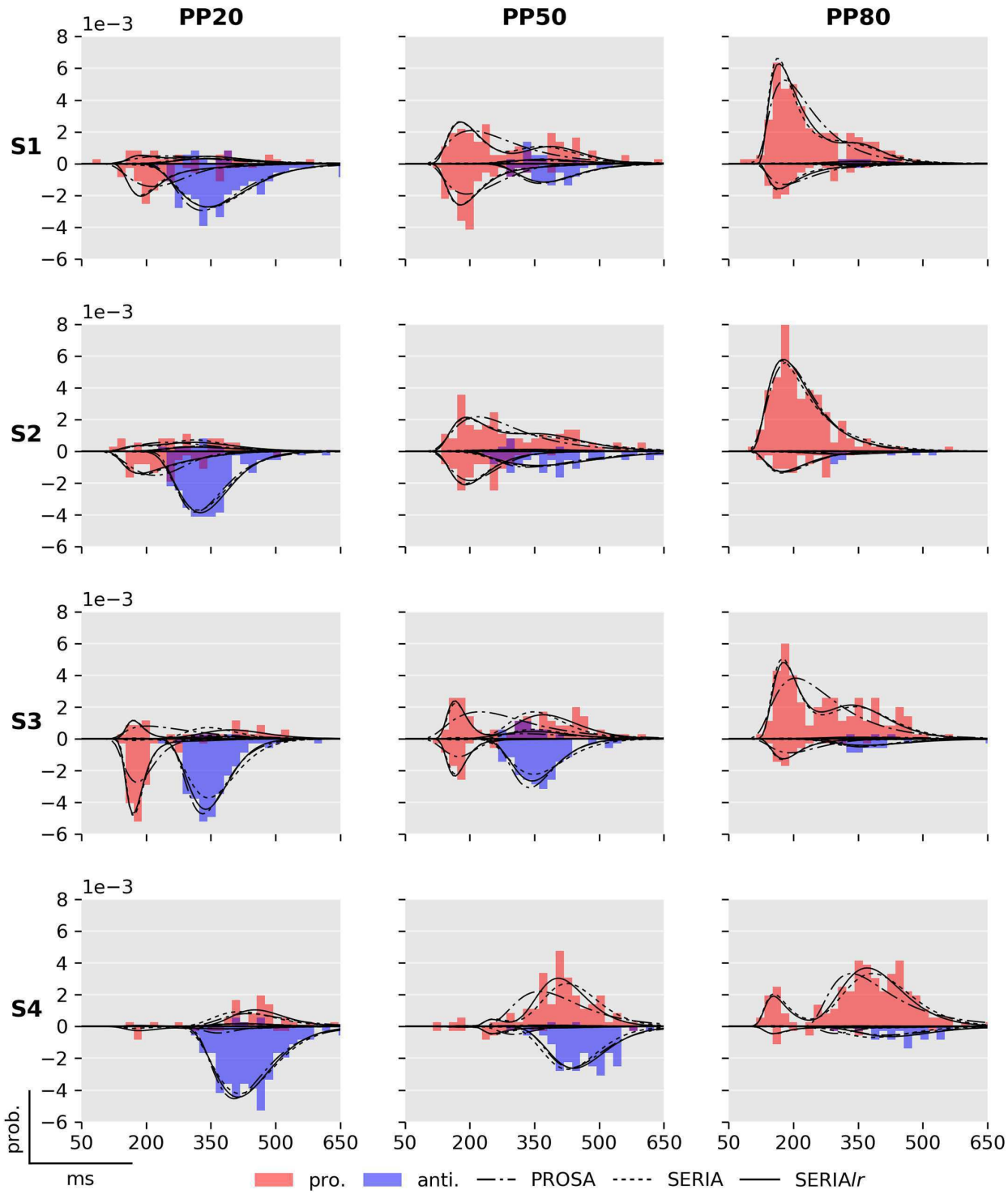


Fig 8. Fits of best PROSA (m_1), SERIA (m_8^c) and SERIA$_{lr}$ (m_{13}^c) models. Columns display the normalized histogram of the RTs of pro- (red) and antisaccades (blue) in each of the conditions. Rows correspond to individual subjects named S1 to S4 for display purpose. As in Fig 6, prosaccade

trials are displayed on the upper half plane, whereas antisaccade trials are displayed in the lower half plane. The predicted RT distributions based on the samples from the posterior distribution are displayed in solid ($SERIA_{lr}$), broken (SERIA), and dash-dotted (PROSA) lines. Note that data from subject 3 in the PP50 condition is the same as shown in Fig 6. Early outliers are not displayed.

https://doi.org/10.1371/journal.pcbi.1005692.g008

four separate ANOVA tests, which revealed that PP had a significant effect on all four units: early unit ($F_{2,138} = 9.2$, $p < 10^{-3}$), late antisaccade ($F_{2,138} = 26.6$, $p < 10^{-3}$), late prosaccade ($F_{2,138} = 19.6$, $p < 10^{-3}$), and inhibitory unit ($F_{2,138} = 30.9$, $p < 10^{-3}$). We then explored the differences across conditions through planned post hoc tests on each condition for each of the units (Table 6). The arrival times of the early unit did not change significantly between condition PP20 and PP50, but decreased significantly in the PP80 condition as compared to the PP50 block. The response times of late antisaccades increased significantly between the PP20 and the PP50 conditions but not so between the PP50 and PP80. Late prosaccades followed the opposite pattern, showing only a significant decrease in response time between the PP50 and PP80 conditions. Finally, the inhibitory unit changed significantly across all conditions.

Finally, we examined how the probability of a late antisaccade $p(U_a > U_p)$ (Fig 12, right) depended on PP and TT. The estimated parameters for both pro- and antisaccade trials were analyzed with a model with factors SUBJECT, TT, PP and the interaction between TT and PP. An ANOVA test demonstrated that both PP ($F_{2,276} = 51.2$, $p < 10^{-3}$) and TT ($F_{1,276} = 985.0$, $p < 10^{-3}$) had a significant effect, but there was no evidence for an interaction between the two factors ($F_{2,276} = 1.5$, $p < 0.23$), suggesting that PP affected the probability of a late antisaccade equally in pro- and antisaccade trials.

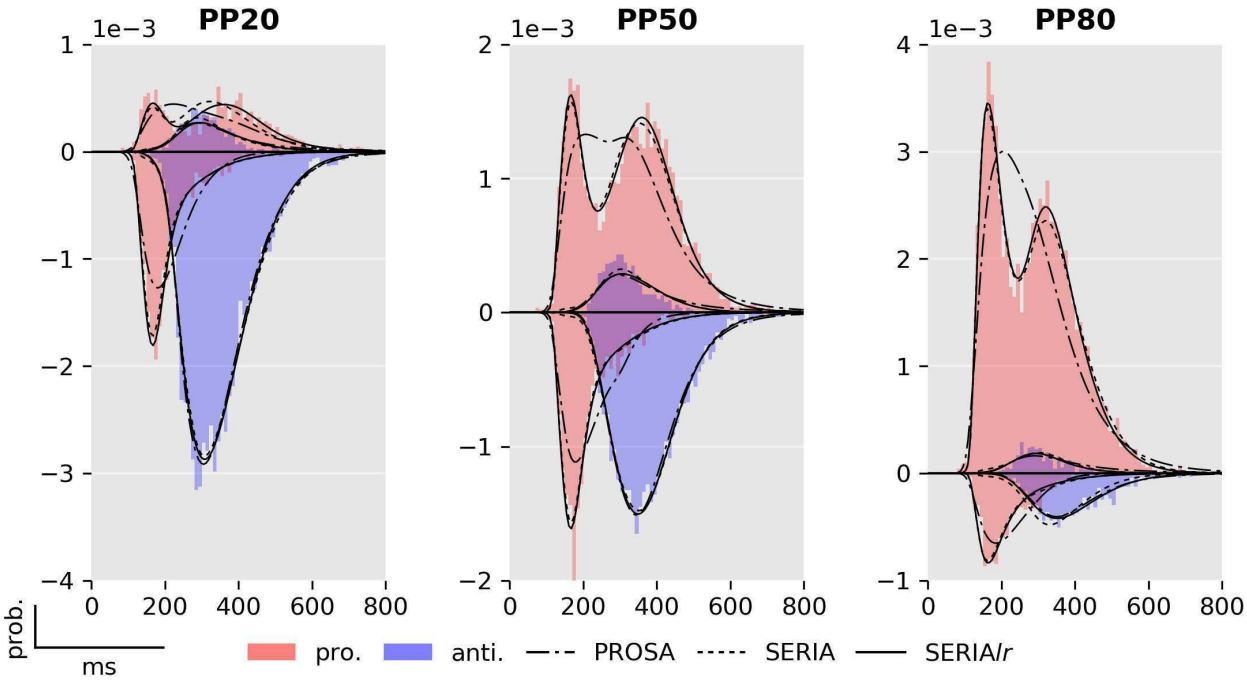


Fig 9. Fits from the best models in each family (m_1, m_8^c, m_{13}^c). Model fits and RT histograms for each condition collapsed across subjects. For more details see Figs 6 and 8.

https://doi.org/10.1371/journal.pcbi.1005692.g009

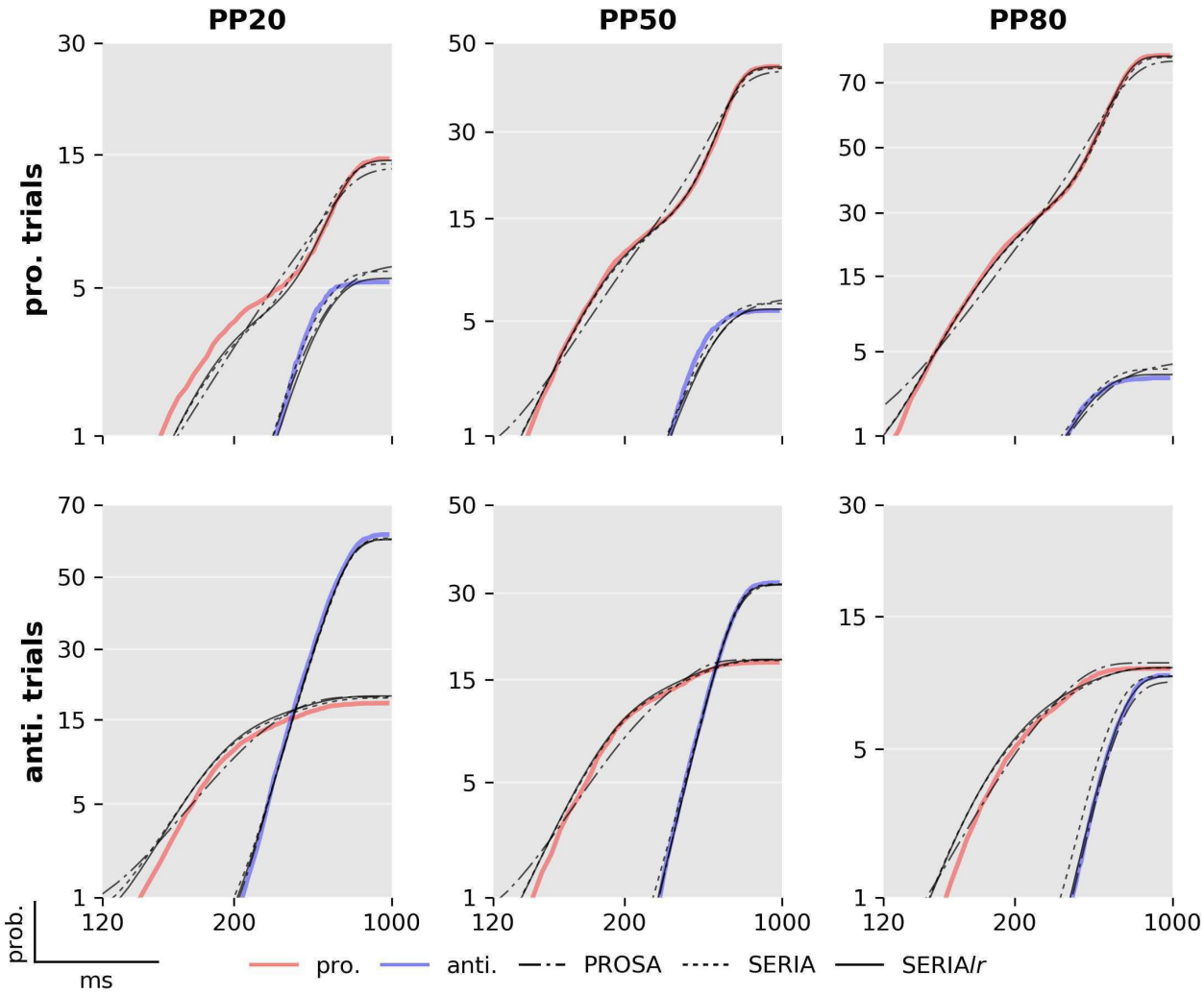


Fig 10. Reciprobit plot of best models. Predicted and empirical cumulative density function of the reciprocal RT in the probit scale for each condition and model collapsed across all subjects. The data shown are the same as in Fig 9, but split for trial types and illustrated as cumulative distributions. Note that the y-axis is in the probit scale and that nearly all differences between the model and the data occur at very small probability values of 5% or below.

https://doi.org/10.1371/journal.pcbi.1005692.g010

Subject specific parameters

Finally, we investigated how some of the parameters of the model were related to each other across subjects. Because it has been commonly reported that schizophrenia is related with higher ER, but also with increased antisaccade RT, an interesting question is whether higher late-action response times are correlated with the percentage of late errors and inhibition failures, i.e., early saccades that are not stopped. We found that the response time of late pro ($F_{1,135} = 13.6$, $p < 0.001$) and antisaccades ($F_{1,135} = 7.1$, $p < 0.01$) was negatively correlated with the probability of a late error (Fig 13), but no significant interaction between PP and response time was found (pro: $F_{2,135} = 1.7$, $p = 0.19$; anti: $F_{2,135} = 0.3$, $p = 0.76$). Hence, late responders tended to make fewer late errors, suggesting a speed/accuracy trade-off in addition to the main effect of PP. We further considered the question whether the percentage of

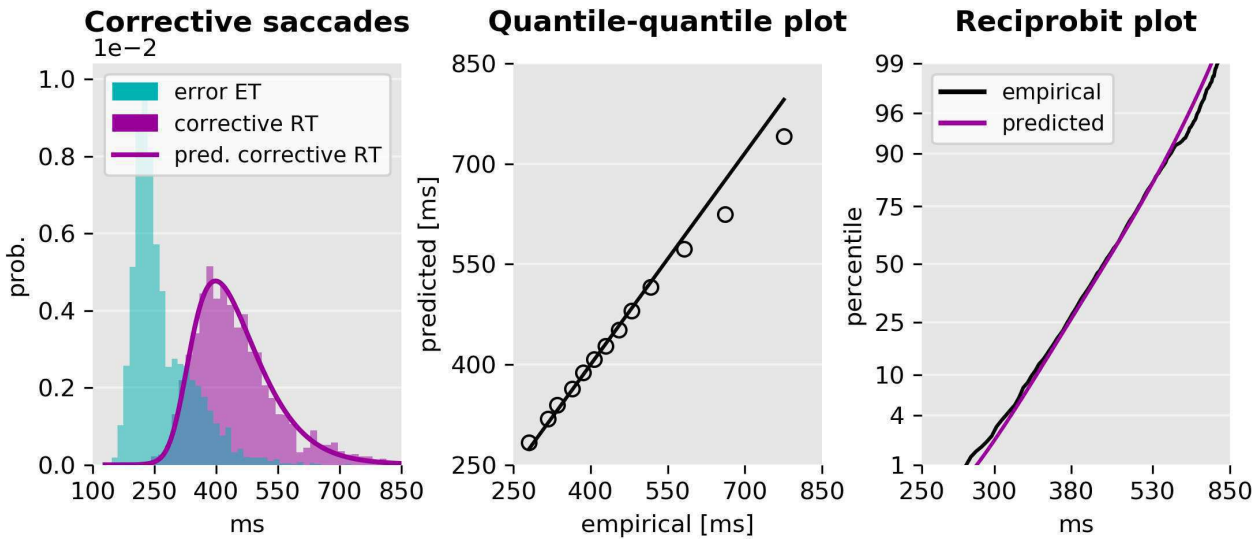


Fig 11. Empirical and predicted RT of corrective antisaccades. Left: End time of erroneous prosaccades, RTs of corrective antisaccades, and time shifted predicted response time distribution of late antisaccades. The time shift was selected to be the difference between the empirical and predicted mean response time. Center: Quantile-quantile plot of the predicted and empirical distribution of corrective antisaccades, and a linear fit to the central 98% quantiles. There is a small deviation only at the tail of the distribution. Right: Reciprobit plot of the empirical and predicted cumulative density functions of the RT of corrective antisaccades. The scale of the horizontal axis is proportional to the reciprocal RT. The vertical axis is in the probit scale.

https://doi.org/10.1371/journal.pcbi.1005692.g011

inhibition failures was correlated with the expected arrival time of the late antisaccade unit in antisaccade trials (Fig 13 right). Note that the number of inhibition failures is the same in both trial types in a constrained model, but inhibition failures are errors in antisaccade trials and correct early reactions in prosaccade trials. We found that these parameters were not

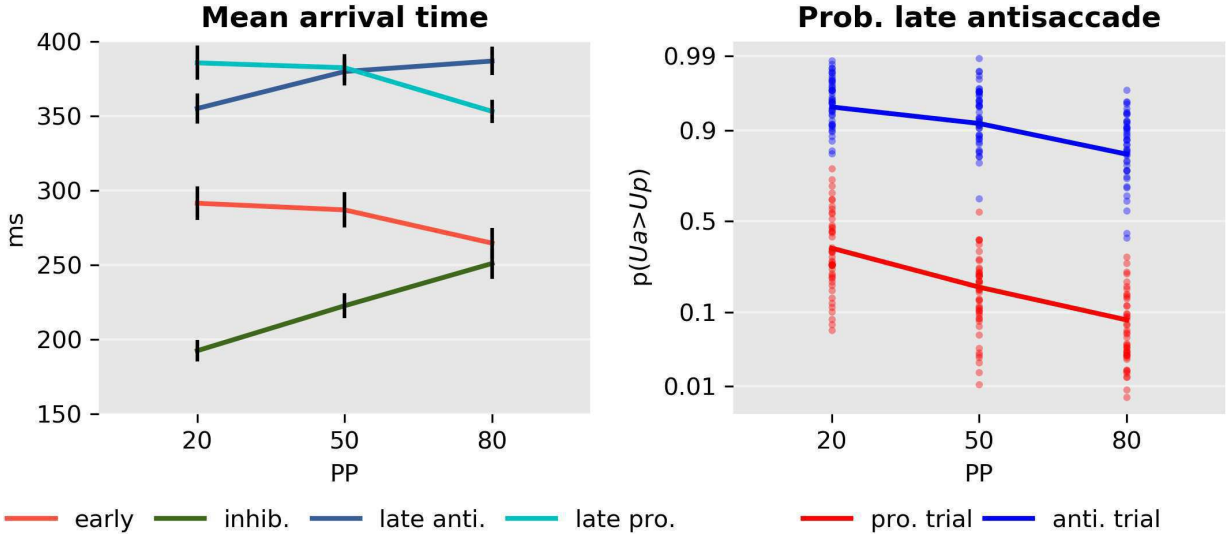


Fig 12. Model parameters. Left: Mean arrival or response time and standard error of the early and inhibitory units and late pro- and antisaccades. Right: Probability of a late antisaccade $p(U_a > U_p)$ in prosaccade (red) and antisaccade (blue) trials in each condition in the probit scale.

https://doi.org/10.1371/journal.pcbi.1005692.g012

Table 6. Post hoc comparison of the effect of PP.

	Early			Inhib.			Late pro.			Late anti.		
	Mean [ms]	t_{138}	p	Mean [ms]	t_{138}	p	Mean [ms]	t_{138}	p	Mean [ms]	t_{138}	p
PP20–PP50	4	0.6	0.50	-30	-4.0	<0.001*	3	0.5	0.55	-24	-36	<0.001*
PP20–PP80	26	3.9	<0.001*	-58	-7.8	<0.001*	32	5.7	<0.001*	-32	-5.4	<0.001*
PP50–PP80	22	3.3	0.001*	-28	-3.8	<0.001*	29	5.1	<0.001*	-7	-1.5	0.13

Effect of PP on the mean arrival time of the early and inhibitory unit, and for late pro- and antisaccade units in the corresponding trial type.
*, $p < 0.05$.of a two-tailed t-test.

https://doi.org/10.1371/journal.pcbi.1005692.t006

significantly correlated ($F_{2,135} = 1.2$, $p = 0.26$). This was also the case when we considered the expected response time of late prosaccades in prosaccade trials (not displayed; $F_{2,135} = 0.0$, $p = 0.98$).

Fig 14 illustrates the posterior distribution of late errors and inhibition failures of two representative subjects as estimated using MCMC. Clearly, PP induced strong differences in the percentage of inhibition failures and late errors in prosaccade trials in both subjects. The effect of PP is less pronounced in late errors in antisaccade trials. The posterior distributions also illustrate how the $SERIA_{lr}$ model can capture individual differences: For example, the percentage of late prosaccade errors in the PP80 condition and the percentage of inhibition failures across all conditions are clearly different in each subject.

Discussion

In this study, we provided a formal treatment of error rates (ER) and reaction times (RT) in the antisaccade task using probabilistic models. We applied these models to data from an experiment consisting of 3 mixed blocks with different probabilities of pro- and antisaccades trials. Model comparison showed that a novel model that allows for late pro- and antisaccades

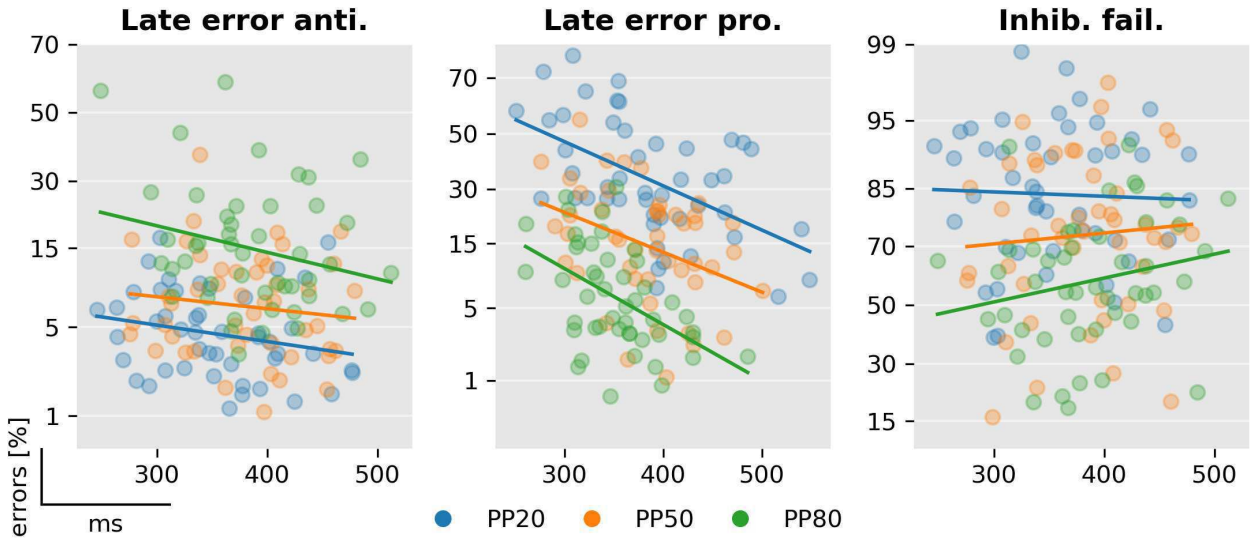


Fig 13. Correlation between late arrival times and errors. Left: Percentage of late errors against late antisaccades' response times in antisaccade trials. Center: Percentage of late errors against late prosaccades' response time in prosaccade trials. Left: Percentage of inhibitory failures against late antisaccades' response time in antisaccade trials. The vertical axis is in the probit scale.

https://doi.org/10.1371/journal.pcbi.1005692.g013

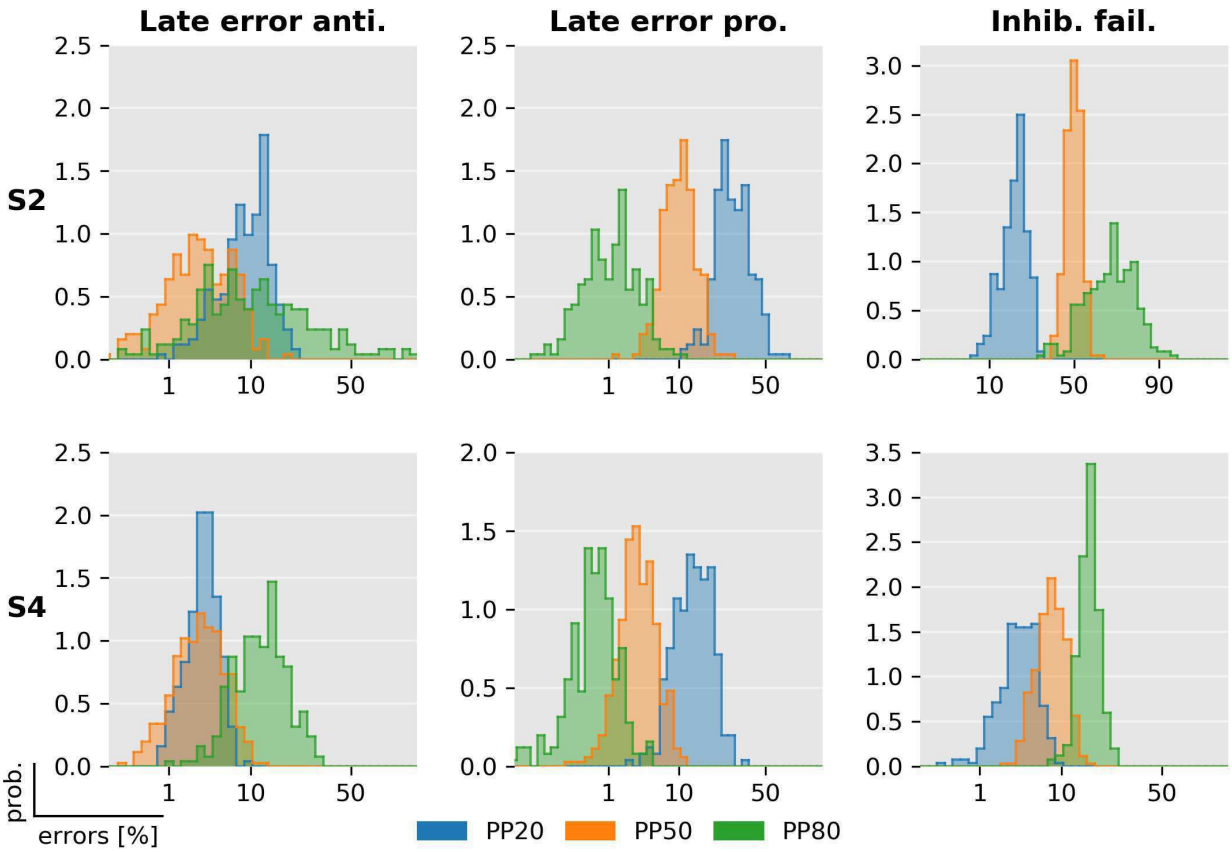


Fig 14. Posterior distribution of late errors and inhibition failures. The posterior distribution of the percentage of late and inhibition failures of two exemplary subjects (see Figs 8 and 13). Samples from the posterior distribution were obtained using MCMC. Histograms display the distributions of the samples in probit scale (horizontal axis). For these two subjects, the posterior distribution of late prosaccade and inhibitory failures clearly discriminates between the three PP conditions.

https://doi.org/10.1371/journal.pcbi.1005692.g014

explains our experimental findings better than a model in which all late responses are assumed to be antisaccades. The parameter estimates of the hidden units of the model showed that changes in the inhibitory unit and the late decision process explained most of the overt changes in behavior caused by our experimental manipulation, i.e., differences in trial type probability. Moreover, we found that all units were sensitive to the PP in a block, although late responses tended to plateau when the corresponding trial type was not highly frequent.

Our main finding is that two decision processes are necessary to properly model the antisaccade task: on one hand, an early race between a prepotent response towards a target and an endogenously generated signal to cancel this action, and, on the other hand, a secondary late race between two units encoding the cue-action mapping. Although the late decision process can be closely approximated by assuming that RT and actions are independent (at least in our experimental design), Bayesian model comparison demonstrated that late decisions are more accurately described by a race between two units representing different actions. The two decision processes are the sources of early errors–fast prosaccades in antisaccade trials- and late errors–late actions incongruent with the cue presented. The late decision process displays a speed/accuracy tradeoff and is biased by the probability of a trial type in a block. Moreover,

this decision process predicts the RT distribution of corrective antisaccades that follow early errors. Because the extra latency of these corrective antisaccades (80ms) is relatively short, it is unlikely that corrective antisaccades are due to a restart in the decision process. Rather these are late actions that overwrite early errors.

Influence of trial type probability on reaction times and error rates

Our results show that both RT and ER depend on PP. While this was a highly significant factor in our study, there are mixed findings in previous reports. ER in antisaccade trials was found to be correlated with TT probability in several studies [29,46,47]. However, this effect might depend on the exact implementation of the task [47,48]. Changes in prosaccade ER similar to our study have been reported by [29] and [48]. Studies in which the type of saccade was signaled at fixation prior to the presentation of the peripheral cue do not always show this effect [47]. The results on RTs are less consistent in the literature. Our findings of increased anti- and decreased prosaccade RTs with higher PP are in line with the overall trend in [29], and with studies in which the cue was presented centrally [47]. Often, there is an additional increase in RT in the PP50 condition [29,47], which was visible in our data as a slight increase in RT in the PP50 condition on top of the linear effect of PP. Overall, RTs in our study were relatively slow compared to studies in which the TT cue was separated from the spatial cue [46,47]. However, a study with a similar design and added visual search reported even slower RTs in both pro- and antisaccades [29].

Interpretation of model comparison results

Formal comparison of generative models can offer insight into the mechanisms underlying eye movement behavior [11] and might be relevant in translational neuromodeling applications, such as computational psychiatry [49–53]. Here, we have presented what is, to our knowledge, the first formal statistical comparison of models of the antisaccade task. For this, we formalized the model introduced in [17] and proceeded to develop a novel model that relaxes the one-to-one association of early and late responses with pro- and antisaccades, respectively. All models and estimation techniques presented here are openly available under the GPLv3.0 license as part of the open source package TAPAS (www.translationalneuromodeling.org/tapas).

Bayesian model comparison yielded four conclusions at the family level. First, the SERIA models were clearly favored when compared to the PROSA models. Second, including a late race between actions representing late pro- and antisaccades ($SERIA_{lr}$) resulted in an increase in model evidence, compared to a model not including a late race (SERIA). Third, models in which the race parameters of the early and inhibitory unit were constrained to be equal across TT had a higher LME than models in which all parameters were free. Hence, the effect of the cue in a single trial was limited to the late action, and did not affect the race between an early and inhibitory process. This constitutes an important external validation, as it means that model comparison does favor a model which respects the temporal order of the experiment: Information about TT is only available after the stimulus was presented and, thus, it is unlikely to have an impact on fast reactive responses. Fourth, early responses were nearly always pro-saccades. Crucially, these four conclusions are based on family-wise comparison across all parametric distribution of the increase rate of the units.

A further consequence of our findings is that two independent and qualitatively different decision processes lead to an antisaccade: the race process between early and inhibitory units, and the secondary decision process that generates late responses. A separation of decisions into a 'where' and a 'when' component has been proposed by [54], but mainly in conceptual

terms. However, model comparison showed that these two components ('where' and 'when') cannot be completely dissociated and that time plays a role in late decisions. Nevertheless, the assumption that action type and arrival time of late responses were independent yielded a good fit to this particular data set, suggesting that it is, in many cases, an acceptable approximation to assume a time-independent late decision process. The most obvious difference between the SERIA and SERIA_{lr} can be observed in prosaccade trials in the PP20 condition (left panel, upper half plane Fig 9), in which late prosaccades are slower than antisaccades. We discuss this point in more detail below.

Parametric distribution of reaction times. The parametric distribution of oculomotor RTs has been discussed in great detail in the literature (e.g., [13,55]). Here, we did not aim at determining the most suitable distribution, but rather opted for a practical approach by evaluating different models with a reduced number of parametric distributions. We then based our conclusions on the model with the highest LME. Nevertheless, one can consider the connection of the models presented here with other families of parametric distributions. In particular, the linear relationship

$$\frac{s_i}{r_i} = t \tag{38}$$

could be seen as formally inconsistent with the observation that RT are likely to be explained by stochastic accumulation processes (see for example [56,57], but [58]). This is a weaker constraint than one would expect, because under low noise conditions, for example, a linear relationship can be a good approximation of neural activity. Even if the relationship is not linear, for any continuous function ϕ with an inverse function ϕ^{-1}, the model can be recasted as [59]:

$$s_i = \phi(tr_i), \tag{39}$$

$$\frac{\phi^{-1}(s_i)}{r_i} = t. \tag{40}$$

In any case, linear accumulation models have been shown to yield similar conclusions to stochastic accumulation models [58].

More generally, it can be shown that if RTs follow a generalized inverse normal distribution (GIN) of the form

$$GIN(t;\lambda,\kappa,\psi) = \frac{\left(\frac{\psi}{\kappa}\right)^{\lambda}}{2K_{\lambda}(\sqrt{\kappa\psi})} t^{\lambda-1}\exp\left(-\frac{1}{2}(\kappa t^{-1}+\psi t)\right) \tag{41}$$

where $\lambda \leq 0$, and K_λ is a modified Bessel function of the second kind, there exists a continuous diffusion process whose first hit distribution (FHD) follows the GIN [60]. A particular case of this distribution is the Wald distribution for which $\lambda = -\frac{1}{2}, \kappa = 0$. It is the FHD of the Brownian diffusion process with drift

$$X_t = -\sqrt{\sigma}\psi t + \sigma W_t \tag{42}$$

where W_t denotes a Wiener process, $x_0 > 0$, and the absorbing boundary a is zero. More relevant here, when $\psi = 0$ the distribution reduces to an inverse gamma distribution, the FHD of the process

$$X_t = \sqrt{\sigma}(2\lambda - 1)t^{-1} + \sigma W_t \tag{43}$$

with $x_0 > 0$ and boundary $a = 0$ (for a detailed mathematical treatment see [60]). Thus, if the rates of a ballistic, linear processes are assumed to be gamma distributed, the RTs follow a distribution that is formally equivalent to a first hit model with stochastic updates and fixed rates. While the model presented here can be seen as a ballistic accumulation model, this equivalence suggests that it is *compatible* with a diffusion process with infinitesimal mean change proportional to t^{-1}.

Other antisaccade models. In broad terms, three families of antisaccade models can be distinguished (reviewed in [61]). The first set of models is based on a race process with independent saccadic and stop units. These models build on the seminal work of [16] on the stop-signal paradigm. According to this model, a 'GO' signal triggers a stochastic 'race' process that generates a response once it reaches threshold. Critically, a stop signal triggers a second process that inhibits the first 'GO' response if it is the first to reach threshold. Moreover, the rates of both units are assumed to be independent. This model was further extended for the antisaccade task by [17] (but see [14,21], and the review in [20]), who included a third unit such that an antisaccade is generated when a reflexive prosaccade is inhibited by an endogenously-triggered stop process. Note that the original 'horse-race' model has also been modified [62] to account for different competing response actions, similarly as in the antisaccade task. The models proposed here belong to this family.

A second type of model relies on lateral or mutual inhibition of competing pro- and antisaccade units. In this direction, Cutsuridis and colleagues [61,63,64] proposed that lateral inhibition is implemented by inhibitory connections in the intermediate layers of the superior colliculus. Thus, saccades are the result of accumulation processes, but these are not independent of each other. Crucially, no veto-like stop signal is required. Although no formal model-fitting has been proposed for this model, qualitative agreement with data suggests that it might capture behavioral patterns relevant in translational applications [64,65]. Since no probabilistic version of this model is available, it is not yet possible to decide on the grounds of model comparison whether mutually dependent or independent race processes best explain current behavioral findings.

Finally, several models that incorporate detailed physiological mechanisms have been proposed [23,66–68]. These models cannot be easily assigned to one of the above categories, as they often employ both an inhibitory mechanism that stops or withholds the reactive responses as well as competition between actions. In addition, while more realistic models possess a more fine-grained representation of the underlying neurobiology, they rely on a large number of parameters and for this reason, it is difficult to fit them to behavioral data (for discussion, see [11]).

Regarding neurobiologically realistic models, the model proposed by [23] is the most similar to the SERIA model. It posits two different mechanisms that interact in the generation of antisaccades: an action-selection module and a remapping module that controls the cue-action mapping. As a consequence, this model allows for the generation of late errors that follow a similar RT distribution as correct antisaccades. Consistent with this observation, the SERIA model can quantitatively distinguish between inhibition and late cue-action mapping errors (Fig 15, left panel). A less obvious similarity between the SERIA model and [23] is that different cues do not lead *directly* to different dynamics in the action module, but only in the so-called 'remapping' module. Furthermore, the incorporation of a late race is conceptually close to the approach proposed by [23], which includes a winner-take-all competition in what we have referred here as late responses. Similarly, our model comparison results show that different cues (i.e., trial types) do not affect the GO/NO-GO process but only the late cue-action mapping.

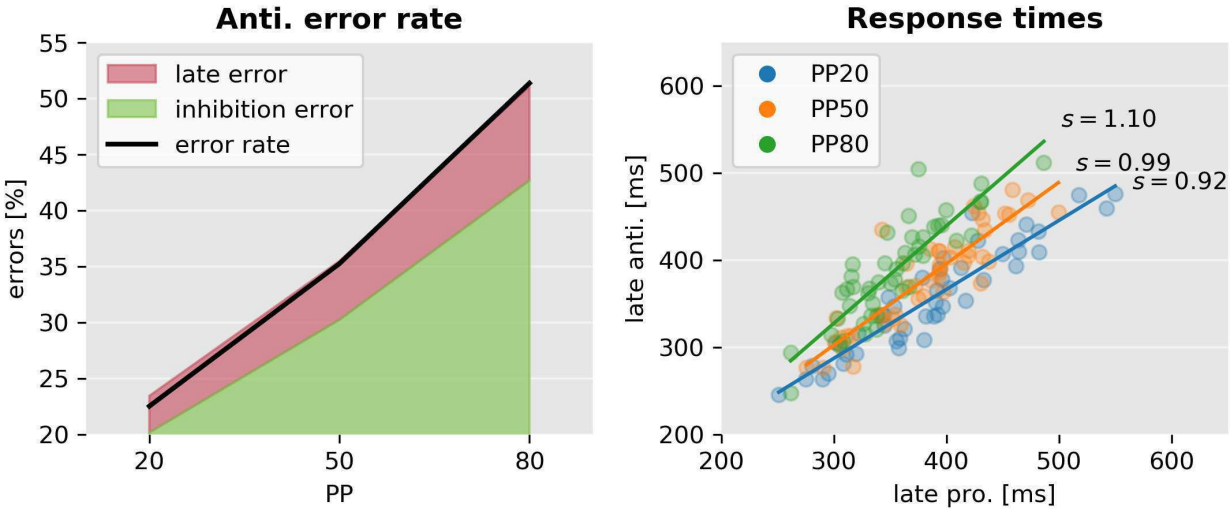


Fig 15. Error sources and the correlation of response times. Left: Error rate (black line) split into the two causes predicted by the model. Inhibition errors are early actions that always trigger prosaccades. Similarly as described by [23], late errors occur when a late response leads to a prosaccade. Right: Correlation between correct antisaccades' and late prosaccades' response times according to the best $SERIA_{lc}$ model. The best linear fit is depicted as a solid line. The mean ratio of pro- and antisaccade response times (s) is displayed on the right. Although late pro- and antisaccade response times are highly correlated, their ratio is different in each condition (interaction PP and late prosaccade response time $F = 9.2$, $p < 0.001$).

https://doi.org/10.1371/journal.pcbi.1005692.g015

Parameter changes across trial types

One of the most salient results presented here is that models in which the parameters of the units were constrained to be equal across trial types had a larger LME than models in which all the parameters were free, suggesting that the early and inhibitory race units were not affected by the cue presented in a single trial. While visual inspection of the predicted likelihood under the posterior parameters showed that most of the prominent characteristics of the data were explained correctly, some more subtle effects were not captured accurately by the SERIA model. This is particularly clear in the PP20 condition, in which the SERIA model displays a large bias in prosaccades trials in the PP20 condition. One possible explanation is that restricting the parameters across trial types made the model too rigid to capture this effect. Fig 16 compares the fitted RT distributions for models m_8 (SERIA) and m_{13} ($SERIA_{lr}$), in which no constraint on the parameters was imposed. Both models are qualitatively almost identical, although as shown in Fig 7, the LME favored the $SERIA_{lr}$ model. Thereby, the qualitative similarity between both models indicates that, in our experiment, the RT of late decisions is only weakly dependent on time. In conclusion, although removing the constraint on the parameters did improve the fit, the differences are marginal and thus did not justify the additional model complexity. As mentioned above this is consistent with the notion that the information about trial type is only available to a subject once the peripheral stimulus (green bar) has been processed, presumably tens of milliseconds after the stimulus onset. In fact, this example illustrates the protection against overfitting provided by the LME, as this is a case in which simpler models were preferred over more complex models despite of slightly less accurate fits.

Arguably, the constrained SERIA model fails to fully capture the RT of late prosaccade in the PP20 and PP50 conditions because of the assumption that late prosaccades have the same arrival time as late antisaccades. As shown in Fig 15, although the response time of late pro-

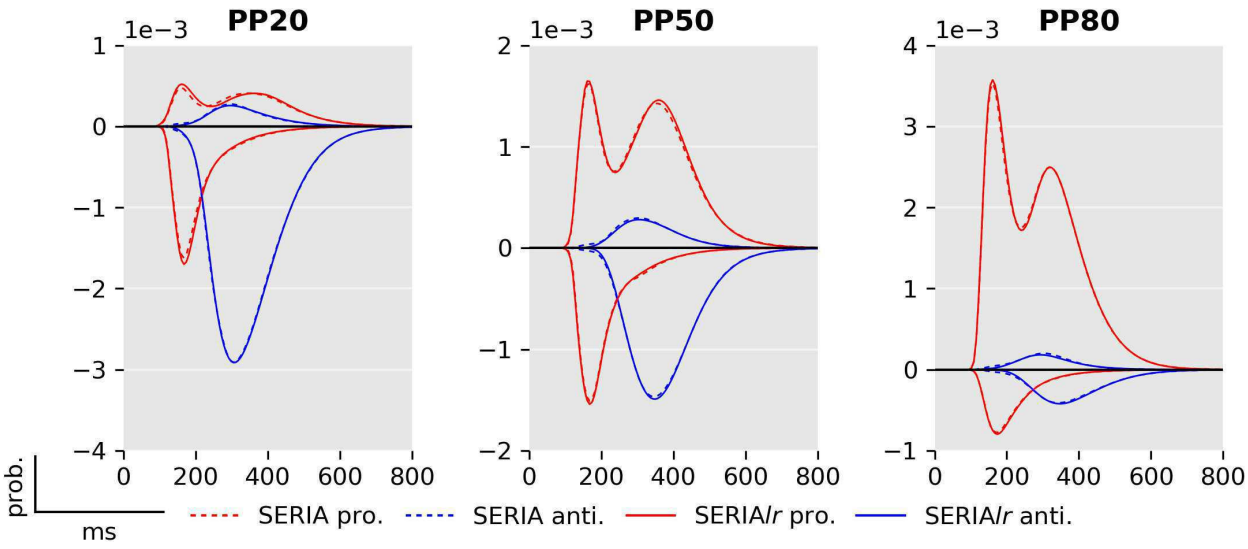


Fig 16. Comparison between unconstrained SERIA and SERIA$_{lr}$ models. Comparison between models m_8 (broken lines; SERIA model) and m_{13} (solid lines; late race SERIA$_{lr}$ model.).

https://doi.org/10.1371/journal.pcbi.1005692.g016

and antisaccades are strongly correlated, the average ratio of the response times changes across conditions.

The effect of trial type probability

It is far from obvious why TT probability affects RT and ER in the antisaccade task. One possible explanation is that increased probability leads to higher preparedness for either pro- or antisaccades. Such a theory posits an intrinsic trade-off between preparations for one of the two action types that leads to higher RTs and ERs in low probability trials. Thus, a trade-off theory predicts that the arrival times of early and late responses should be negatively correlated. Although this hypothesis can explain our behavioral findings in terms of summary statistics, our model suggests a more complicated picture.

The main explanation of our results is the effect of TT probability on the inhibitory unit and the probability of a late prosaccade. A higher probability of antisaccade trials leads to faster inhibition and to a higher number of late prosaccades. This resulted in higher mean RT in prosaccade trials when PP is low. In the case of antisaccades, although the mean arrival times of the late unit increased in the PP50 condition, the increased arrival time of the inhibitory unit on the PP80 condition skewed the antisaccade distribution towards higher RTs. Nevertheless, the SERIA$_{lr}$ implies the anticorrelation of late pro- and antisaccades in a single trial type, as these are the results of a GO-GO race.

Action inhibition

The biological implementation of action inhibition in the antisaccade and other countermanding tasks has received a lot of attention and is still debated [69–73]. Our work adds evidence to the theory that the antisaccade task requires a process that inhibits prepotent responses and is independent of the initiation of a late action [20]. Recent evidence from electrophysiological recordings in the rat brain ([74] reviewed by [71]) suggests that the hypothesized race between

GO and inhibitory responses might be implemented by different pathways in the basal ganglia [68]. In addition to the basal ganglia, microstimulation of the supplementary eye fields tends to facilitate inhibition of saccades in the countermanding task [75].

Corrective antisaccades

Although not a primary goal of our model, we considered the question of predicting corrective antisaccades. This problem has received some attention recently [18,61,65,76], as more sophisticated models of the antisaccade task have been developed. We speculated that corrective antisaccades are generated by the same mechanism as late responses. Thus, their RT distribution should follow a similar distribution. Our results strongly suggest that this is the case (see Fig 11). Moreover, the time delay of the corrective antisaccades indicates that, on average, these actions are not the result of the late unit being restarted at the end time of the erroneous prosaccade, as this would lead to much higher RTs. Rather, the planning of a corrective antisaccade might be started much before the end of the execution of an erroneous prosaccade, in accordance with the parallel planning model of the antisaccade task [46] and the 'GO–STOP +GO' model in [21].

Translational applications

Despite the large number of studies of clinical patients using the antisaccade task, an important question remains open: What are the causes of the errors in different neurological and psychiatric conditions? For example [77,78] argued that errors in schizophrenia might be explained, at least partially, by a failure to generate a secondary late action based on several modifications of the antisaccade task. However, it was also proposed that the increased ER in schizophrenia is due to high tonic dopamine levels in the basal ganglia, that lead to decreased inhibition of early responses [68]. More generally, different neurological and psychiatric diseases, or even patients with the same condition, might be characterized by a different source of errors. For example, there is intriguing evidence [79] that patients with different diseases such as attention deficits disorders [80], Parkinson's disease [81], and amyotrophic lateral sclerosis [82] might be characterized by different ratios of early and late errors. An interesting experimental finding in our study related to this is the considerable amount of erroneous antisaccades in prosaccade trials. An increased number of such errors could be caused by reduced cognitive flexibility leading to impaired shifting between tasks as observed for example in obsessive compulsive disorder [83]. The ability to quantify different types of errors through computational modeling might help to further characterize these diseases.

Summary

Here we have presented a novel model of the antisaccade task. While the basic structure of the model follows the layout of a previous model [17], we have introduced two crucial advancements. First, we postulated that late responses could trigger both pro- and antisaccades, which are selected by an independent decision process. Second, the generative nature of our model allows for Bayesian model inversion, which enables the comparison of different models and families of models on formal grounds. To our knowledge this has not been done for any of the previous models of the antisaccade task, which is of relevance for translational applications that aim at better understanding psychiatric diseases by means of computational modeling.

The application of the model to a large data set yielded several novel results. First, the early and inhibitory race processes triggered by different cues are almost identical. Moreover, different PP had very different effects on the individual units, which was not obvious from the linear analysis of the mean RT and ER. Crucially, our modeling approach allowed us to look at a

mechanistic explanation or the effects of PP by examining the individual units. In future work we aim to disentangle the mechanisms of behavioral differences caused by neuromodulatory drugs and psychiatric illnesses using formal Bayesian inference.

Supporting information

S1 Dataset. Table of data. Spreadsheet including all reaction times, actions and errors that entered the analysis. More details are included in the file.
(XLSM)

S1 Fig. Reaction time and error rate in all conditions for each subject. Mean reaction times and error rates are displayed as solid lines.
(TIFF)

Acknowledgments

We thank Saee Paliwal for her remarks and helpful comments on an earlier version of this manuscript.

Author Contributions

Conceptualization: Eduardo A. Aponte, Dario Schöbi, Klaas E. Stephan, Jakob Heinzle.

Data curation: Eduardo A. Aponte.

Formal analysis: Eduardo A. Aponte, Jakob Heinzle.

Funding acquisition: Klaas E. Stephan.

Investigation: Eduardo A. Aponte.

Methodology: Eduardo A. Aponte, Dario Schöbi, Jakob Heinzle.

Project administration: Jakob Heinzle.

Resources: Klaas E. Stephan.

Software: Eduardo A. Aponte, Dario Schöbi.

Supervision: Jakob Heinzle.

Visualization: Eduardo A. Aponte, Jakob Heinzle.

Writing – original draft: Eduardo A. Aponte, Jakob Heinzle.

Writing – review & editing: Eduardo A. Aponte, Dario Schöbi, Klaas E. Stephan, Jakob Heinzle.

References

1. Hallett PE. Primary and secondary saccades to goals defined by instructions. Vision Res. 1978; 18: 1279–1296. PMID: 726270
2. Munoz DP, Everling S. Look away: the anti-saccade task and the voluntary control of eye movement. Nat Rev Neurosci. 2004; 5: 218–228. https://doi.org/10.1038/nrn1345 PMID: 14976521
3. Hutton SB, Ettinger U. The antisaccade task as a research tool in psychopathology: a critical review. Psychophysiology. 2006; 43: 302–313. https://doi.org/10.1111/j.1469-8986.2006.00403.x PMID: 16805870
4. Fukushima J, Fukushima K, Chiba T, Tanaka S, Yamashita I, Kato M. Disturbances of voluntary control of saccadic eye movements in schizophrenic patients. Biol Psychiatry. 1988; 23: 670–677. PMID: 3370264

5. Curtis CE, Calkins ME, Grove WM, Feil KJ, Iacono WG. Saccadic disinhibition in patients with acute and remitted schizophrenia and their first-degree biological relatives. Am J Psychiatry. 2001; 158: 100–106. https://doi.org/10.1176/appi.ajp.158.1.100 PMID: 11136640
6. Harris MS, Reilly JL, Keshavan MS, Sweeney JA. Longitudinal studies of antisaccades in antipsychotic-naive first-episode schizophrenia. Psychol Med. 2006; 36: 485–494. https://doi.org/10.1017/S0033291705006756 PMID: 16388703
7. Reilly JL, Frankovich K, Hill S, Gershon ES, Keefe RS, Keshavan MS, et al. Elevated antisaccade error rate as an intermediate phenotype for psychosis across diagnostic categories. Schizophr Bull. 2014; 40: 1011–1021. https://doi.org/10.1093/schbul/sbt132 PMID: 24080895
8. Radant AD, Millard SP, Braff DL, Calkins ME, Dobie DJ, Freedman R, et al. Robust differences in anti-saccade performance exist between COGS schizophrenia cases and controls regardless of recruitment strategies. Schizophr Res. 2015; 163: 47–52. https://doi.org/10.1016/j.schres.2014.12.016 PMID: 25553977
9. Crawford TJ, Sharma T, Puri BK, Murray RM, Berridge DM, Lewis SW. Saccadic eye movements in families multiply affected with schizophrenia: the Maudsley Family Study. Am J Psychiatry. 1998; 155: 1703–1710. https://doi.org/10.1176/ajp.155.12.1703 PMID: 9842779
10. Radant AD, Dobie DJ, Calkins ME, Olincy A, Braff DL, Cadenhead KS, et al. Antisaccade performance in schizophrenia patients, their first-degree biological relatives, and community comparison subjects: data from the COGS study. Psychophysiology. 2010; 47: 846–856. https://doi.org/10.1111/j.1469-8986.2010.01004.x PMID: 20374545
11. Heinzle J, Aponte EA, Stephan KE. Computational models of eye movements and their application to schizophrenia. Current Opinion in Behavioral Sciences. 2016; 11: 21–29. https://doi.org/https://doi.org/10.1016/j.cobeha.2016.03.008
12. Wolfe JM, Palmer EM, Horowitz TS. Reaction time distributions constrain models of visual search. Vision Res. 2010; 50: 1304–1311. https://doi.org/10.1016/j.visres.2009.11.002 PMID: 19895828
13. Palmer EM, Horowitz TS, Torralba A, Wolfe JM. What are the shapes of response time distributions in visual search? J Exp Psychol Hum Percept Perform. 2011; 37: 58–71. https://doi.org/10.1037/a0020747 PMID: 21090905
14. Noorani I, Carpenter RH. Full reaction time distributions reveal the complexity of neural decision-making. Eur J Neurosci. 2011; 33: 1948–1951. https://doi.org/10.1111/j.1460-9568.2011.07727.x PMID: 21645090
15. Huys QJ, Maia TV, Frank MJ. Computational psychiatry as a bridge from neuroscience to clinical applications. Nat Neurosci. 2016; 19: 404–413. https://doi.org/10.1038/nn.4238 PMID: 26906507
16. Logan GD, Cowan WB, Davis KA. On the ability to inhibit simple and choice reaction time responses: a model and a method. J Exp Psychol Hum Percept Perform. 1984; 10: 276–291. PMID: 6232345
17. Noorani I, Carpenter RH. Antisaccades as decisions: LATER model predicts latency distributions and error responses. Eur J Neurosci. 2013; 37: 330–338. https://doi.org/10.1111/ejn.12025 PMID: 23121177
18. Noorani I, Carpenter RH. Re-starting a neural race: anti-saccade correction. Eur J Neurosci. 2014; 39: 159–164. https://doi.org/10.1111/ejn.12396 PMID: 24168375
19. Noorani I. LATER models of neural decision behavior in choice tasks. Front Integr Neurosci. 2014; 8: 67. https://doi.org/10.3389/fnint.2014.00067 PMID: 25202242
20. Noorani I. Towards a unifying mechanism for cancelling movements. Philos Trans R Soc Lond, B, Biol Sci. 2017; 372.
21. Camalier CR, Gotler A, Murthy A, Thompson KG, Logan GD, Palmeri TJ, et al. Dynamics of saccade target selection: race model analysis of double step and search step saccade production in human and macaque. Vision Res. 2007; 47: 2187–2211. https://doi.org/10.1016/j.visres.2007.04.021 PMID: 17604806
22. Carpenter RH, Williams ML. Neural computation of log likelihood in control of saccadic eye movements. Nature. 1995; 377: 59–62. https://doi.org/10.1038/377059a0 PMID: 7659161
23. Lo CC, Wang XJ. Conflict Resolution as Near-Threshold Decision-Making: A Spiking Neural Circuit Model with Two-Stage Competition for Antisaccadic Task. PLoS Comput Biol. 2016; 12: e1005081. https://doi.org/10.1371/journal.pcbi.1005081 PMID: 27551824
24. Brown SD, Heathcote A. The simplest complete model of choice response time: linear ballistic accumulation. Cogn Psychol. 2008; 57: 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002 PMID: 18243170
25. Gorea A, Rider D, Yang Q. A unified comparison of stimulus-driven, endogenous mandatory and "free choice" saccades. PLoS ONE. 2014; 9: e88990. https://doi.org/10.1371/journal.pone.0088990 PMID: 24586474

26. Edelman JA, Valenzuela N, Barton JJ. Antisaccade velocity, but not latency, results from a lack of saccade visual guidance. Vision Res. 2006; 46: 1411–1421. https://doi.org/10.1016/j.visres.2005.09.013 PMID: 16260025

27. Zhang M, Barash S. Neuronal switching of sensorimotor transformations for antisaccades. Nature. 2000; 408: 971–975. https://doi.org/10.1038/35050097 PMID: 11140683

28. Sato TR, Schall JD. Effects of stimulus-response compatibility on neural selection in frontal eye field. Neuron. 2003; 38: 637–648. PMID: 12765614

29. Chiau HY, Tseng P, Su JH, Tzeng OJ, Hung DL, Muggleton NG, et al. Trial type probability modulates the cost of antisaccades. J Neurophysiol. 2011; 106: 515–526. https://doi.org/10.1152/jn.00399.2010 PMID: 21543748

30. Stampe D. Heuristic filtering and reliable calibration methods for video-based pupil-tracking systems. Behavior Research Methods, Instruments, \& Computers. Springer-Verlag; 1993; 25: 137–142. https://doi.org/10.3758/BF03204486

31. Robert C, Casella G. Monte Carlo statistical methods. Springer Science \& Business Media; 2013.

32. Shaby B, Wells MT. Exploring an adaptative Metropolis algorithm. Durham, NC, USA: Department of statistical science. Duke Universitiy; 2010.

33. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian Data Analysis. Chapman and Hall/CRC; 2003.

34. Aponte EA, Raman S, Sengupta B, Penny WD, Stephan KE, Heinzle J. mpdcm: A toolbox for massively parallel dynamic causal modeling. J Neurosci Methods. 2016; 257: 7–16. https://doi.org/10.1016/j.jneumeth.2015.09.009 PMID: 26384541

35. Ben Calderhead, Girolami MA. Estimating Bayes factors via thermodynamic integration and population MCMC. Computational Statistics \& Data Analysis. 2009; 53: 4028–4045. https://doi.org/10.1016/j.csda.2009.07.025

36. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Statistical Science. JSTOR; 1992;: 457–472.

37. MacKay DJC. Information Theory, Inference, and Learning Algorithms. Cambridge University Press; 2003.

38. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. Bayesian model selection for group studies. Neuroimage. 2009; 46: 1004–1017. https://doi.org/10.1016/j.neuroimage.2009.03.025 PMID: 19306932

39. Gelman A, Meng XL. Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling. Statistical Science. Institute of Mathematical Statistics; 1998; 13: 163–185. https://doi.org/10.2307/2676756

40. Kass RE, Raftery AE. Bayes factors. Journal of the american statistical association. Taylor \& Francis Group; 1995; 90: 773–795.

41. Rigoux L, Stephan KE, Friston KJ, Daunizeau J. Bayesian model selection for group studies—revisited. Neuroimage. 2014; 84: 971–985. https://doi.org/10.1016/j.neuroimage.2013.08.065 PMID: 24018303

42. Penny WD, Stephan KE, Daunizeau J, Rosa MJ, Friston KJ, Schofield TM, et al. Comparing families of dynamic causal models. PLoS Comput Biol. 2010; 6: e1000709. https://doi.org/10.1371/journal.pcbi.1000709 PMID: 20300649

43. Brodersen KH, Schofield TM, Leff AP, Ong CS, Lomakina EI, Buhmann JM, et al. Generative embedding for model-based classification of fMRI data. PLoS Comput Biol. 2011; 7: e1002079. https://doi.org/10.1371/journal.pcbi.1002079 PMID: 21731479

44. Stephan KE, Schlagenhauf F, Huys QJ, Raman S, Aponte EA, Brodersen KH, et al. Computational neuroimaging strategies for single patient predictions. Neuroimage. 2017; 145: 180–199. https://doi.org/10.1016/j.neuroimage.2016.06.038 PMID: 27346545

45. Patterson TNL. The optimum addition of points to quadrature formulae. Mathematics of Computation. 1968; 22: 847–856.

46. Massen C. Parallel programming of exogenous and endogenous components in the antisaccade task. Q J Exp Psychol A. 2004; 57: 475–498. https://doi.org/10.1080/02724980343000341 PMID: 15204137

47. Pierce JE, McDowell JE. Effects of preparation time and trial type probability on performance of anti- and pro-saccades. Acta Psychol (Amst). 2016; 164: 188–194.

48. Pierce JE, McDowell JE. Modulation of cognitive control levels via manipulation of saccade trial-type probability assessed with event-related BOLD fMRI. J Neurophysiol. 2016; 115: 763–772. https://doi.org/10.1152/jn.00776.2015 PMID: 26609113

49. Montague PR, Dolan RJ, Friston KJ, Dayan P. Computational psychiatry. Trends Cogn Sci (Regul Ed). 2012; 16: 72–80.

50. Wang XJ, Krystal JH. Computational psychiatry. Neuron. 2014; 84: 638–654. https://doi.org/10.1016/j.neuron.2014.10.018 PMID: 25442941

51. Stephan KE, Mathys C. Computational approaches to psychiatry. Curr Opin Neurobiol. 2014; 25: 85–92. https://doi.org/10.1016/j.conb.2013.12.007 PMID: 24709605

52. Paulus MP, Huys QJ, Maia TV. A Roadmap for the Development of Applied Computational Psychiatry. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging. Elsevier; 2016.

53. Huys QJ, Maia TV, Paulus MP. Computational Psychiatry: From Mechanistic Insights to the Development of New Treatments. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging. Elsevier; 2016; 1: 382–385.

54. Findlay JM, Walker R. A model of saccade generation based on parallel processing and competitive inhibition. Behav Brain Sci. 1999; 22: 661–674. PMID: 11301526

55. Feng G. Is there a common control mechanism for anti-saccades and reading eye movements? Evidence from distributional analyses. Vision Res. 2012; 57: 35–50. https://doi.org/10.1016/j.visres.2012.01.001 PMID: 22260785

56. Gold JI, Shadlen MN. The neural basis of decision making. Annu Rev Neurosci. 2007; 30: 535–574. https://doi.org/10.1146/annurev.neuro.29.051605.113038 PMID: 17600525

57. Ratcliff R, Smith PL, Brown SD, McKoon G. Diffusion Decision Model: Current Issues and History. Trends Cogn Sci (Regul Ed). 2016; 20: 260–281.

58. Donkin C, Brown S, Heathcote A, Wagenmakers EJ. Diffusion versus linear ballistic accumulation: different models but the same conclusions about psychological processes? Psychon Bull Rev. 2011; 18: 61–69. https://doi.org/10.3758/s13423-010-0022-4 PMID: 21327360

59. Moscoso del Prado Martin F. A theory of reaction time distributions. 2008.

60. Barndorff-Nielsen O, BlÆsild P, Halgreen C. First hitting time models for the generalized inverse Gaussian distribution. Stochastic Processes and their Applications. Elsevier; 1978; 7: 49–54.

61. Cutsuridis V. Behavioural and computational varieties of response inhibition in eye movements. Philos Trans R Soc Lond, B, Biol Sci. 2017; 372.

62. Logan GD, Van Zandt T, Verbruggen F, Wagenmakers EJ. On the ability to inhibit thought and action: general and special theories of an act of control. Psychol Rev. 2014; 121: 66–95. https://doi.org/10.1037/a0035230 PMID: 24490789

63. Cutsuridis V, Smyrnis N, Evdokimidis I, Perantonis S. A neural network model of decision making in an antisaccade task by the superior colliculus. Neural Networks. 2007; 20: 690–704. https://doi.org/10.1016/j.neunet.2007.01.004 PMID: 17446043

64. Cutsuridis V, Kumari V, Ettinger U. Antisaccade performance in schizophrenia: a neural model of decision making in the superior colliculus. Front Neurosci. 2014; 8: 13. https://doi.org/10.3389/fnins.2014.00013 PMID: 24574953

65. Cutsuridis V. Neural competition via lateral inhibition between decision processes and not a STOP signal accounts for the antisaccade performance in healthy and schizophrenia subjects. Front Neurosci. 2015; 9: 5. https://doi.org/10.3389/fnins.2015.00005 PMID: 25688183

66. Brown JW, Bullock D, Grossberg S. How laminar frontal cortex and basal ganglia circuits interact to control planned and reactive saccades. Neural Netw. 2004; 17: 471–510. https://doi.org/10.1016/j.neunet.2003.08.006 PMID: 15109680

67. Heinzle J, Hepp K, Martin KA. A microcircuit model of the frontal eye fields. J Neurosci. 2007; 27: 9341–9353. https://doi.org/10.1523/JNEUROSCI.0974-07.2007 PMID: 17728448

68. Wiecki TV, Frank MJ. A computational model of inhibitory control in frontal cortex and basal ganglia. Psychol Rev. 2013; 120: 329–355. https://doi.org/10.1037/a0031542 PMID: 23586447

69. Carpenter R, Noorani I. Movement suppression: brain mechanisms for stopping and stillness. Philos Trans R Soc Lond, B, Biol Sci. 2017; 372.

70. Boucher L, Palmeri TJ, Logan GD, Schall JD. Inhibitory control in mind and brain: an interactive race model of countermanding saccades. Psychol Rev. 2007; 114: 376–397. https://doi.org/10.1037/0033-295X.114.2.376 PMID: 17500631

71. Schmidt R, Berke JD. A Pause-then-Cancel model of stopping: evidence from basal ganglia neurophysiology. Philos Trans R Soc Lond, B, Biol Sci. 2017; 372.

72. Bissett PG. The countermanding task revisited: mimicry of race models. J Neurosci. 2013; 33: 12150–12151. https://doi.org/10.1523/JNEUROSCI.2091-13.2013 PMID: 23884923

73. Schall JD, Palmeri TJ, Logan GD. Models of inhibitory control. Philos Trans R Soc Lond, B, Biol Sci. 2017; 372.

74. Schmidt R, Leventhal DK, Mallet N, Chen F, Berke JD. Canceling actions involves a race between basal ganglia pathways. Nat Neurosci. 2013; 16: 1118–1124. https://doi.org/10.1038/nn.3456 PMID: 23852117

75. Stuphorn V, Schall JD. Executive control of countermanding saccades by the supplementary eye field. Nat Neurosci. 2006; 9: 925–931. https://doi.org/10.1038/nn1714 PMID: 16732274

76. Pouget P, Murthy A, Stuphorn V. Cortical control and performance monitoring of interrupting and redirecting movements. Philos Trans R Soc Lond, B, Biol Sci. 2017; 372.

77. Reuter B, Rakusan L, Kathmanna N. Poor antisaccade performance in schizophrenia: an inhibition deficit? Psychiatry Res. 2005; 135: 1–10. https://doi.org/10.1016/j.psychres.2004.12.006 PMID: 15893384

78. Reuter B, Jager M, Bottlender R, Kathmann N. Impaired action control in schizophrenia: the role of volitional saccade initiation. Neuropsychologia. 2007; 45: 1840–1848. https://doi.org/10.1016/j.neuropsychologia.2006.12.006 PMID: 17258779

79. Coe BC, Munoz DP. Mechanisms of saccade suppression revealed in the anti-saccade task. Philos Trans R Soc Lond, B, Biol Sci. 2017; 372.

80. Hakvoort Schwerdtfeger RM, Alahyane N, Brien DC, Coe BC, Stroman PW, Munoz DP. Preparatory neural networks are impaired in adults with attention-deficit/hyperactivity disorder during the antisaccade task. Neuroimage Clin. 2012; 2: 63–78. https://doi.org/10.1016/j.nicl.2012.10.006 PMID: 24179760

81. Cameron IG, Pari G, Alahyane N, Brien DC, Coe BC, Stroman PW, et al. Impaired executive function signals in motor brain regions in Parkinson's disease. Neuroimage. 2012; 60: 1156–1170. https://doi.org/10.1016/j.neuroimage.2012.01.057 PMID: 22270353

82. Witiuk K, Fernandez-Ruiz J, McKee R, Alahyane N, Coe BC, Melanson M, et al. Cognitive deterioration and functional compensation in ALS measured with fMRI using an inhibitory task. J Neurosci. 2014; 34: 14260–14271. https://doi.org/10.1523/JNEUROSCI.1111-14.2014 PMID: 25339740

83. Remijnse PL, van den Heuvel OA, Nielen MM, Vriend C, Hendriks GJ, Hoogendijk WJ, et al. Cognitive inflexibility in obsessive-compulsive disorder and major depression is associated with distinct neural correlates. PLoS ONE. 2013; 8: e59600. https://doi.org/10.1371/journal.pone.0059600 PMID: 23637737

Chapter 7

In Chapter 6, we introduced the SERIA model for the antisaccade task. Its most remarkable characteristic is the postulate that prosaccades can be either early responses, or voluntary saccades analogous to antisaccades. Through formal model comparison, we showed that SERIA accounted for experimental data better than models in which prosaccades were assumed to be always pre-potent, fast responses.

One important limitation of these findings is that the experimental design used in the previous chapter has been rarely investigated in humans, and only in a few physiological studies in the macaque monkey. The most conspicuous aspect of our design is that the peripheral cue that signals the correct (or incorrect) saccade direction, also signals the action to be performed (a pro- or antisaccade). By contrast, most antisaccade experiments are organized in blocks of the same trial type. Even in mixed designs, subjects are usually presented a central cue that signals the trial type before a peripheral stimulus is presented. Thereby, it is questionable whether the distinction between early and late errors introduced in the previous chapter is relevant when considering other designs of the antisaccade task. In other words*, is there any evidence of late decision processes and late errors when task demand and saccade direction cues are presented one after the other*?

In this chapter, we investigate whether voluntary prosaccades can be observed in the antisaccade task when task demands are presented in advance of the peripheral cue. To operationalize this hypothesis, we compare the SERIA model against a model that posits that all prosaccades are early responses (the PROSA model in Chapter 6) Moreover, we contrast this task with the same experimental procedure as used in the previous chapter.

Our results indicate that ERs and RTs were much lower when subjects were cued in advance about task demands compared to the design used in Chapter 6. In addition, the prosaccade RT distribution was not

bimodal, as reported in the previous chapter. Nevertheless, our results indicate that regardless of the experimental procedure, prosaccade distributions are better accounted for by SERIA. Model parameters also suggest that around a third of all errors in antisaccade trials can be categorized as late errors, with substantial variability across subjects. Furthermore, despite of the large absolute differences in ER and RT, we found that subjects' ERs were strongly correlated across the two conditions. This observation is important because it implies that the results obtained in Chapter 6 are indeed comparable to findings previously reported in the literature.

In the next chapter, we will extend SERIA to account for inter-trial effects, showing that alternating between pro- and antisaccade trials generates a degradation in performance that can be captured and partially explained by our generative model.

This chapter is publicly available as *Inhibition and late errors in the antisaccade task: Influence of task design; Eduardo A. Aponte, Dominic G. Tschan, Klaas E. Stephan, Jakob Heinzle bioRxiv 270165; doi: https://doi.org/10.1101/270165*.

Inhibition failures and late errors in the antisaccade task: Influence of cue delay

Eduardo A. Aponte[1,*], *Dominic G. Tschan*[1], *Klaas E. Stephan*[1,2], *Jakob Heinzle*[1]

[1] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich. Wilfriedstrasse 6, 8004, Zurich, Switzerland.

[2] Wellcome Centre for Neuroimaging, University College London. 12 Queen Square, London WC1N 3BG, UK.

[*]Corresponding author: Eduardo A. Aponte; aponte@biomed.ee.ethz.ch

Abstract

In the antisaccade task participants are required to saccade in the opposite direction of a peripheral visual cue (PVC). This paradigm is often used to investigate inhibition of reflexive responses as well as voluntary response generation. However, it is not clear to what extent different versions of this task probe the same underlying processes. Here, we explored with the *Stochastic Early Reaction, Inhibition, and late Action* (SERIA) model how the delay between task cue and PVC affects reaction time (RT) and error rate (ER) when pro- and antisaccade trials are randomly interleaved. Specifically, we contrasted a condition in which the task cue is presented prior to the PVC with a condition in which the PVC serves also as task cue. Summary statistics indicate that ER and RTs are reduced, and contextual effects largely removed when the task is signaled before the PVC appears. The SERIA model accounts for RT and ER in both conditions, and better so than other candidate models. Modeling demonstrates that voluntary pro- and antisaccades are frequent in both conditions, showing that not all prosaccades are reflex-like actions. Moreover, early task cue presentation results in better control of reflexive saccades leading to fewer fast antisaccade errors and more rapid correct prosaccades. Finally, wrong voluntary actions, i.e. late errors, are prevalent in both conditions. In summary, SERIA provides an explanation for the differences in the delayed and non-delayed antisaccade task.

Keywords: antisaccades, eye movements, SERIA model, reaction time, error rate

Introduction

The antisaccade task (Hallett 1978) is an oculomotor paradigm widely used in psychiatric research (Hutton and Ettinger, 2006; Gooding and Basso 2008; Bittencourt et al. 2013), in which participants are required to saccade in the opposite direction of a peripheral visual cue (PVC). This paradigm probes both the ability to inhibit reflexive responses – i.e., (pro)saccades towards a visual cue – and the ability to initiate voluntary actions – i.e., (anti)saccades in the direction contralateral to the peripheral stimulus (Everling and Fischer 1998). Since the seminal study of Hallett (1978), it is known that participants commit errors in the antisaccade task, defined as saccades towards the PVC. The clinical relevance of this paradigm originates from the fact that both error rates (ER) and reaction times (RT) are altered in many psychiatric and neurological diseases. For example, ERs are elevated not only in schizophrenic patients (Gooding and Basso 2008), but also in their first order relatives as well as in related psychiatric populations, such as schizoaffective disorder patients (Calkins et al. 2004; Reilly et al. 2014; Myles et al. 2017).

Errors in this task have often been attributed to deficits in inhibitory control (e.g. Levy et al. 1998; Broerse et al. 2001; Calkins et al. 2004), while other proposals view them as deficits in voluntary action initiation (Reuter and Kathmann 2004; Reuter et al. 2005). Initially, Fischer et al. (2000) proposed to differentiate between inhibition errors and volitional errors based on a factor analysis on the ER of a large cohort of subjects, sub-classed by the number of 'expressed saccades' performed in prosaccade trials, a proxy of inhibitory control. This analysis revealed two main factors that predicted ER, which Fischer and colleagues interpreted as inhibitory control and voluntary action initiation. Using a similar argument, Klein and Fischer (2005) proposed to extend the distinction between express and 'normal-range' saccades to antisaccade errors, and used indirect statistical evidence to suggest that these evolve differently during development and are correlated with different psychometric constructs (Klein et al. 2010). Reuter and colleagues (Reuter and Kathmann 2004; Reuter et al. 2005), based on the parallel programming model proposed by Massen (2004), hypothesized that at least some fraction of the errors observed in this paradigm are caused by failures to initiate a voluntary action. More recently, Lo and Wang

(2016) incorporated the idea of two sources of antisaccade errors into a biophysical model of eye movement control and speculated that the mechanisms behind prosaccade errors with unusual high latency might be of interest in psychiatric research. In that spirit, Coe and Munoz (2017) suggested that the ratio between early and late errors could distinguish between control and patient populations, such as Parkinson's disease and lateral amyotrophic sclerosis patients.

Recently (Aponte et al. 2017), using the *Stochastic Early Reaction, Inhibition, and late Action* (SERIA) model, we presented quantitative and qualitative evidence that errors in the antisaccade task can be divided into fast, reflex-like prosaccades and voluntary but erroneous late prosaccades. SERIA is a generative model that extends the LATER model for antisaccades initially proposed by Noorani and Carpenter (2013) and builds on the idea that RTs are distributed as the threshold hit times of linear, ballistic accumulation processes (Noorani and Carpenter 2016). In this family of models, pro- and antisaccades are generated by two competing but independent accumulators. In addition, a third unobservable process can stop reflexive prosaccades, similarly as in the model used for the countermanding saccade task (Logan et al. 1984).

Conceptually, SERIA extends Noorani and Carpenter's work by introducing a further decision process that can generate late prosaccades and competes with the (late) antisaccade process. Errors can therefore be divided into early errors, explained as inhibition failures, and late errors, explained as the result of a late race between voluntary pro- and antisaccades. Moreover, according to SERIA, late errors on prosaccade trials can occur when an early response is inhibited, but an antisaccade overwrites a late prosaccade. Thus, our model provides a unified account of all types of errors observed in the antisaccade task.

One limitation of the study reported in Aponte et al. (2017) is that the version of the antisaccade task used there originated from non-human primate studies (e.g., Sato and Schall 2003) but has not been extensively investigated in humans (Weber 1995; Irving et al. 2009; Liu et al. 2010; Chiau et al. 2011; Weiler and Heath 2012). Concretely, in Aponte et al. (2017) subjects performed interleaved pro- and antisaccade trials, in which a PVC signaled both the trial type and the target location (see Fig 1A). We refer to this version of the antisaccade task as a *synchronous cue* (SC) design.

In humans, the antisaccade task is most often administered in a block design (Antoniades et al. 2013) in which subjects perform a single trial type throughout a block and are informed in advance about the task to be performed. Even when different trial types are interleaved, participants are usually informed about the task demands before the PVC is presented (e.g., Cherkasova et al. 2002; Massen 2004; O'Driscoll et al. 2005; Reuter et al. 2006; Pierce et al. 2015; Pierce and McDowell 2016a; 2016b). We refer to this paradigm as the *asynchronous cue* (AC) design. This version of the task has been used in monkey experiments as well (e.g., Amador et al. 1998; Johnston et al. 2014; Koval et al. 2014; Vijayraghavan et al. 2016).

The main goal of the present study was to test whether the conclusions drawn in our previous experiment generalize to the AC, the most commonly used version of the antisaccade task in humans. We acquired data from twenty-four participants in both the SC and AC conditions and compared RT and ER as well as SERIA model parameters estimated from the data. We were interested in three main questions: First, we investigated whether in an AC design it was necessary to postulate a late race between voluntary pro- and antisaccades. Hence, we compared models that incorporated a late race against models in which all late saccades were antisaccades. Second, we were interested in differences in the probability of inhibition failures (i.e. early errors in antisaccades) and late errors in the two task designs. In particular, we investigated if and in what proportions late errors occurred in SC and AC tasks. Finally, we tested whether the effects of trial type probability reported in Aponte et al. (2017) could be replicated, and whether these effects generalized to the AC design.

Methods

Participants

Twenty-five healthy male volunteers (age: 21.4±2.0 y) participated in the study approved by the local ethics board of the Canton of Zurich, Switzerland (KEK-ZH-Nr.2014-0246) and conducted according to the Declaration of Helsinki. Because this experiment was part of a larger pharmacological study, only male participants were included. All subjects had normal or corrected to normal vision and gave their written informed consent to participate. One subject had to be excluded because of incomplete data. Hence, twenty-four subjects were included in the final analysis.

Apparatus

The experiment took place in a dimly illuminated room. Subjects viewed a CRT screen (41.4x30cm; Philips 20B40) operating at 85Hz from a distance of 60cm, while their gaze was recorded with an infrared eye tracker (Eyelink 1000, SR Research, Ottawa, Canada). Head position was stabilized using a chin rest. Gaze position was recorded at a sampling rate of 1000Hz. Every block started with a 5-points calibration procedure. Absolute calibration error was aimed to be below 1°. The experiment was programmed in the Python programming language (2.7) using the *PsychoPy* (1.82.02) package (Peirce 2007, 2008). The experiment was controlled by a personal computer (Intel Core i7 4740K) equipped with a Nvidia GTX760 graphics card.

Experimental design

The experimental design used here is an extension of the design used in Aponte et al. (2017). Subjects participated in 6 blocks of mixed pro- and antisaccade trials. Each block consisted of 200 randomly interleaved pro- and antisaccade trials, from which either 20, 50 or 80% were prosaccade trials. In addition to trial type probability, we also manipulated the temporal order in which the trial type cue and the saccade direction cue were presented: Subjects were either simultaneously informed about the trial type and saccade direction using one peripheral cue (SC condition), or they were informed about the trial type before being presented with the peripheral cue (AC condition). Both conditions are explained in detail below. All task instructions were given to the participants in written format prior to the experiment.

The experiment followed a within-subject, 3x2 factorial design, with factors *prosaccade trial probability* (PP) with levels PP20, PP50, and PP80 and *cue type* (CUE) with levels SC and AC. The blocks belonging to one of the CUE conditions were administered consecutively. The order of presentation of the blocks was pseudo-randomized and counterbalanced across subjects. The same sequence of pro- and antisaccade trials was used for each PP condition independently of the CUE condition. The peripheral cue was presented randomly on the right and left side of the screen. Again, the same random sequence was used across subjects.

Before participating in the main experiment, subjects underwent a training block for each condition. These consisted of 100 trials, from which the first half were prosaccade trials, followed by 50 antisaccade trials. During training, participants received automatic feedback after each trial indicating whether they had made a saccade in the correct direction. In order to urge participants to respond quickly, saccades with a latency above 500ms were signaled as errors.

Synchronous cue (SC) condition

Throughout the experiment, two red circles of 0.25° of radius were presented at +12° to the left and right of the center of the screen. Each trial started with a cross (0.6x0.6°) displayed at the center of the screen. Subjects were required to fixate for at least 500ms. If their gaze drifted outside a 3° window, the fixation interval was restarted. The fixation target was presented for a further random interval (500-1000ms), after which a green bar (3.48x0.8°) centered on one of the peripheral red circles was displayed for 500ms (Fig. 1A). The bar was presented in either horizontal or vertical orientation. A horizontal bar indicated a saccade to the cued stimulus, and a vertical bar indicated a saccade to the uncued stimulus. The next trial started 1000ms after the peripheral cue was removed.

Asynchronous cues (AC) condition

The start of the AC condition (Fig. 1B) was identical to the SC condition, but after the initial fixation period a green bar (3.48x0.8°) was displayed for 700ms, centered on the fixation cross. The bar could be in horizontal or vertical orientation, cueing a pro- or antisaccade trial, respectively. The fixation cross and the green bar were removed at the end of the 700ms period and subsequently a green square (1.74x1.74°) was presented on one of the peripheral red circles for 500ms. Subjects were

instructed to saccade to the circle ipsilateral to the green square if the task cue was a horizontal bar, and to saccade to the contralateral circle if it was a vertical bar. The next trial started 1000ms after the green square was removed.

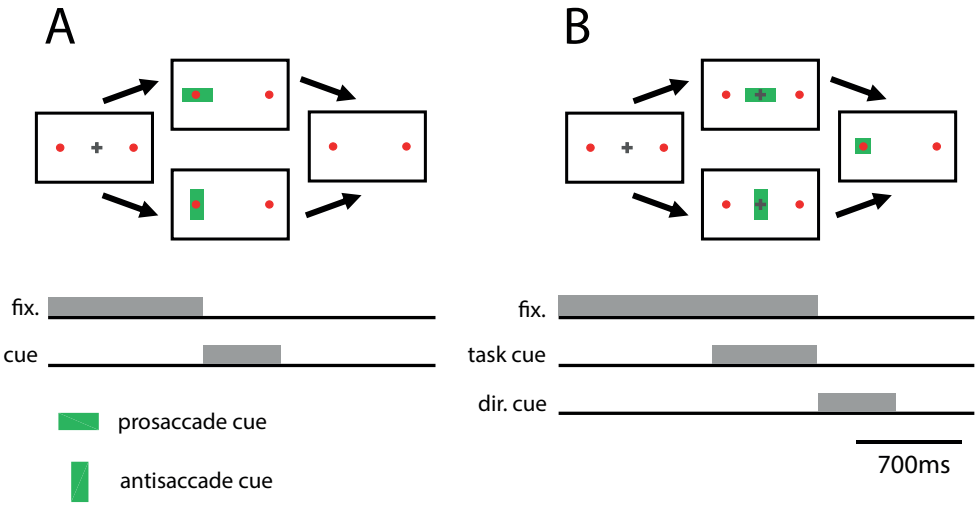


Fig. 1: Task design. A. *Synchronous cue (SC) condition.* Similarly to Aponte et al., 2017, subjects were instructed to fixate a central cross for 500-1000ms, while two red circles (0.25° radius) were displayed at ±12°. Immediately after the fixation period, a green bar (3.4x0.8°) was displayed centered on one of the red circles for 500ms. Participants were instructed to saccade as fast as possible to the red circle ipsilateral to a horizontal green bar, and to saccade to the contralateral circle when a vertical bar was displayed. **B.** *Asynchronous cues (AC) condition.* As in the SC condition, subjects were instructed to fixate a central cross for 500 to 1000ms. After the initial fixation period, a green bar (3.4x0.8°) was displayed at the center of the screen for 700ms. Immediately afterwards, the fixation cross and the green bar were removed and a green square (1.74x1.74°) was displayed centered on one of the circles. Subjects were instructed to saccade to the circle ipsilateral to the peripheral cue if a horizontal bar was presented, and to saccade to the contralateral circle otherwise.

Data preprocessing

Data was preprocessed using the Python programming language (2.7). Saccades were detected using the algorithm provided by the eye tracker manufacturer (Stampe 1993), which uses velocity and acceleration thresholds of 22°/s and 3800°/s^2, respectively. Saccades with a magnitude lower than 2° were ignored. RT was defined as the latency of the first saccade after the fixation cross was removed (henceforth, the *main saccade*). Trials were discarded if any of the following conditions

was true: if a blink occurred between the start of the fixation period and the end of the main saccade; if subjects failed to maintain fixation; if a saccade had a latency above 800ms or below 50ms, and in the case of an antisaccade, a latency below 95ms. Corrective antisaccades were defined as saccades contralateral to the peripheral cue that followed an error in an antisaccade trial. Corrective saccades were only included in the analysis if they occurred within 900ms after cue presentation and if their end location was within a 4° window around the correct target.

Classical statistical analysis

Mean RTs, ERs and parameter estimates of the model (see below) were analyzed using a generalized mixed effects linear model. The independent variables were PP with levels PP20, PP50, PP80; CUE with levels SC and AC; SUBJECT entered as a random effect, and, when pro- and antisaccade trials were analyzed together, *trial type (TT)*. All regressors were entered as categorical variables. ERs were analyzed using a binomial regression model with the probit function as link function. When probabilities were analyzed, a fixed effects Beta regression model (Cribari-Neto and Zeileis 2009) was used, because a mixed effect model proved numerically unstable. For RT, we report tests based on the *F* statistic, whereas for ER and probabilities we report tests based on the X^2 statistic, as this is more appropriate in model were the dispersion parameter is not estimated from the data. Statistical significance was asserted at $\alpha=0.05$. All statistical tests were performed with the *R* programming language (3.4.2) using the functions *lmer, glmer,* and *glmmadmb* (Beta regression model) from the packages *lme4, lmerTest*, and *glmmADMB*.

Modeling

Two models (described in detail in Aponte et al. 2017) were fitted to actions (pro- or antisaccades) and RTs. First, we fitted the *PRO-, Stop and Antisaccade* (PROSA) model, which structurally resembles the model described in Noorani and Carpenter (2013). According to this model, three linear race decision units determine RTs and ERs in the antisaccade task. Each unit triggers or stops different types of action depending on the order and time at which they hit threshold (henceforth *hit time*): The first *early unit* triggers a prosaccade if it hits threshold before all other units. These fast reactions can be stopped by the inhibitory unit, if the latter hits threshold before the *early unit*. If an early response is

inhibited, the third unit triggers an *antisaccade* once it hits threshold. Note that we assume that the antisaccade unit is delayed with respect to the early and inhibitory unit, a phenomenon denominated the antisaccade cost (Hallet 1978). This model represents the hypothesis that all voluntary or late responses are antisaccades.

More formally, we assume three independent stochastic accumulation processes or units that represent early responses (u_e), a unit that inhibits them (u_i), and a unit that trigger antisaccades (u_a). The threshold hit time of the units can be represented by the random variables U_e, U_i and U_a, respectively. According to PROSA, a prosaccade is generated at time t if the early unit hits threshold at time t before all other units

$$p(A = pro, T = t) = p(U_e = t)p(U_i > t)p(U_a > t). \quad (1)$$

Here the probability on the left hand side of the equation is the probability that the action prosaccade ($A = pro$) is generated at time ($T = t$). An antisaccade at time t is elicited when the antisaccade unit hits threshold at time t before all other units,

$$p(U_a = t)\, p(U_e > t)p(U_i > t) \quad (2)$$

or the inhibitory unit hit threshold before the early unit.

$$p(U_a = t) \int_0^t p(U_i = \tau)p(U_e > \tau)d\tau\,. \quad (3)$$

From this it follows that

$$\begin{aligned} p(A = anti, T = t) &= p(U_a = t)\, p(U_e > t)p(U_i > t) \\ &+ p(U_a = t) \int_0^t p(U_i = \tau)p(U_e > \tau)d\tau\,. \end{aligned} \quad (4)$$

Note that according to PROSA, all early reactions are prosaccades, which can be stopped by the inhibitory unit u_i.

Second, we fitted the SERIA model see Fig. 2, which extends PROSA by including a fourth unit that can trigger late, voluntary prosaccades. Hence, SERIA distinguishes between reflexive, early prosaccades, and voluntary late prosaccades.

Formally, to account for late prosaccade we model a fourth unit u_p and its threshold hit time U_p. A prosaccade at time t can be generated when the early unit hits threshold before all other units

$$p(U_e = t)p(U_a > t)p(U_i > t)p\big(U_p > t\big) \quad (5)$$

or the late prosaccade unit hits threshold before all other units

$$p\big(U_p = t\big)p(U_a > t)p(U_i > t)p(U_e > t) \tag{6}$$

or the inhibitory unit stops early reaction and the late prosaccade unit hits threshold before the antisaccade unit

$$p\big(U_p = t\big)p(U_a > t)\int_0^t p(U_i = \tau)p(U_e > \tau)d\tau\,. \tag{7}$$

Finally, antisaccades are generated either when the antisaccade unit hits threshold before all other units

$$p(U_a = t)p\big(U_p > t\big)p(U_e > t)p(U_i > t) \tag{8}$$

or the early prosaccade unit is stopped, and the late prosaccade unit hits threshold after the antisaccade unit

$$p(U_a = t)p\big(U_p > t\big)\int_0^t p(U_i = \tau)p(U_e > \tau)d\tau\,. \tag{9}$$

As for the PROSA model, the probability of a specific action at time *t* can be calculated by summing the probabilities of the different cases that can trigger the corresponding action.

SERIA distinguishes two types of errors in antisaccade trials: inhibition failures, when the early unit hits threshold before all other units, and volitional or late errors when the late prosaccade unit hits threshold before the antisaccade unit. An error on a prosaccade trial occurs when an early response is stopped, but the antisaccade unit hits threshold before the late prosaccade unit. Note that the model used here corresponds to the SERIA model with late race (SERIA_{lr}) introduced in Aponte et al. (2017).

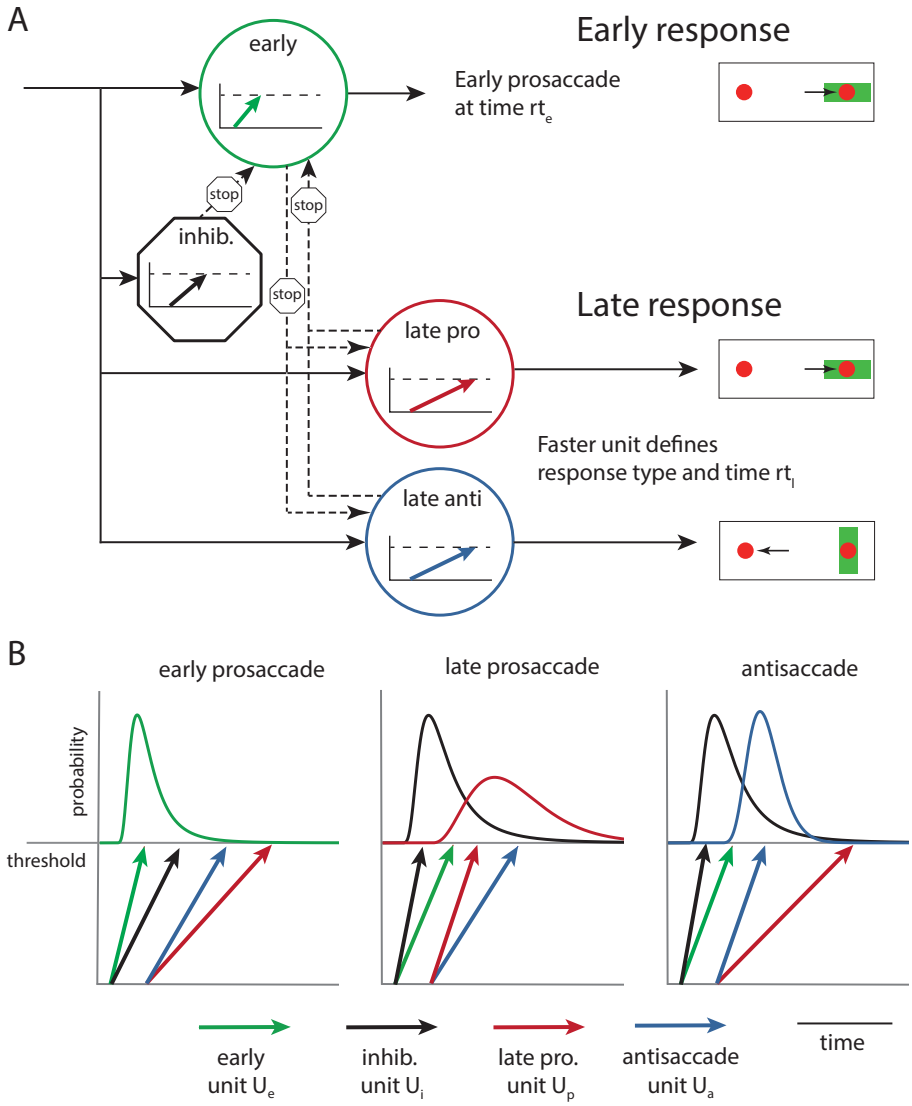


Fig. 2 The SERIA model: A) The SERIA model consists of four units with different arrival time distributions. A reactive, early response is triggered if the early unit (green) hits threshold before all other units. If the early unit is inhibited by the inhibitory unit (black), a late decision process is decided between the late pro (red) and late anti (blue) units. The unit arriving at threshold first, defines the action and reaction time. Figure adapted with permission from Aponte et al. (2017). **B)** The order and the hit times of the units determine the RT and action performed in a trial. The increase rate of each of the units is assumed to be stochastic. Colors correspond to subfigure A). For simplicity, units are shown sharing the same threshold, although this assumption is not necessary. Note that in the PROSA model, there is no late prosaccade unit and thereby prosaccades can only be generated by the early unit. **Left:** An early prosaccade is generated when the early unit hits threshold before all other units. **Middle:** A late prosaccade is generated when the inhibitory unit hits threshold before all other units, and the late prosaccade unit hit threshold before the late antisaccade unit. **Right:** An antisaccade is generated when early reactions are inhibited and the antisaccade unit hit threshold before the late prosaccade unit.

To fit the models to empirical data, we evaluated three different parametric distributions for the increase rate (or reciprocal hit time) of each of the units: We either assumed that the increase rate of all the units were truncated Gaussian distributed in analogy to the LATER model (Noorani and Carpenter 2016), or that the increase rates of the early and inhibitory unit were Gamma distributed, but the increase rate of the late units was inverse Gamma distributed. We refer to this model as the mixed Gamma model. Finally, we considered a model in which the increase rate of all the units was Gamma distributed.

We assumed that a different set of parameter values for each of the units was necessary for each trial type. However, we also considered a constrained version of the SERIA model in which the early and inhibitory units followed the same distribution on pro- and antisaccade trials, but where the late units had different parameter values across trial types (Aponte et al. 2017). For PROSA, we investigated a model in which the early unit followed the same distribution across trial types but others were allowed to differ (Noorani and Carpenter 2013; Aponte et al. 2017). A summary of the model space is presented in Table 1. More details on the model space can be found in Aponte et al. (2017).

Table 1: Model space

Model	Parametric dist.	No. parameters
Unconstrained/Constrained	PROSA	
m_1/m_2	Truncated Normal	15/13
m_3/m_4	Mixed Gamma	15/13
m_5/m_6	Gamma	15/13
	SERIA	
m_7/m_8	Truncated Normal	19/15
m_9/m_{10}	Mixed Gamma	19/15
m_{11}/m_{12}	Gamma	19/15

List of models with corresponding increase rate distributions and number of free parameters. In constrained models, some of the parameters are assumed to be equal across trial types. Note that besides the parameters of the units, all models include three nuisance parameters that account for no-response time, late response cost, and the frequency of outliers, i.e., saccades with latencies below the no-response time. Further details can be found in Aponte et al. (2017).

We fitted the data of all subjects and PP conditions simultaneously using a Bayesian hierarchical model (Gelman et al. 2003), in which the prior distribution of the parameters for each subject was informed by the population distribution. The two CUE conditions were analyzed independently, because our goal was to evaluate whether different models were favored under different task designs. The population distribution was modeled using a linear mixed effects model with PP as fixed effect and SUBJECT as a random effect. Details are provided in the Supplementary Methods and in Supplementary Figure S1.

Models were fitted using Markov chain Monte Carlo (MCMC) sampling via the Metropolis-Hastings algorithm. Model evidence was computed with thermodynamic integration (Gelman and Meng 1998; Aponte et al. 2016), with 32 chains and a 5^{th} order temperature schedule (Calderhead and Girolami 2009). To increase the efficiency of the algorithm, we incorporated a 'swap-step' according to population MCMC's accept/reject rule (Calderhead and Girolami 2009). The algorithm was run for 16×10^4 iterations, and the first 6×10^4 samples were discarded

as 'burn-in' samples. The code was executed on a computer cluster running Linux (CentOS 7.4.1708), MATLAB R2015a (8.5.0.197613), and GSL 1.16. The software implemented here is publicly available as part of the TAPAS toolbox (http://translationalneuromodeling.org/tapas/; see software note).

The statistic used to compare models was the difference in log model evidence (LME), which correspond to log Bayes factors (Kass and Raftery 1995). Because our main hypothesis was related to families of models (SERIA and PROSA), we used Bayesian family model comparison (Penny et al. 2010) implemented in the SPM12 software package (release 6470, function *spm_compare_families.m*). Building on random effects Bayesian model selection (Stephan et al. 2009), this method pools the evidence of models which are assumed to belong to the same family and returns the posterior probability of each family.

Results

A total of 28815 main saccades were collected from 24 subjects. 1079 saccades (or 3.7%) were discarded due mainly to eye blinks (330), fixation failures (458), and with a latency below 50ms (162). Only a minority of saccades (14) had a latency above 800ms. For the analysis of corrective saccades, 983 and 696 trials were included in the SC and AC conditions, respectively.

Error rate (ER)

Fig. 3A and 3B display the mean ER in all conditions and trial types. Pro- and antisaccade ERs were submitted to two independent tests using PP and CUE as explanatory variables. ERs were higher in the SC condition, regardless of trial type (prosaccade trials: $\mathrm{X}^2(2, N = 144) = 402.75, p < 10^{-5}$, antisaccade trials: $\mathrm{X}^2(2, N = 144) = 257.06, p < 10^{-5}$). Moreover, there was a significant interaction between the factors PP and CUE in both trial types, demonstrating that PP had a much more pronounced effect in the SC condition (prosaccade trials: $\mathrm{X}^2(2, N = 144) = 43.00, p < 10^{-5}$; antisaccade trials: $\mathrm{X}^2(2, N = 144) = 63.43, p < 10^{-5}$). Next, we submitted ERs in the two CUE conditions to two separate tests with explanatory variables TT and PP. Thus, we could test whether PP had a significantly different effect on pro- and antisaccade trials. We found that in the two CUE conditions, the interaction between PP and TT was significant (SC: $\mathrm{X}^2(2, N = 144) = 700.46, p < 10^{-5}$, AC: $\mathrm{X}^2(2, N = 144) = 41.24, p < 10^{-5}$).

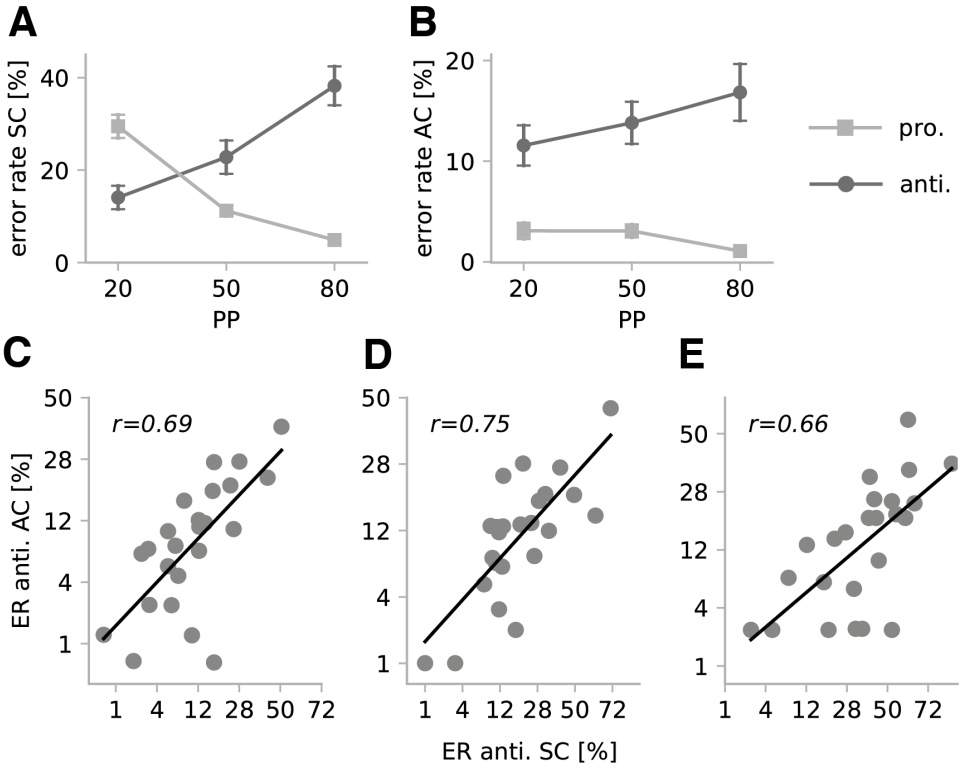


Fig. 3: A. Mean ER vs. prosaccade probability (PP), SC condition. **B.** Mean ER vs. prosaccade probability (PP), AC condition. Error bars depict the standard error of the mean (sem). **C-E.** ER correlation between the AC and SC conditions in the PP20, PP50 and PP80 conditions, respectively. ER are displayed in the probit scale.

Next, we investigated the correlation of ER from the two CUE conditions (Fig. 3C-E). The probit transformed ERs in each PP block were analyzed separately. For numerical reasons, zero percent ERs were set to a non-zero value, pretending that the respective subjects had committed a single error. There was a significant correlation ($R^2 > 0.43, p < 0.001$) between ERs on antisaccade trials for all three PPs, but we found no comparable results on prosaccade trials ($R^2 < 0.01, p > 0.56$).

Reaction times (RT)

Mean RTs of correct saccades are displayed in Fig. 4. First, RT on pro- and antisaccade trials were submitted to two separate models with PP and CUE as independent variables. Clearly, RTs in the SC condition were much higher than in the AC condition (prosaccades: $F_{1,115} = 815.05, p < 10^{-5}$; antisaccades: $F_{1,115} = 789.90, p < 10^{-5}$). The factor PP was significant in both pro- ($F_{2,115} = 3.46, p = 0.03$) and antisaccade trials ($F_{2,115} = 4.32, p = 0.01$). However, there was a significant interaction

between the factors CUE and PP on antisaccade ($F_{2,115} = 11.25, p < 10^{-3}$), but not on prosaccade trials ($F_{2,115} = 1.79, p = 0.17$).

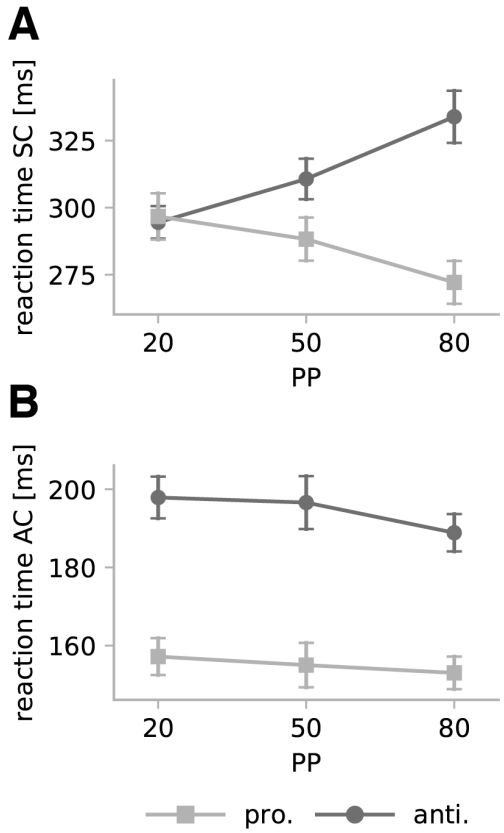


Figure 4: A. Mean RT vs. prosaccade probability (PP), SC condition. **B.** Mean RT vs. prosaccade probability (PP), AC condition. Only the mean RTs of correct trials are displayed. Error bars depict the sem.

We then investigated both CUE conditions separately in a model with factors PP and TT. In the AC condition, pro- and antisaccade RTs decreased with PP, as previously reported by Pierce et al. (2015). However, neither the main effect of PP ($F_{1,115} = 2.40, p = 0.09$) nor the interaction PP*TT was significant ($F_{2,115} = 0.48, p = 0.61$), although the main effect of TT was significant ($F_{1,115} = 238.93, p < 10^{-5}$). In the SC condition, PP had the opposite effect on pro- and antisaccades which resulted in a significant interaction between PP and TT ($F_{2,115} = 12.99, p = 10^{-5}$).

Model comparison

In order to compare models, we used the differences in LME or log Bayes factors between the hierarchical models fitted to our data (Table 2). The expected log likelihood or accuracy of each model is reported in

Supplementary Table S1. This measure is closely related to the R^2 statistic and reflects the un-penalized goodness of fit of a model.

Table 2: Differences in log model evidence (LME)

Model	Parametric family	SC	AC
	PROSA		
m_1	T. normal	103.0	295.0
m_2	T. normal	0.0	0.0
m_3	Mixed Gamma	540.7	291.7
m_4	Mixed Gamma	518.3	92.3
m_5	Gamma	572.5	364.1
m_6	Gamma	557.4	140.2
	SERIA		
m_7	T. normal	1162.7	740.1
m_8	T. normal	1177.8	717.7
m_9	Mixed Gamma	1230.4	**874.8**
m_{10}	Mixed Gamma	1264.9	542.6
m_{11}	Gamma	1248.0	795.3
m_{12}	Gamma	**1291.5**	769.9

Model comparison. Log evidences are given relative to the worst model (zero) in each condition. The models with the highest evidence are highlighted in bold font.

We first compared families of models in each of the conditions separately. In the SC condition, the SERIA family was favored when compared to the PROSA family (posterior probability nearly 1). In the SERIA family, constrained models were favored when compared to models in which the early and inhibitory unit were allowed to differ across trial types (posterior probability nearly 1). When we considered each model independently (Table 2), analogously to the findings in

Aponte et al. (2017), a constrained SERIA model (m_{12}) obtained the highest evidence ($\Delta LME > 26.6$).

In the AC condition, while the SERIA model family was favored when compared to the PROSA model family (posterior probability approx. 1), SERIA models in which the early and stop units were not constrained obtained the highest evidence (posterior probability approx. 1). When comparing models individually, the unconstrained mixed Gamma SERIA model (m_9) was favored among all possibilities ($\Delta LME > 79.5$).

In order to facilitate the comparison across CUE conditions and our previous study (Aponte et al. 2017), in the following we report the parameter estimates obtained using mixed Gamma models (SC condition: m_{10}; AC condition: m_9).

Model fits

To qualitatively evaluate the PROSA and SERIA models (Gelman et al. 2003; Gelman and Shalizi 2013), we plotted the histogram of RTs of all saccades and the fit of the best model in each family (Fig. 5). For the PROSA model, we used model m_5 in both conditions. Fits were computed by weighting the expected probability density function in a given block by the corresponding number of trials. Fits for representative subjects are displayed in Supplementary Figures S2 and S3.

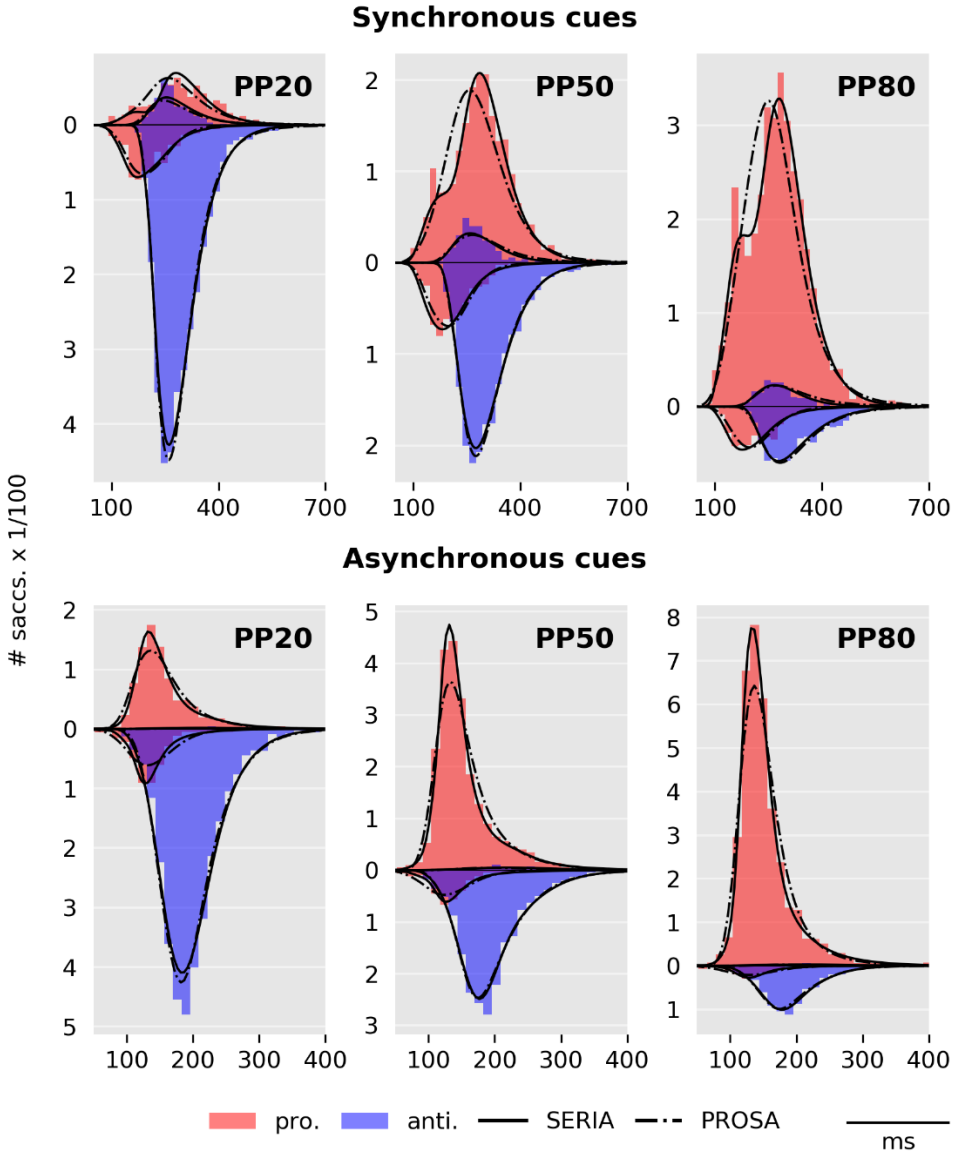


Fig. 5: Histogram of RTs and model fits. For comparison, prosaccade trials are displayed in the positive half plane while antisaccades trials are displayed in the negative half plane. The histogram of prosaccade responses is displayed as red bars, whereas antisaccades are displayed in blue. Hence, errors in prosaccade trials (antisaccades) are displayed in blue in the positive half plane, whereas errors in antisaccade trials (prosaccades) are displayed in red in the negative half plane.

Replicating our previous findings (Aponte et al. 2017) the RT distribution of correct prosaccades in the SC condition was bimodal, and could not be captured by the PROSA model, but was accounted for by the SERIA model. More importantly, since this is the first time that

SERIA is applied to data from an AC task design, the RT distributions in the AC condition were also fitted better by the SERIA model. This was particularly visible for correct prosaccades in the PP50 and PP80 condition (Fig. 5, bottom row, middle and right panels).

To further investigate the fits of the SERIA model, Fig. 6 displays the empirical and predicted cumulative density function (cdf) of the reciprocal RT[1] of correct pro- and antisaccades. Cdfs are displayed on the probit scale (Noorani and Carpenter 2016) but in contrast to previous accounts (Aponte et al. 2017; Noorani and Carpenter 2013), we did not normalize by the total number of saccades.

[1] Reciprocal RT are often used to compare cumulative RT distributions. In these plots the x-axis is rescaled proportionally to 1/RT and flipped in order to have RT increasing from left to right. A detailed description of reciprobit plots can be found elsewhere (Noorani and Carpenter 2016).

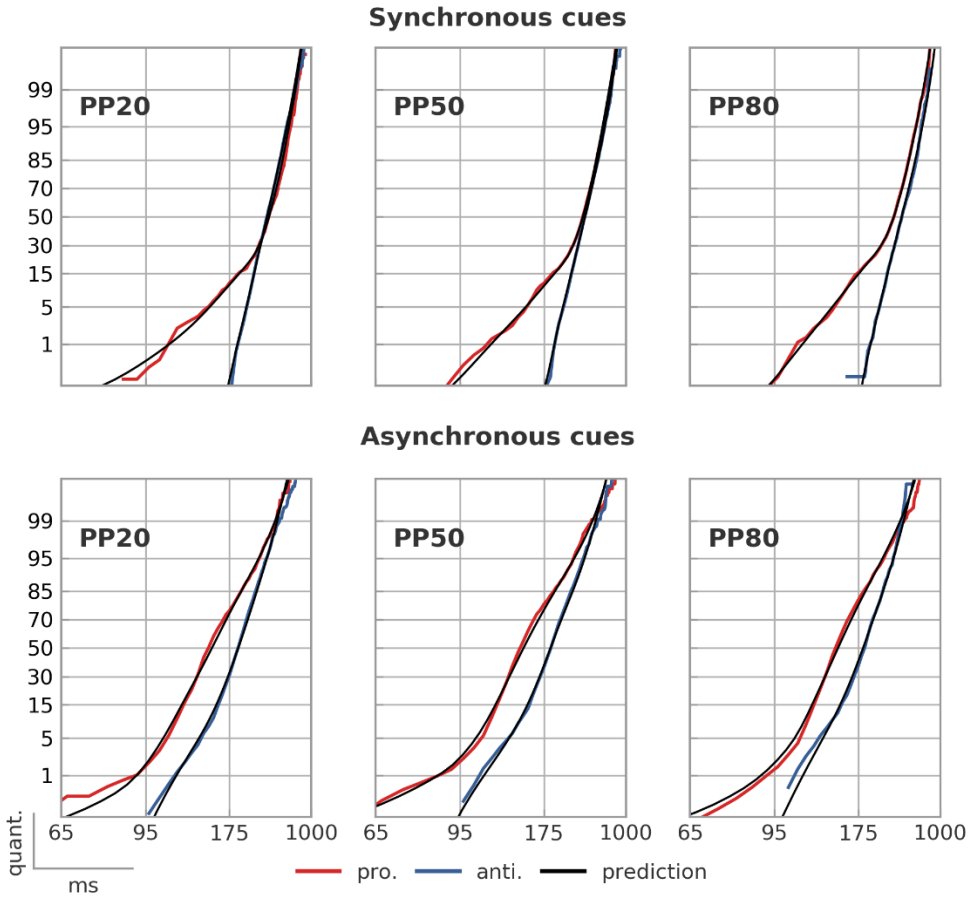


Fig. 6: Empirical and predicted reciprobit of RTs in correct trials. In the SC condition, the SERIA model clearly captured the apparent bimodality of the RT distributions. Please note the deflection in the prosaccade cdf, which demonstrates a bimodal distribution. In the AC condition, the SERIA model accounted for most of the relevant aspects of the RT distribution, including left and right tails.

The distribution of reciprocal (inverse) RTs on correct trials in the SC condition echoed the findings by Carpenter and Williams (1995), and suggest that prosaccades are the results of two processes (Noorani and Carpenter 2016). Moreover, the RT distribution of late prosaccades converges to the distribution of correct antisaccades. This provides further evidence for the hypothesis that late prosaccades are the result of a slow accumulation process analogous to the one used to model antisaccades (Aponte et al., 2017).

Importantly, SERIA also yielded accurate fits in the AC condition. Although the RT distribution of pro- and antisaccades deviated from the linear behavior observed in the SC condition, the model correctly

predicted the empirical cdfs. Arguably, because late responses have latencies as low as 95ms, early and late prosaccades are disguised in a single unimodal distribution that does not follow the linear pattern observed in the SC condition. The PROSA model yielded less accurate fits (cf. Supplementary Figure S4).

RT distribution of corrective antisaccades

In order to predict the RT distribution of corrective antisaccades, the distribution of the hit time of the late antisaccade unit of each subject in each condition was weighted by the corresponding number of corrective antisaccades. See Aponte et al. (2017) for a similar analysis. The estimated distribution was time-shifted to optimize the predictive fit, i.e., we tried to predict the shape of the RT distribution, *not* its mean. Fig. 7 displays the predicted distributions in the SC (*time-shift=93ms*) and AC (*time-shift=63ms*) conditions. Visual inspection suggests that SERIA predicted correctly the shape of the distribution of corrective antisaccades.

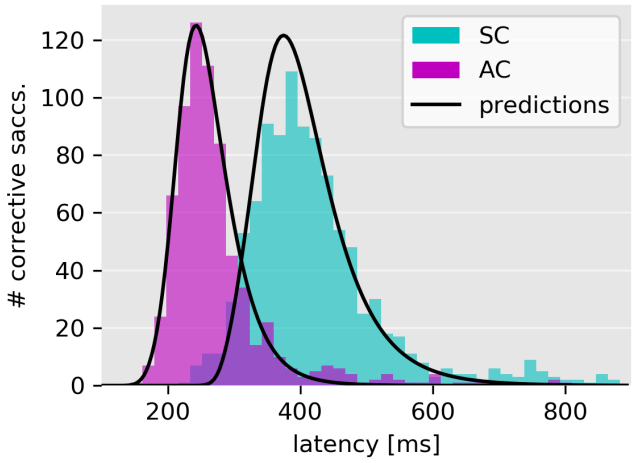


Fig. 7: Histogram of corrective antisaccades and model predictions. Depicted are the distributions of the hit times of the antisaccade unit and the histogram of the corrective antisaccades' RT. The location or time-shift of the predicted distributions was optimized using the data.

Model parameters: Inhibition failures and late errors

We then turned our attention to inhibition and volitional or late errors. These are defined by the probability that the late prosaccade unit hits threshold before the antisaccade unit on an antisaccade trial, and the

probability that the antisaccade unit hits threshold before the late prosaccade unit on a prosaccade trial, respectively. We also investigated the probability of an inhibition failure, i.e., the probability that the early unit hits threshold before all other units. On an antisaccade trial, an inhibition failure corresponds to a reflexive error.

In the SC condition (Fig. 8A), the findings were in line with our previous results (Aponte et al. 2017). While the probability of a late error on a prosaccade trial was negatively correlated with PP ($\mathrm{X}^2(2, N = 72) = 156.66, p < 10^{-5}$), the opposite behavior was observed for the probability of an inhibition failure on antisaccade trials ($\mathrm{X}^2(2, N = 72) = 22.5, p < 10^{-3}$) and late errors on antisaccade trials ($\mathrm{X}^2(2, N = 72) = 23.50, p < 10^{-5}$).

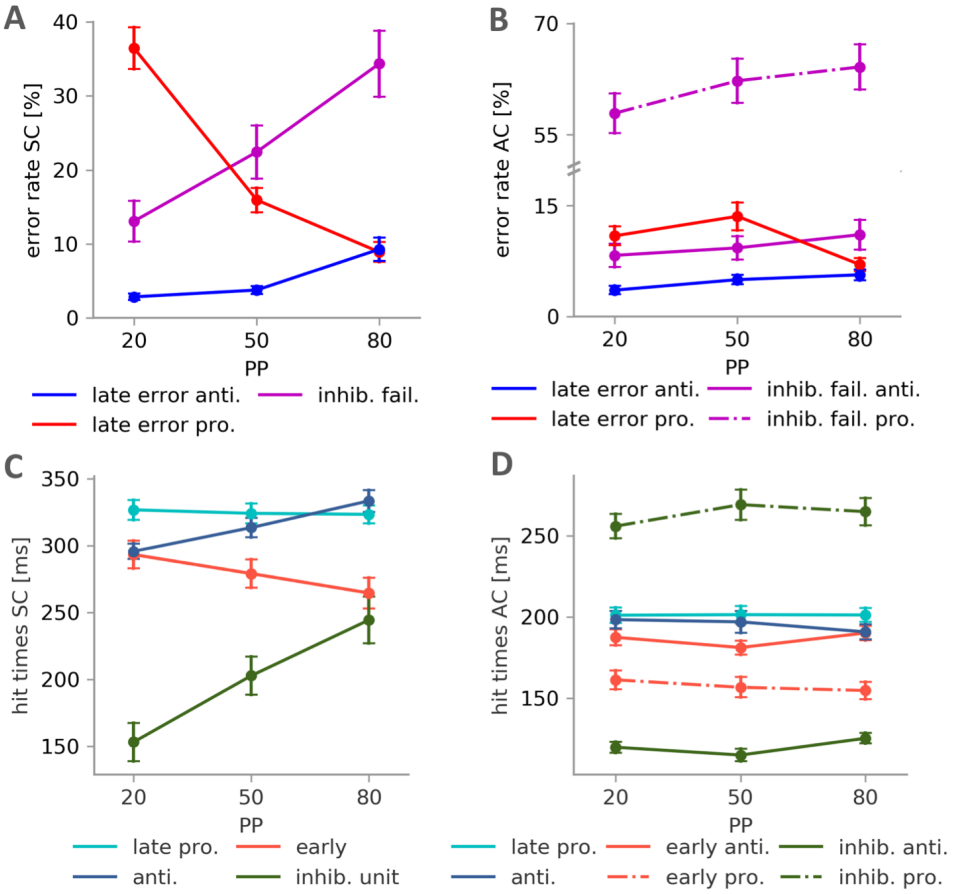


Fig. 8: A. Probability of late errors and inhibition failures in the SC condition. Late errors occur when an early prosaccade is stopped by the inhibitory unit, but the incorrect late action is performed. Non-stopped early reactions are called inhibition failures. **B. Probability of late errors and inhibition failures in the AS condition. C. Expected hit time of the units in the SC condition.** Note that we report a single estimate for the early and inhibitory unit because in a constrained model both units are assumed to have the same behavior across trial types. **D. Expected hit time of the units in the AC condition.**

By contrast, in the AC condition it was necessary to consider the number of inhibition failures on pro- and antisaccade trials separately because model comparison favored models in which the early and inhibitory units behaved differently across trial types. We found (Fig. 8B) that the probability of an inhibition failure on prosaccade trials (mean 61%, std. 11) was much higher than on antisaccade trials (mean 9%, std. 8), indicating that most correct prosaccades were early, reflexive responses. When we considered the effect of PP in the AC condition, we found only

a significant effect on the probability of a late error in antisaccade trials ($\mathrm{X}^2(2, N = 72) = 6.31, p = 0.04$).

The percentage of late responses in prosaccade trials was estimated to be approximately 39% of all trials in the AC condition (see Fig. 8B and Table 3). In antisaccade trials, the percentage of inhibition failures was estimated to be 9% of all trials, or 61% of all errors. Hence, 39% of all errors could be attributed to the late decision process. In the SC condition, the number of antisaccade errors predicted by the model was approximately 2% higher than the empirical error rate. On average 21% of all errors in antisaccades were cataloged as late decision errors. To assess the posterior predictions of the model, we report the correlation coefficient between the empirical and predicted ER in Table 3.

Table 3

	Empirical and fitted error rates					
	PP20	PP50	PP80	PP20	PP50	PP80
	Antisaccade trials					
	SC			AC		
Empirical error rate [%]	14.06	22.79	38.22	11.56	13.81	16.83
Predicted error rate [%]	15.18	24.90	40.53	11.50	13.82	16.07
Correlation coefficient	0.99	0.97	0.98	0.99	0.94	0.99
Inhib. failures [%]	12.65	22.00	34.63	8.27	9.28	11.06
$\frac{100*\text{Late errors}}{\text{Late errors + inhib.fail.}}$	26.82	18.62	22.13	39.44	40.01	37.28
	Prosaccade trials					
Empirical error rate [%]	29.11	11.18	4.88	3.07	3.07	1.07
Predicted error rate [%]	30.91	11.70	5.21	3.13	3.14	1.12
Correlation coefficient	0.98	0.96	0.99	0.97	0.98	0.90
Inhib. failures [%]	13.08	22.42	34.34	57.86	62.23	64.10

Empirical and predicted error rate, inhibition failures, and late errors. In order to evaluate the error rate estimates, we display the correlation coefficient between the predicted and observed error rates. Please note that inhibition failures on prosaccade trials correspond to correct early prosaccades. Errors on prosaccade trials can only be explained as late, volitional errors.

We then investigated whether the percentage of inhibition failures in the SC condition was correlated with the percentage of inhibition failures on antisaccade trials in the AC condition. Results are displayed in Fig. 9. In each of the PP conditions, we found a significant correlation ($p < 0.005$), with a correlation coefficient between 0.67-0.77 (Fig. 9). This indicates

that the tendency of individual subjects to respond with an early saccade was comparable across task designs.

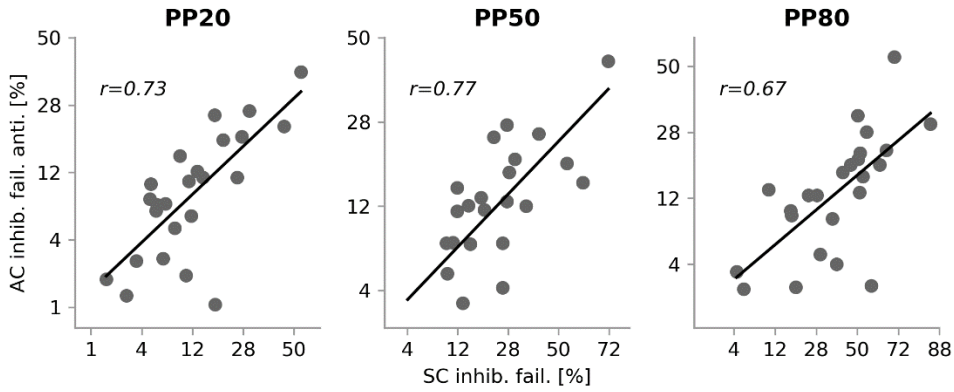


Fig. 9: Correlation of inhibition failures in antisaccade trials. Values are displayed on the probit scale. There was a significant and strong correlation between the percentage of inhibition failures across task designs and conditions.

Model parameters: Hit times

Finally, we investigated the effect of PP on the expected hit times of the units. In the SC condition (Fig. 8C), the early ($F_{2,46} = 7.39, p = 0.001$), as well as the antisaccade ($F_{2,46} = 36.34, p < 10^{-5}$) and inhibitory units ($F_{2,46} = 18.12, p < 10^{-5}$) were significantly affected by PP: High prosaccade trial probability led to slower inhibition, slower antisaccades, and faster early responses. However, we did not find a significant effect of PP on the hit times of the late prosaccade unit ($F_{2,46} = 0.22, p = 0.79$).

In the AC condition (Fig. 8D), most of the units had a much shorter hit time compared to the SC condition. Moreover, the fitted parameters suggested that most differences between pro- and antisaccade trials could be attributed to changes in the hit time of the inhibitory unit, which was over 100ms higher on prosaccade trials than in antisaccade trials. To further support this observation, we fitted a mixed Gamma SERIA model in which the early prosaccade unit (but not the inhibitory unit) was set to be equal across trial types. This is analogous to the restricted model originally proposed by (Noorani and Carpenter, 2013). This post-hoc model obtained the highest evidence in the AC condition ($\Delta LME > 7$ log units). Crucially, this model was also better than one in which the early unit but not the inhibitory unit was allowed to change

across trial types ($\Delta LME > 80$). Thus, most variance in the probability of early prosaccades could be explained by changes in the inhibitory unit, which indicates that cuing the trial type in advance of the saccade direction cue mainly influenced the inhibition of early responses.

There was no significant effect of PP on the hit time of the late pro- and antisaccade units (late pro: $F_{2,46} = 0.00, p = 0.99$; anti: $F_{2,46} = 2.08, p = 0.13$). However, we found a significant effect of PP on the inhibitory unit regardless of the trial type (pro. trials: $F_{2,46} = 3.23, p = 0.04$; anti. trials: $F_{2,46} = 14.11, p < 10^{-3}$). Finally, there was a significant effect of PP on the early unit in antisaccade trials ($F_{2,46} = 8.62, p = 10^{-3}$), but not on prosaccade trials ($F_{2,46} = 2.15, p = 0.12$). Taken together, our results suggest that manipulating the trial type probability in AC task had only an effect on the early and inhibitory units, and this effect was weak in prosaccade trials.

Discussion

The present study resulted in four main finding. First, the SERIA model better accounted for RT and ER than the PROSA model in both the SC and AC conditions. This indicates that even in AC designs, the prosaccade RT distribution is best described by more than one process. Second, according to the model fits, a significant proportion of errors on antisaccade trials were late errors, irrespective of the CUE condition. Third, we found that in the AC condition, the main factor explaining the differences in ER and RT between pro- and antisaccade trials was the hit time of the inhibitory unit and, consequently, the probability of inhibiting an early response. Finally, we found that the effects of manipulating the probability of a trial type were almost completely abolished when subjects were cued about task demands in advance of the peripheral cue. This suggests that SC task designs are more appropriate for studies interested in probability-dependent effects. Moreover, all effects of trial type probability were restricted to the early and inhibitory unit in the AC condition. We proceed to discuss these findings.

SERIA accounts for antisaccade behavior regardless of CUE condition

Arguably, the main novelty of the SERIA model is the distinction between early responses, which are always directed toward the PVC (i.e., a prosaccade) and can be inhibited by a stop process, and voluntary, late responses which can trigger both pro- and antisaccades. The units that trigger this type of saccades can generate rule guided behavior (e.g., an antisaccade), at the cost of higher RTs. Moreover, voluntary saccades are also subject to a race-to-threshold decision process (Aponte et al. 2017).

By contrast, involuntary and voluntary saccades are often distinguished by the paradigm in which these are elicited (Walker et al. 2000) and not by the mechanism that generates them: On one hand, involuntary saccades are associated with paradigms in which a suddenly displayed stimulus elicits a saccade. On the other hand, voluntary saccades are associated with paradigms in which the target needs to be retrieved from memory or it depends on specific task instructions, such as in the antisaccade task.

Because the SERIA model accounts for both reflex-like and 'voluntary' saccades towards a visual cue, the distinction between voluntary and

involuntary saccades can be reformulated in terms of the processes that generates them. Accordingly, the antisaccade 'cost' (Hallett 1978) might be also understood as a 'voluntary' saccade cost (ignoring remapping costs). This reconceptualization might explain the finding that under certain circumstances pro- and antisaccades exhibit the same latency (Liu et al. 2010; Weiler and Heath 2014); if all early responses are inhibited, pro- and antisaccades can have the same latency.

Qualitatively, evidence for the SERIA model can be easily observed in histograms of RT for the SC condition (cf. Fig. 5): RTs of correct prosaccades follow a bimodal distribution, and their late component resembles the distribution of correct antisaccades. Moreover, errors on prosaccades trials are relatively common in this version of the antisaccade task, and their latency is similar to the latency of correct antisaccades.

The main question that we addressed in this study is whether a similar mechanism could explain RT and ER distributions in an AC task design. Although correct prosaccade RTs do not show a bimodal distribution and errors in prosaccade trials are rare (<4%), model comparison and qualitative checks clearly indicate that prosaccade RT distributions in the AC condition can be better explained by a model that postulates early and voluntary prosaccades. Moreover, this model can predict the RT distribution of corrective antisaccades with surprising accuracy in both conditions.

Our data supports the idea that prosaccades do not appear to be bimodally distributed in the AC condition because voluntary prosaccades are fast enough to overlap with early prosaccades. This is obvious in Fig. 6 (bottom row), in which the distribution of correct prosaccades deviates from the linear pattern usually observed in other conditions (see Fig. 6 top row and Noorani and Carpenter 2016).

Early and late errors on antisaccade trials

SERIA provides a formal account of errors in the antisaccade task which distinguishes it from the two most prominent models in the literature. On the one hand, the model in Noorani and Carpenter (2013) does not incorporate a late decision process and thereby it explains all errors as inhibition failures. On the other hand, lateral inhibition models (Cutsuridis et al. 2007, 2014; Cutsuridis, 2015) explain errors as the

result of connected accumulators that represent pro- and antisaccades, without the intervention of a third inhibitory unit. Accordingly, an error occurs when a voluntary action does not inhibit a reflex-like prosaccade. Along this line, Reuter and colleagues (Reuter and Kathmann 2004) have argued that deficits in the ability to initiate an antisaccade contribute to the elevated ER observed in patients with schizophrenia.

The SERIA model is closer to the idea proposed by Fischer and colleagues (Fischer et al. 2000; Klein and Fischer 2005), who extended the distinction between 'express' and 'normal latency' saccades to antisaccade errors. Although conceptually similar to the approach presented here, these authors used a simple time threshold to distinguish between the two types of saccades (Klein and Fischer 2005). In this context, SERIA offers a model-based, statistically sound separation between early and late errors that goes beyond simple thresholding of RTs.

Hence, an important conclusion from our analysis is that late errors are a significant fraction of all errors regardless of task design. Concretely, in the present sample, approx. 39% of the errors on antisaccade trials in the AC condition were quantified as late errors, with large variability across subjects (Fig. 9). This number was estimated to be 21% in the SC condition. This is of significance, as the ability to separate between early and late errors might be of relevance in computational psychiatry and future patient studies (Fischer et al. 2000; Heinzle et al. 2016; Lo and Wang 2016; Coe and Munoz 2017).

AC vs. SC designs

The most obvious difference between the AC and SC conditions was an overall reduction in RT and ER in the AC task. This observation replicates previous findings (Weber 1995; Weiler and Heath 2014).

There are two main explanations for these differences. First, in the SC condition the mapping between a cue and an action can only be started once the peripheral stimulus is presented. Thus, one would expect robust inhibition of reactive saccades, that affords enough time to select the correct action (pro, or antisaccade) (Weber 1995). Second, in the AC condition subjects could anticipate the presentation of the peripheral cue because the task cue was always displayed for 700ms. Despite this

general reduction in RT, ERs were lower in the AC condition when compared to the SC condition.

Model comparison suggests differences in the type of anticipatory preparation in the two tasks: whereas in the SC condition the early and inhibitory unit followed a similar hit time distribution across trial types, this was not the case in the AC condition. Furthermore, a model in which the prosaccade unit was fixed across trial types obtained the highest model evidence, indicating that most of the differences in the number of early responses could be accounted for by changes in inhibitory control.

One interpretation of these findings is that in the SC condition, the peripheral cue does not influence the inhibition of early responses, because it is integrated in the decision-making process too late to strongly affect the early and inhibitory units. Nevertheless, contextual information about trial type probability can be exploited by the participants to drive inhibitory control. By contrast, in the AC condition early prosaccade inhibition is almost entirely determined by the trial type cue and only weakly modulated by the probability of a trial type, as discussed below.

Importantly, the probability of antisaccade errors was correlated between both CUE conditions. Thus, relative ERs were consistent across the two tasks, suggesting that the same cognitive processes are involved in both conditions. In conclusion, SC designs are likely to provide more variability in terms of ER and RT, for example under different PP conditions, while probing the same cognitive processes involved in an AC paradigm.

The effect of trial type probability

Our results replicate the finding that in the SC condition the probability of a trial type has a large impact on both ER and RT (Chiau et al. 2011; Aponte et al. 2017). Concretely, RTs of correct responses were negatively correlated with the corresponding trial type probability. These effects were strongly reduced in the AC condition, as reported before (Massen 2004; Pierce et al. 2015; Pierce and McDowell 2016a). Modeling indicated no significant effect of PP on late responses and a significant but relatively small effect on the early and inhibitory units. One interpretation of this is that the early presentation of the task cue in the AC condition, essentially removes all uncertainty about the task,

rendering the probabilistic manipulation largely un-effective, especially for late responses. This is in contrast to the SC condition, in which contextual information is of relevance for optimal, i.e. fast execution of the task. Thus, we conclude that the effects of contextual or prior information in the antisaccade task are best studied using the SC design.

Summary

This study investigated whether and to what extent cue presentation order (task cue and spatial cue) influenced ER and RT in the antisaccade task. Overall, we found that the impact of trial type probability was strongly reduced in the AC condition compared to the SC condition. From a modeling perspective, our results demonstrate that the combination of an early and a late race between voluntary pro- and antisaccades better accounts for RT and ER in an AC design, as compared to models that incorporate only an early race. Furthermore, modeling revealed that early inhibitory processes are strongly influenced by trial type in the AC condition, but not in the SC condition. By contrast, trial type probability had a strong effect on early units in the SC condition, but not in the AC condition. SERIA also provided a good prediction of the shape of the distribution of corrective antisaccades in both tasks. Finally, our quantitative analysis supports the hypothesis that a non-negligible fraction of errors in the antisaccade task can be categorized as late errors, irrespective of task design.

Software note

The models used here are available under the GPL license as part of the TAPAS toolbox (http://translationalneuromodeling.org/tapas/).

Acknowledgments

This work was supported by the René and Susanne Braginsky Foundation (KES) and the University of Zurich.

References

Amador N, Schlag-Rey M, Schlag J. Primate antisaccades. I. Behavioral characteristics. *J Neurophysiol* 80: 1775–1786, 1998.

Antoniades C, Ettinger U, Gaymard B, Gilchrist I, Kristjansson A, Kennard C, John Leigh R, Noorani I, Pouget P, Smyrnis N, Tarnowski A, Zee DS, Carpenter RH. An internationally standardised antisaccade protocol. *Vision Res* 84: 1–5, 2013.

Aponte EA, Raman S, Sengupta B, Penny WD, Stephan KE, Heinzle J. mpdcm: A toolbox for massively parallel dynamic causal modeling. *J Neurosci Methods* 257: 7–16, 2016.

Aponte EA, Schobi D, Stephan KE, Heinzle J. The Stochastic Early Reaction, Inhibition, and late Action (SERIA) model for antisaccades. *PLoS Comput Biol* 13: e1005692, 2017.

Calderhead B., Girolami MA. Estimating Bayes factors via thermodynamic integration and population MCMC. *Computational Statistics \& Data Analysis* 53: 4028–4045, 2009.

Bishop CM. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc, 2006.

Bittencourt J, Velasques B, Teixeira S, Basile LF, Salles JI, Nardi AE, Budde H, Cagy M, Piedade R, Ribeiro P. Saccadic eye movement applications for psychiatric disorders. *Neuropsychiatr Dis Treat* 9: 1393–1409, 2013.

Broerse A, Crawford TJ, Boer den JA. Parsing cognition in schizophrenia using saccadic eye movements: a selective overview. *Neuropsychologia* 39: 742–756, 2001.

Calkins ME, Curtis CE, Iacono WG, Grove WM. Antisaccade performance is impaired in medically and psychiatrically healthy biological relatives of schizophrenia patients. *Schizophr Res* 71: 167–178, 2004.

Camalier CR, Gotler A, Murthy A, Thompson KG, Logan GD, Palmeri TJ, Schall JD. Dynamics of saccade target selection: race model analysis of double step and search step saccade production in human and macaque. *Vision Res* 47: 2187–2211, 2007.

Carpenter RH, Williams ML. Neural computation of log likelihood in control of saccadic eye movements. *Nature* 377: 59–62, 1995.

Cherkasova MV, Manoach DS, Intriligator JM, Barton JJ. Antisaccades and task-switching: interactions in controlled processing. *Exp Brain Res* 144: 528–537, 2002.

Chiau HY, Tseng P, Su JH, Tzeng OJ, Hung DL, Muggleton NG, Juan CH. Trial type probability modulates the cost of antisaccades. *J Neurophysiol* 106: 515–526, 2011.

Coe BC, Munoz DP. Mechanisms of saccade suppression revealed in the anti-saccade task. *Philos Trans R Soc Lond, B, Biol Sci* 372, 2017.

Cribari-Neto F, Zeileis A. Beta regression in R.

Cutsuridis V, Kumari V, Ettinger U. Antisaccade performance in schizophrenia: a neural model of decision making in the superior colliculus. *Front Neurosci* 8: 13, 2014.

Cutsuridis V, Smyrnis N, Evdokimidis I, Perantonis S. A neural network model of decision making in an antisaccade task by the superior colliculus. *Neural Networks* 20: 690–704, 2007.

Cutsuridis V. Neural competition via lateral inhibition between decision processes and not a STOP signal accounts for the antisaccade performance in healthy and schizophrenia subjects. *Front Neurosci* 9: 5, 2015.

Cutsuridis V. Behavioural and computational varieties of response inhibition in eye movements. *Philos Trans R Soc Lond, B, Biol Sci* 372, 2017.

Everling S, Fischer B. The antisaccade: a review of basic research and clinical studies. *Neuropsychologia* 36: 885–899, 1998.

Fischer B, Gezeck S, Hartnegg K. On the production and correction of involuntary prosaccades in a gap antisaccade task. *Vision Res* 40: 2211–2217, 2000.

Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2003.

Gelman A, Meng XL. Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling. *Statistical Science* 13: 163–185, 1998.

Gelman A, Shalizi CR. Philosophy and the practice of Bayesian statistics. *Br J Math Stat Psychol* 66: 8–38, 2013.

Gooding DC, Basso MA. The tell-tale tasks: a review of saccadic research in psychiatric patient populations. *Brain Cogn* 68: 371–390, 2008.

Hallett PE. Primary and secondary saccades to goals defined by instructions. *Vision Res* 18: 1279–1296, 1978.

Heinzle J, Aponte EA, Stephan KE. Computational models of eye movements and their application to schizophrenia. *Current Opinion in Behavioral Sciences* 11: 21–29, 2016.

Hutton SB, Ettinger U. The antisaccade task as a research tool in psychopathology: a critical review. *Psychophysiology* 43: 302–313, 2006.

Irving EL, Tajik-Parvinchi DJ, Lillakas L, Gonzalez EG, Steinbach MJ. Mixed pro and antisaccade performance in children and adults. *Brain Res* 1255: 67–74, 2009.

Johnston K, Koval MJ, Lomber SG, Everling S. Macaque dorsolateral prefrontal cortex does not suppress saccade-related activity in the superior colliculus. *Cereb Cortex* 24: 1373–1388, 2014.

Kass RE, Raftery AE. Bayes factors. *Journal of the American statistical association* 90: 773–795, 1995.

Klein C, Fischer B. Developmental fractionation and differential discrimination of the anti-saccadic direction error. *Exp Brain Res* 165: 132–138, 2005.

Klein C, Rauh R, Biscaldi M. Cognitive correlates of anti-saccade task performance. *Exp Brain Res* 203: 759–764, 2010.

Koval MJ, Hutchison RM, Lomber SG, Everling S. Effects of unilateral deactivations of dorsolateral prefrontal cortex and anterior cingulate cortex on saccadic eye movements. *J Neurophysiol* 111: 787–803, 2014.

Levy DL, Mendell NR, LaVancher CA, Brownstein J, Krastoshevsky O, Teraspulsky L, McManus KS, Lo Y, Bloom R, Matthysse S, Holzman PS. Disinhibition in antisaccade performance in schizophrenia In M. F. Lenzenweger & R. H. Dworkin (Eds.), Origins and development of schizophrenia: Advances in experimental psychopathology (pp. 185-210). Washington, DC, US: American Psychological Association, 1998.

Liu CL, Chiau HY, Tseng P, Hung DL, Tzeng OJ, Muggleton NG, Juan CH. Antisaccade cost is modulated by contextual experience of location probability. *J Neurophysiol* 103: 1438–1447, 2010.

Lo CC, Wang XJ. Conflict Resolution as Near-Threshold Decision-Making: A Spiking Neural Circuit Model with Two-Stage Competition for Antisaccadic Task. *PLoS Comput Biol* 12: e1005081, 2016.

Logan GD, Cowan WB, Davis KA. On the ability to inhibit simple and choice reaction time responses: a model and a method. *J Exp Psychol Hum Percept Perform* 10: 276–291, 1984.

Massen C. Parallel programming of exogenous and endogenous components in the antisaccade task. *Q J Exp Psychol A* 57: 475–498, 2004.

Myles JB, Rossell SL, Phillipou A, Thomas E, Gurvich C. Insights to the schizophrenia continuum: A systematic review of saccadic eye movements in schizotypy and biological relatives of schizophrenia patients. *Neurosci Biobehav Rev* 72: 278–300, 2017.

Noorani I, Carpenter RH. Antisaccades as decisions: LATER model predicts latency distributions and error responses. *Eur J Neurosci* 37: 330–338, 2013.

Noorani I, Carpenter RH. The LATER model of reaction time and decision. *Neurosci Biobehav Rev* 64: 229–251, 2016.

Noorani I. Towards a unifying mechanism for cancelling movements. *Philos Trans R Soc Lond, B, Biol Sci* 372, 2017.

O'Driscoll GA, Depatie L, Holahan AL, Savion-Lemieux T, Barr RG, Jolicoeur C, Douglas VI. Executive functions and methylphenidate response in subtypes of attention-deficit/hyperactivity disorder. *Biol Psychiatry* 57: 1452–1460, 2005.

Peirce JW. PsychoPy--Psychophysics software in Python. *J Neurosci Methods* 162: 8–13, 2007.

Peirce JW. Generating Stimuli for Neuroscience Using PsychoPy. *Front Neuroinform* 2: 10, 2008.

Penny WD, Stephan KE, Daunizeau J, Rosa MJ, Friston KJ, Schofield TM, Leff AP. Comparing families of dynamic causal models. *PLoS Comput Biol* 6: e1000709, 2010.

Pierce JE, McCardel JB, McDowell JE. Trial-type probability and task-switching effects on behavioral response characteristics in a mixed saccade task. *Exp Brain Res* 233: 959–969, 2015.

Pierce JE, McDowell JE. Effects of preparation time and trial type probability on performance of anti- and pro-saccades. *Acta Psychol (Amst)* 164: 188–194, 2016a.

Pierce JE, McDowell JE. Modulation of cognitive control levels via manipulation of saccade trial-type probability assessed with event-related BOLD fMRI. *J Neurophysiol* 115: 763–772, 2016b.

Reilly JL, Frankovich K, Hill S, Gershon ES, Keefe RS, Keshavan MS, Pearlson GD, Tamminga CA, Sweeney JA. Elevated antisaccade error rate as an intermediate phenotype for psychosis across diagnostic categories. *Schizophr Bull* 40: 1011–1021, 2014.

Reuter B, Kathmann N. Using saccade tasks as a tool to analyze executive dysfunctions in schizophrenia. *Acta Psychol (Amst)* 115: 255–269, 2004.

Reuter B, Philipp AM, Koch I, Kathmann N. Effects of switching between leftward and rightward pro- and antisaccades. *Biol Psychol* 72: 88–95, 2006.

Reuter B, Rakusan L, Kathmanna N. Poor antisaccade performance in schizophrenia: an inhibition deficit? *Psychiatry Res* 135: 1–10, 2005.

Sato TR, Schall JD. Effects of stimulus-response compatibility on neural selection in frontal eye field. *Neuron* 38: 637–648, 2003.

Stampe D. Heuristic filtering and reliable calibration methods for video-based pupil-tracking systems. Behavior Research Methods, Instruments, \& Computers 25: 137–142, 1993.

Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. Bayesian model selection for group studies. Neuroimage 46: 1004–1017, 2009.

Trappenberg TP, Dorris MC, Munoz DP, Klein RM. A model of saccade initiation based on the competitive integration of exogenous and endogenous signals in the superior colliculus. *J Cogn Neurosci* 13: 256–271, 2001.

Vijayraghavan S, Major AJ, Everling S. Dopamine D1 and D2 Receptors Make Dissociable Contributions to Dorsolateral Prefrontal Cortical

Regulation of Rule-Guided Oculomotor Behavior. *Cell Rep* 16: 805–816, 2016.

Walker R, Walker DG, Husain M, Kennard C. Control of voluntary and reflexive saccades. *Exp Brain Res* 130: 540–544, 2000.

Weber H. Presaccadic processes in the generation of pro and anti saccades in human subjects--a reaction-time study. *Perception* 24: 1265–1280, 1995.

Weiler J, Heath M. Task-switching in oculomotor control: unidirectional switch-cost when alternating between pro- and antisaccades. *Neurosci Lett* 530: 150–154, 2012.

Weiler J, Heath M. Oculomotor task switching: alternating from a nonstandard to a standard response yields the unidirectional prosaccade switcitch-cost. *J Neurophysiol* 112: 2176–2184, 2014.

Supplementary

Supplementary Methods and Supplementary Figure S1

To infer the model parameters of all subjects we used the likelihood function of the PROSA (Eq. 1-5, in the main text) and SERIA (Eq. 5-9 in the main text) models and assumed a hierarchical prior, such that the parameters of all subjects in each CUE condition were estimated simultaneously. Figure S1 summarizes entire model as a graphical model following the conventions in Bishop (2006).

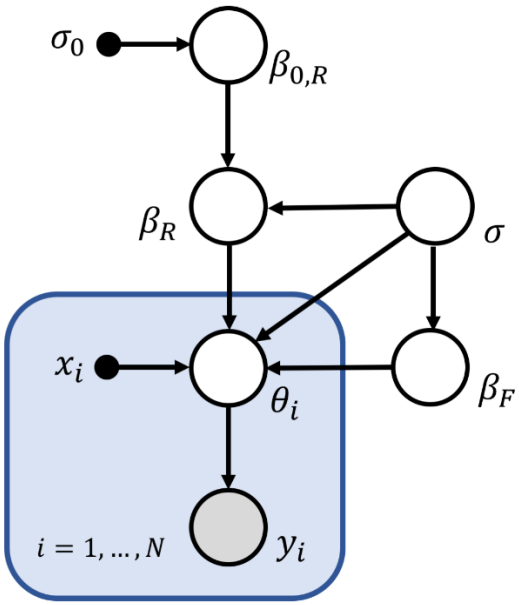


Figure S1: Graphical representation of the hierarchical model. Graphical summary of the statistical model using the convention in Bishop (2006). Briefly, each circle represents a probability distribution and arrows indicate conditional dependence. Black dots represent fixed parameters. The equations describing the distributions of each node are given below. The most important feature of the model is that the prior distribution of each set of parameters $\boldsymbol{\theta_i}$ is parametrically defined by a set of explanatory variables $\boldsymbol{x_i}$ and coefficients $\boldsymbol{\beta}$ with variance $\boldsymbol{\sigma^2}$. These coefficients are estimated from the population distribution. We partition parameters $\boldsymbol{\beta}$ into fixed ($\boldsymbol{\beta_F}$) and random effects ($\boldsymbol{\beta_R}$), such that the latter have a prior mean estimated again from the population distribution. For the present study, random effects represent subject specific intercepts, while their mean (or global intercept) is modeled by $\boldsymbol{\beta_{0,R}}$, whose prior distribution is assumed to be centered at zero with variance $\boldsymbol{\sigma_0^2}$.

In the following, we present the mathematical description of the model depicted in Figure S1. We will write down the conditional probabilities

that specify individual nodes of the graphical model, starting from the likelihood and then moving up the hierarchy. To simplify notation, we assume that the data in each run i, i.e., the RT and actions in a single block of each subject, are represented by a vector y_i and a parameter θ_i. The extension to a multivariate model is straightforward under the assumption that different parameters are conditionally independent. Although this independence assumption can be questioned, it facilitates the construction of a hierarchical extension of the SERIA and PROSA models.

The likelihood of the model (represented by the gray shaded circle in Figure S1) is given by the product of the likelihood of all runs, i.e. all sessions in all subjects:

$$p(y_1, \dots, y_N|\theta_1, \dots, \theta_N) = \Pi_{i=1}^{N} p(y_i|\theta_i). \quad (S1)$$

The precise definition of the likelihood is given by the equations in the main text (Eq. 1-9). The prior distribution of parameters θ_i is given by

$$p(\theta_i|x_i, \beta, \sigma) = \mathcal{N}(\theta_i; x_i \cdot \beta, \sigma^2). \quad (S2)$$

$X = (x_1, \dots, x_N)$ is a design matrix of size $M \times N$ that codes M explanatory variables, such as SUBJECT, PP, etc., and β is a vector of dimension $M \times 1$ that represents the effect of each explanatory variable. The prior distribution of β is given by

$$p(\beta|\beta_0, \sigma) = \mathcal{N}(\beta; \beta_0, \sigma^2) \quad (S3)$$

and the prior probability of σ is

$$p(\sigma^{-2}) = \Gamma(\sigma^{-2}; a, b). \quad (S4)$$

We distinguish between two types of β coefficients in analogy to the concepts of fixed and random effects. For fixed effects β_F, we assumed that the coefficients have a fixed prior mean $\beta_{0,F} = 0$. For random effects β_R, we assume that the prior mean $\beta_{0,R}$ is a random variable drawn from a distribution that represents the sample population, i.e. all subjects,

$$p(\beta_{0,R}|\beta_R, \sigma, \sigma_0) \propto p(\beta_R|\beta_{0,R}, \sigma) p(\beta_{0,R}|0, \sigma_0) \quad (S5)$$

$$\propto \mathcal{N}(\beta_R; \beta_{0,R}, \sigma^2) \mathcal{N}(\beta_{0,R}; 0, \sigma_0^2) \quad (S6)$$

The conditional posterior of $\beta_{0,R}$ can be computed analytically and is given by

$$p(\beta_{0,R}|\beta_R, \sigma, \sigma_0) = \mathcal{N}(\beta_{0,R}; \mu_R, \pi_R^{-1}), \quad (S7)$$

where

$$\mu_R = \frac{\sigma^{-2}\sum_{i=1}^{N}\beta_{i,R}}{N\sigma^{-2}+\sigma_0^{-2}}, \quad (S8)$$

$$\pi_R = (N\sigma^{-2}+\sigma_0^{-2}). \quad (S9)$$

The rationale for including a random effect is to account for the idiosyncrasies of each subject with the parameters β_R while modeling a population wide intercept $\beta_{0,R}$.

All equations defined above are linear and rely on conjugate priors. Hence, it is possible to derive Gibbs steps to sample from the conditional posterior distributions of all parameters with the exception of $\theta_{1,..,N}$, which are sampled from a Gaussian kernel centered at the previous sample.

Supplementary Table S1: Model comparison by accuracy

Table S1

Model accuracy			
Model	**Parametric family**	**SC**	**AC**
	PROSA		
m_1	T. normal	224.6	338.9
m_2	T. normal	0.0	0.0
m_3	Mixed Gamma	607.1	376.3
m_4	Mixed Gamma	499.0	72.4
m_5	Gamma	613.2	347.8
m_6	Gamma	531.2	126.1
	DORA		
m_7	T. normal	1302.8	1011.0
m_8	T. normal	1291.4	812.3
m_9	Mixed	1351.4	**998.7**
m_{10}	Mixed	1260.8	600.7
m_{11}	Gamma	**1390.9**	893.0
m_{12}	Gamma	1305.2	813.6

Expected log likelihood (accuracy) normalized by subtracting the lowest log likelihood (m2) from all estimates. The accuracy of a model is the expected log likelihood of the model. It is tightly related to the unpenalized R^2, or total variance explained, of a model.

Supplementary Figure S2: Model fits for representative subjects – synchronous cue

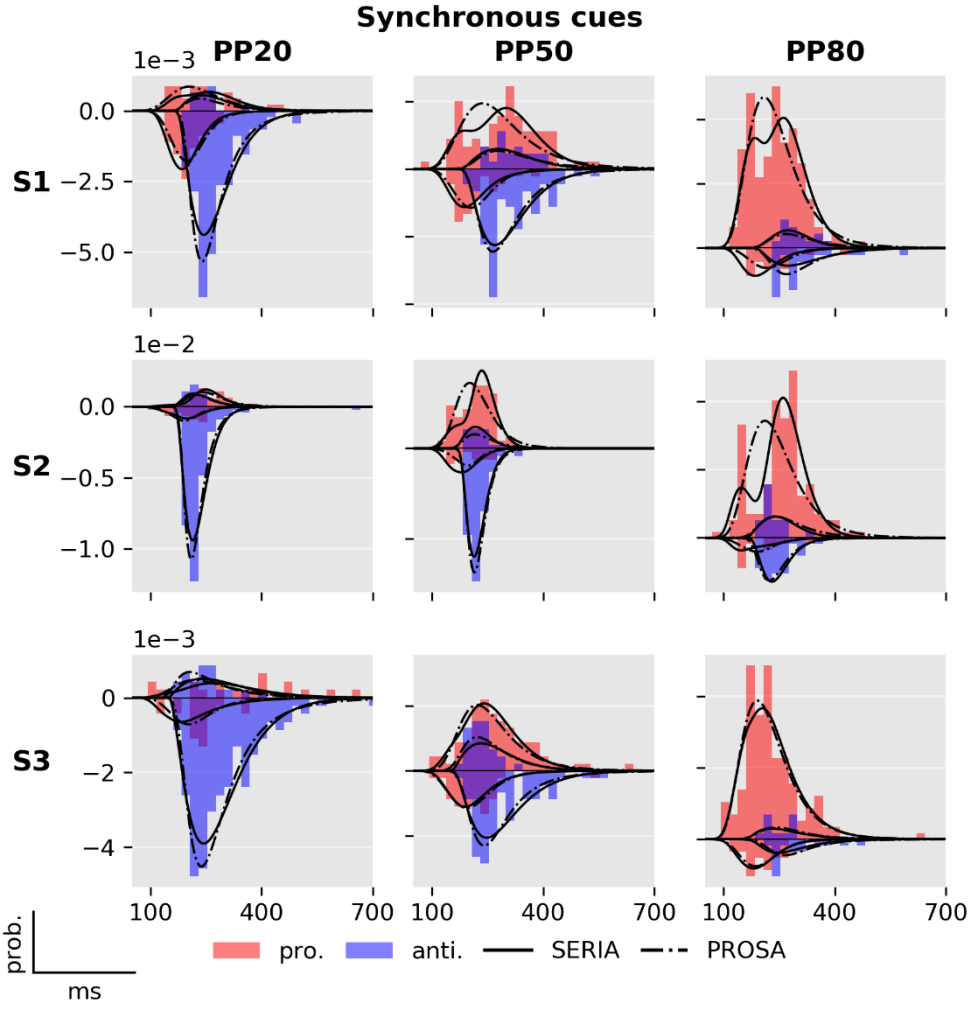


Figure S2: RT distribution and comparison of the SERIA (m_{10}) and PROSA (m_5) models in the SC condition in three representative subjects, each one displayed in a different row. Prosaccade trials are displayed in the upper half plane, antisaccade trials in the bottom half. Note that the two models are the best models of their classes, respectively.

Supplementary Figure S3: Model fits for representative subjects – asynchronous cue

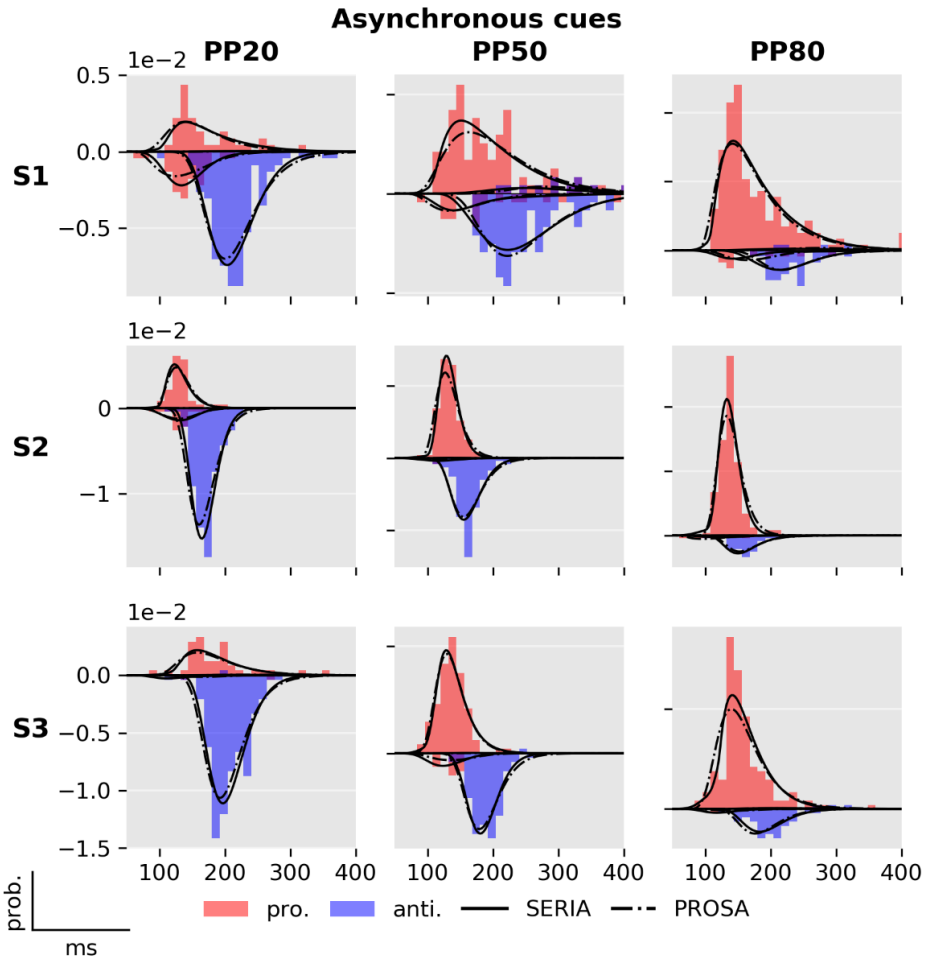


Figure S3: Comparison of the SERIA (m_9) and PROSA (m_5) models in the AC condition in three representative subjects, each one displayed in a different row. Note that the two models are the best in their respective families.

Supplementary Figure S4: Reciprobit plots of the PROSA model

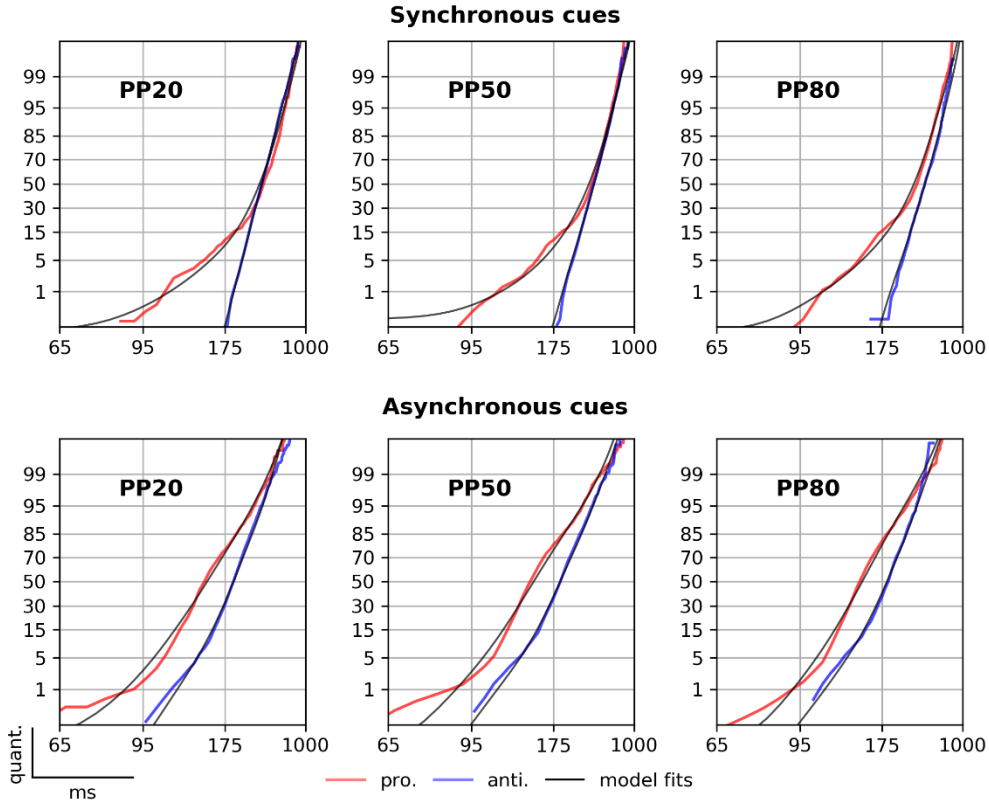


Figure S4.: Empirical and predicted reciprobit plots of the PROSA model. Same as in Figure 6, but using the PROSA model.

Chapter 8

When required to alternate between two tasks, humans tend to be slower and less accurate than when asked to perform the same behavior repeatedly. Although the first study dedicated to this phenomenon dates back to the beginning of the 20th century (Jersild, 1927 reviewed in Monsell, 2003), only since the seminal work of Allport and colleagues (Allport et al., 1994), this topic started to receive attention in the psychological literature.

Around twenty behavioral studies have investigated this phenomenon in the mixed antisaccade task. Surprisingly, while switch costs have been clearly demonstrated when subjects perform a prosaccade after an antisaccade, it is not clear whether the same costs occur in the opposite direction. As discussed later in this chapter, positive, negative, and none switch costs have been reported, sometimes by the same researchers. More importantly, no unified explanation of these dissimilar results has been proposed.

Here, we apply SERIA to a fraction of the data reported in the last chapter with two simple questions in mind: Do switching costs arise in the version of the antisaccade task that we have used in Chapter 6 and 7? More interestingly yet, if switch costs do occur, can we impute them to any of the two main decision processes postulated by SERIA: the inhibition of early responses and the initiation of voluntary actions.

Our results demonstrate RT and ER switching costs regardless of trial type. This is a somewhat surprising outcome in that, from all the studies we are aware of, only a single publication documented a similar effect (Barton et al., 2006). Interestingly, in that study subjects were cued about the task they were supposed to perform (pro- or antisaccade), only 200ms in advance of the peripheral spatial cue. As the synchronous antisaccade task, this design reduces the amount of time that subjects have to prepare how to act in a trial.

Using model comparison, we demonstrate that task switching costs affect both the early and late decision process postulated by SERIA. When we explored this result using the parameters of the model, a subtle picture emerged: On the one hand, switch voluntary pro- and antisaccades were slower and less accurate than their counterparts. On the other hand, antisaccade trials enhanced inhibitory control in the next trial, reducing the number of inhibition failures and decreasing their latency.

As mentioned above, these results seem at odds with the extant literature, in which only the study by Barton and colleagues (Barton et al., 2006) reported a similar effect. The most plausible explanation is that previous studies have masked switching costs in voluntary actions by cueing subjects about the task demands long in advance of the peripheral cue.

The next chapter explores a further application of SERIA to a more clinically relevant question: What are the effects of pro-dopaminergic and pro-cholinergic compounds in the antisaccade task?

Switch costs in inhibitory control and goal-directed behavior: A computational study of the antisaccade task

Eduardo A. Aponte[1,], Klaas E. Stephan[1,2,3], Jakob Heinzle[1]*

[1] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich. Wilfriedstrasse 6, 8004, Zurich, Switzerland.

[2] Wellcome Trust Centre for Neuroimaging, University College London. 12 Queen Square London WC1N 3BG,.

[3] Max Planck Institute for Metabolism Research. Gleueler Strasse 50, 50931, Cologne, Germany.

[*]Corresponding author:

Eduardo A. Aponte; aponte@biomed.ee.ethz.ch

Abstract

When instructed to alternate between different tasks, humans require more time and are less accurate than when they repeat the same task. This decrease in performance is referred to as the 'task switch cost'. Several behavioral studies have investigated this phenomenon in the mixed antisaccade task, in which participants are cued to saccade either in the same or in the opposite direction of a peripheral stimulus. These reports have provided conflicting answers as to whether a cue to saccade away from a target leads to positive, negative, or no switch costs. These contradictory findings are paralleled by two opposing theoretical hypotheses which focus on *task-set inertia* and *oculomotor inhibition*, respectively. Here, we applied a computational, generative model to data from a mixed antisaccade task in which the peripheral visual stimulus also served as a trial type cue. Behaviorally, we found reaction time and error rate switch costs on pro- and antisaccade trials. Modeling revealed that these costs were due to two different effects. First, antisaccade trials enhanced inhibitory control on the following trial, resulting in fewer but faster inhibition failures. Second, goal-directed actions displayed a task-inertia effect, in that rule these were slower on switch trials compared to repeat trials. Our results shed light on the general phenomena of task switching, by demonstrating two types of switch cost that affect differently the inhibition of habitual responses and the initiation of goal-directed actions.

Introduction

One hallmark of higher-order cognition is the ability to alternate between habitual and non-habitual, goal-directed actions (Isoda and Hikosaka, 2008). However, alternating between different tasks engenders costs in terms of reaction time (RT) and error rate (ER), and switching from a non-habitual to a habitual response sometimes leads to larger costs than the opposite transition (Allport et al., 1994). An attractive paradigm to study these phenomena in the oculomotor domain is the antisaccade task (Hallett, 1978; Munoz and Everling, 2004), in which a habitual response – a prosaccade towards a salient peripheral stimulus – needs to be overwritten by a non-habitual action, i.e., an antisaccade in the opposite direction. Importantly, this paradigm has received much attention in psychiatric research, because changes in ER and RT constitute a stable finding in schizophrenia (Greenwood et al., 2007; Reilly et al., 2014; Radant et al., 2015; Myles et al., 2017).

Behaviorally, switch costs in the antisaccade task have been investigated in great detail (Barton et al., 2002; Cherkasova et al., 2002; Manoach et al., 2002; Bojko et al., 2004; Fecteau et al., 2004; Barton et al., 2006a; 2006b; Rivaud-Pechoux et al., 2007; Ansari et al., 2008; Mueller et al., 2009; Lee et al., 2011; Weiler and Heath, 2012a; 2012b; DeSimone et al., 2014; Weiler and Heath, 2014; Weiler et al., 2014; Heath et al., 2015; Pierce et al., 2015; Weiler et al., 2015; Heath et al., 2016; Chan et al., 2017). Despite the large number of studies, no unified picture of the cost of switching in this paradigm has emerged. Concretely, all studies we are aware of have reported that switch prosaccades, i.e. correct prosaccades that follow an antisaccade trial, have higher latencies than repeat trials, whereas the costs associated with switch antisaccades are less clear. For example, some studies have indicated that switch antisaccades display lower RTs (e.g. Cherkasova et al., 2002), while others have reported both lower and higher RTs (e.g., Barton et al., 2006a), and yet others indicate no switch costs (e.g., Weiler and Heath, 2012b).

One path to clarify the relationship between seemingly contradictory experimental evidence is the application of generative models to empirical data (Monsell, 2003; Karayanidis et al., 2010; Heinzle et al., 2016) which might help disentangle the underlying mechanisms that explain switch costs. In this direction, we recently developed the

Stochastic Early Reaction, Inhibition and late Action (SERIA) model (Aponte et al., 2017) of the antisaccade task. In essence, SERIA combines the 'horse-race' model usually applied in the context of the countermanding saccade task (Logan et al., 1984; Camalier et al., 2007) to explain the inhibition of habitual, fast prosaccades, with a second race between two voluntary, or goal-directed actions that can generate both pro- and antisaccades. In contrast to previous models (Noorani and Carpenter, 2013), we acknowledge the possibility that prosaccades can be generated not only as the result of reactive (habitual) saccades, but also by a rule-guided decision process.

In two recent studies (Aponte et al., 2017; Chapter 7), we showed that SERIA can be used to account for RT and ER distributions with great accuracy, and that it is able to predict corrective antisaccades. Here, we employ the model to investigate switch costs and whether they could be attributed either to the inhibition of habitual responses or to the generation of voluntary saccades. In other words, we compare the two dominant theories of antisaccade switch costs – the task-set inertia hypothesis (Allport et al., 1994) and the oculomotor inhibition hypothesis (Barton et al., 2006a; Weiler and Heath, 2014)– by a modeling approach that inherently separates these two processes. Our results demonstrate that switch costs are largely explained by the interference between voluntary pro- and antisaccades, irrespective of trial type. In addition, antisaccade trials lead to increased inhibitory control, which reduces the number of reactive saccades following an antisaccade trial, leading to an apparent switch cost on prosaccade trials.

Methods

In this study, we analyzed a subset (one of two tasks) of the data presented in detail in Chapter 7. In the following, we briefly summarize the experimental procedures relevant for the analysis presented here. This study was approved by the ethics board of the Canton of Zurich, Switzerland (KEK-ZH-Nr.2014-0246) and was conducted according to the Declaration of Helsinki.

Participants

We used the data of all twenty-four, healthy male subjects that were analyzed in Chapter 7. All subjects had normal or corrected to normal vision and provided a written inform consent to participate in the study.

Apparatus

The experiment was conducted in a dimly illuminated room. Subjects sat 60cm in front of a computer screen (41.4x30cm; Philips 20B40; refresh rate 85Hz). Eye position was recorded at a sampling rate of 1000Hz with a remote, infrared eye tracker (Eyelink 1000; SR Research, Ottawa, Canada). Head position was stabilized using a chin rest. The experiment was controlled by in-house software written in the Python programming language (2.7) using the PsychoPy package (1.82.02) (Peirce, 2007; 2008).

Experimental design

Here, we only considered the data from the synchronous task condition in Chapter 7. Subjects participated in three blocks of mixed pro- and antisaccade trials. Each block consisted of 200 trials of which either 20, 50, or 80% were prosaccade trials. Before the main experiment, subjects underwent a training block of 50 prosaccade trials followed by 50 antisaccade trials. In the training phase (but not during the main experiment) subjects received feedback about their performance.

As shown in Fig. 4A, two red circles (radius 0.25°) were presented throughout the experiment at an eccentricity of ±12°. Each trial started with a central fixation cross (0.6x0.6°). Subjects were required to fixate for at least 500ms, after which a random interval of 500 to 1000ms started. Completed this period, the fixation cross disappeared, and a green bar (3.48x0.8°) centered on one of the red circles was presented in either horizontal or vertical orientation for 500ms. Subjects were

instructed to saccade to the red circle ipsilateral to a horizontal green bar (prosaccade trials), and to saccade to the contralateral circle in the opposite case (antisaccade trials). After 1000ms the next trial started. Horizontal and vertical bars were randomly interleaved, but the same sequence was presented across all subjects. The location (left of right) of the peripheral cue was also randomly permuted, such that the number of pro- and antisaccade trials in each direction was the same.

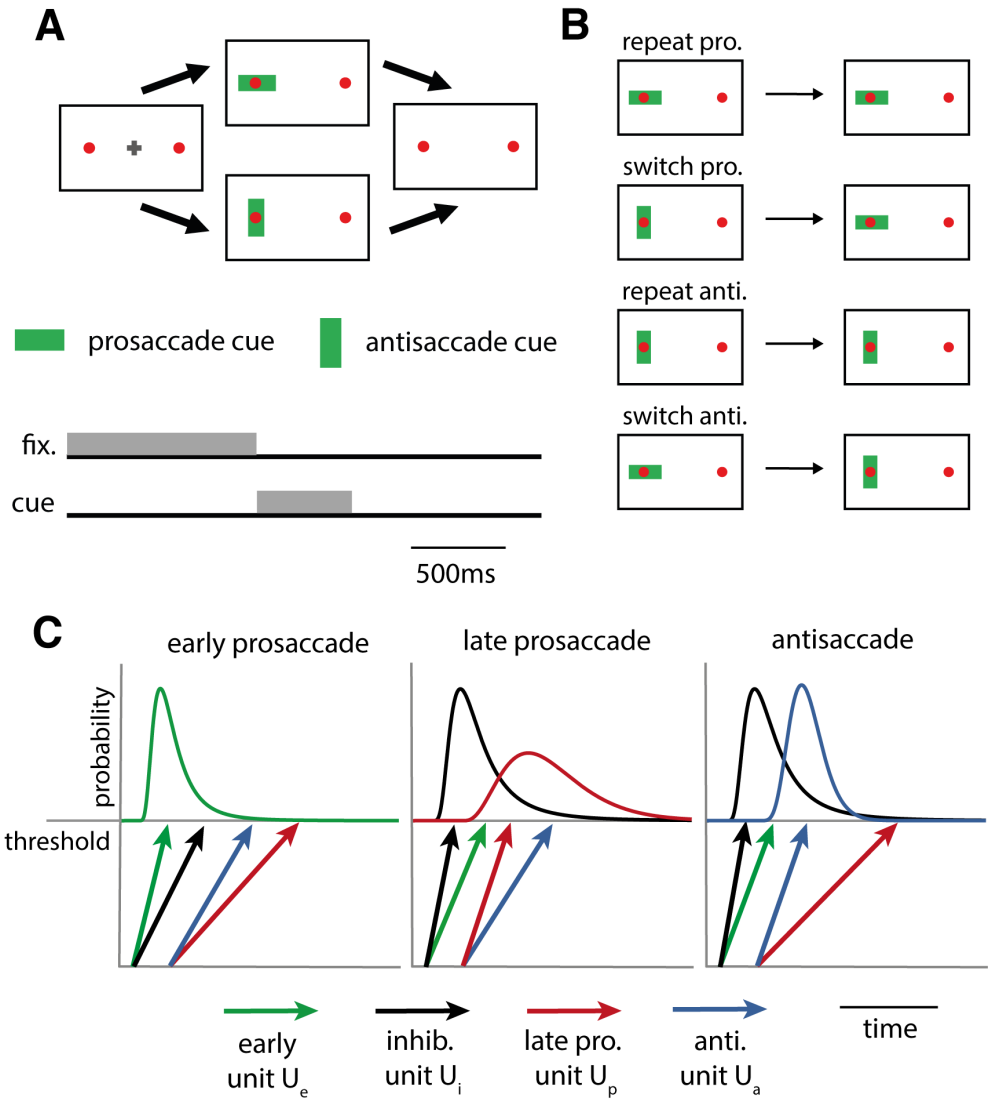


Figure 4: Experimental design and model. Adapted from Chapter 7. **A. Experimental design**. Subjects were instructed to first fixate to a central cross. After a variable interval (500-1000ms), a cue indicating the trial type was presented for 500 ms. **B. Trial types.** Depending on the cue presented in the previous trial, four types of trials were considered. From top to bottom: repeat prosaccade, switch prosaccade, repeat antisaccade, and switch antisaccade. The cue was presented to the right or left pseudorandomly. **C. SERIA model.** The SERIA model is a race model that incorporates four different units: an early prosaccade unit, an inhibitory unit, a late prosaccade and an antisaccade unit (see legend). The RT distributions are a function of the hit time distribution of the four units. Early reactions occur when the early unit hits threshold before all other units. Late reactions occur mainly when early reactions are stopped by the inhibitory unit. Note that late pro- and antisaccade compete with each other in a similar manner. In principle, different trial types could affect the hit time distribution of each unit.

Data processing

Saccades were detected with the software provided by the eye tracker manufacturer (Stampe, 1993), which uses a 22°/s and 3800°/s2 threshold to define the start of a saccade. Only saccades with a magnitude larger than 2° were included in the analysis. Trials were rejected in case of an eye blink, if subjects failed to maintain fixation before the peripheral green bar was presented, and if their latency was above 900ms or below 50ms, and, in the case of antisaccades, below 95ms.

Statistical Analysis

As variables of interest, we investigated mean RT of correct responses and mean ER. These were analyzed with a generalized, linear mixed effect (GLME) model implemented in MATLAB 9.3 (*fitglme.m*). Independent variables were prosaccade probability (PP) with levels 20, 50 and 80%; trial type (TT); last trial (LT) with levels *switch* and *repeat* (Fig. 4B); and SUBJECT entered as a random effect. In the case of ER, the probit function acted as link function in the GLME. The first trial of a block was not included in this analysis.

The SERIA model

Briefly, SERIA (Aponte et al., 2017) models the race of four independent accumulators or units: an early (u_e), an inhibitory (u_i), a late prosaccade (u_p), and an antisaccade (u_a) unit. The pair of an action $A \in \{pro., anti.\}$ and its latency $T \in [0, \infty[$ is treated as a random variable, whose distribution is a function of the hit times of each of the units, U_e, U_i, U_p, and U_a respectively. Conceptually, SERIA can be decomposed into two different competitions: First, the early unit, which models reactive, habitual responses, generates a prosaccade at time t if it hits threshold at time t (i.e., $U_e = t$) and all the other units hit threshold afterwards. An early response can be stopped by the inhibitory unit if the latter hits threshold at some earlier point. In that case, either a late prosaccade or an antisaccade is generated. This second decision process is modeled as the race between the corresponding units, such that, for example, a late prosaccade at time t is generated only if the late prosaccade unit hits threshold at $U_p = t$ before the antisaccade unit (i.e., $U_a > t$).

More concretely, SERIA provides an explicit formula for the probability of an action *A* and its RT. First, a prosaccade at time *t* is generated when

either the early unit u_e hits threshold at time t (i.e., $U_e = t$) before all other units. The probability of this event is given by

$$p(U_e = t)p\big(U_p > t\big)p(U_a > t)p(U_i > t). \quad (1)$$

Furthermore, a prosaccade at time t can be triggered when the late prosaccade unit hits threshold at t and before all other units

$$p\big(U_p = t\big)p(U_e > t)p(U_a > t)p(U_i > t) \quad (2)$$

or an early response is stopped by the inhibitory unit (i.e., $U_i < t$ and $U_i < U_e$), and the late prosaccade unit hits threshold before the antisaccade unit

$$p\big(U_p = t\big)p(U_a > t)\int_0^t p(U_i = \tau)p(U_e > \tau)d\tau\,. \quad (3)$$

Similarly, an antisaccade at time t is generated when the antisaccade unit hits threshold at t ($U_a = t$), before all other units

$$p(U_a = t)p(U_e > t)p\big(U_p > t\big)p(U_i > t) \quad (4)$$

or after an early prosaccade has been stopped

$$p(U_a = t)p\big(U_p > t\big)\int_0^t p(U_i = \tau)p(U_e > \tau)d\tau\,. \quad (5)$$

To fit the model, we assumed a parametric form for the hit times of each of units: the hit times of the early (U_e) and inhibitory unit (U_i) were modeled with an inverse Gamma distribution, while the hit times of the late units (U_p and U_a) were modeled using a Gamma distribution (Aponte et al., 2017). Thus, each unit could be fully characterized by two parameters controlling the mean and variance of the hit times. Accordingly, to fully specify the distribution of actions and RTs in a condition, 8 (unit) parameters were required.

Model space

We aimed to answer three different questions through Bayesian model comparison (Kass and Raftery, 1995; Stephan et al., 2009): First, are models that included information about the previous trial superior in explaining experimental data compared to models which did not include this factor? Second, are inter-trial effects driven by either the trial type, the action, or the correctness of the action performed in the previous trial? Third, can inter-trial effects be accounted for by changes in either the generation of goal-directed actions, inhibitory control, or a combination of both?

To answer these questions, we fitted models that explained the totality of the data not only as a function of the current trial type, but also as a function of the previous trial. For this, all trials were divided into four different conditions, according to the cue displayed (pro- or antisaccade) and the previous trial, as explained below. Although a different set of parameters could operate in each condition, this seems biologically implausible and our goal was to identify which parameters could be fixed across conditions, without compromising the ability of the models to parsimoniously explain the behavioral data. As illustrated in Table 4, we first evaluated models in which the parameters of the units were constant irrespective of the previous trial but could vary depending on the cue presented on the current trial. Second, we evaluated models that accounted for effects of the last trial in three manners: either the parameters of the model depended on (i) the previous trial type (Fig. 4B), (ii) the previous action, or on (iii) whether the previous action was an error or not. In principle, one could also consider interactions among these three factors, but this would require $2x2x2x2x8=128$ parameters. Hence, we limited the analysis to the three options mentioned above and ignored their interaction, acknowledging that such effects have been reported in the literature (DeSimone et al., 2014).

These families of models were divided into three nested versions that represented different types of inter-trial effects: either the parameters of the inhibitory and early unit were fixed across conditions, but the late units were allowed to vary across conditions, or only inhibitory control changed across conditions, or both the late and the inhibitory units could differ in all conditions, but the early units were equal across them.

Table 4

	Model space					
	Late units		**Inhib. units**		**Late+inhib. units**	
	Model	# Pars.	Model	# Pars.	Model	# Pars.
No switch	m_1	15	m_2	13	m_3	17
Switch trial	m_4	23	m_5	21	m_6	29
Last action	m_7	23	m_8	21	m_9	29
Last error	m_{10}	23	m_{11}	21	m_{12}	29

Model space. In addition to the 8 parameters controlling the units, all models included three parameters that accounted for the 'no response' time, the late response delay, i.e., a constant delay associated with all goal-directed saccades and, finally, the probability of an outlier, i.e., the probability of a saccade faster than the no-response-time. In model m_1, we assumed that the early and inhibitory units were equal in pro- and antisaccade trials (4 parameters), but the late units were allowed to vary across pro- and antisaccade trials (4x2 parameters). Thus, the total number of parameters was 3+4+4x2=15. In the case of models m_4, m_7 and m_{10}, the parameters of the late units varied in all four possible conditions, resulting in 2x4 (two switch trials times four parameters) additional degrees of freedom (15+2x4=23). In model m_2, the inhibitory unit was allowed to vary between pro- and antisaccade trials, but the late units were fixed across the two conditions. Thus, m_2 had 13 parameters, i.e., two parameters less than m_1. In models m_5, m_8 and m_{11}, we allowed the inhibitory unit to vary across all four conditions. Note that m_2 had the identical late decision process in pro- and antisaccade trials. We relaxed this severe restriction in models m_5, m_8 and m_{11} and allowed the late units to vary between pro- and antisaccade trials, but kept them fix across switching. Thus, theses three models had 3x2+4+13=21 parameters. Lastly, in the *late+inhib.* family we allowed both the inhibitory and late units to vary, keeping the parameters of the early unit fixed. In model m_3 this implied two more parameters when compared to m_1, while in models m_6, m_9 and m_{12} this led to 3x2 more parameters than e.g. m_4.

Model fitting

The models were estimated using the techniques presented in our previous studies (Aponte et al., 2017). Data from all subjects were entered simultaneously into a hierarchical model presented in Chapter 7. Samples from the posterior distribution were drawn using the Metropolis-Hasting algorithm. The evidence or marginal likelihood of

models was computed using thermodynamic integration with 32 parallel chains ordered according to the temperature schedule suggested by Chapter 7. The algorithm was run for 60000 iterations, from which only the last half was used to compute summary statistics. The software used here to implement the models and inference is available in the TAPAS toolbox (http://translationalneuromodeling.org/tapas/).

We were interested in several model-based statistics derived from the fits. First, we evaluated the probability of an inhibition failure, defined as the probability that the early unit hits threshold before all the other units:

$$p(inhib.fail.) = \int_0^\infty p(U_e = t)p(U_i > t)p\big(U_p > t\big)p(U_a > t)dt\,. \quad (6)$$

Inhibition failures are fast, reflexive prosaccades, which are correct on prosaccade trials and errors on antisaccade trials. The expected RT of an inhibition failure is

$$E[inhib.fail.RT] = \frac{1}{p(inhib.fail.)}\int_0^\infty t\,p(U_e = t)p(U_i > t)p\big(U_p > t\big)p(U_a > t)dt\,. (7)$$

We also report the conditional probability of a late prosaccade, defined as the probability that the late prosaccade unit hits threshold before the antisaccade unit:

$$p(late\ pro.) = \int_0^\infty p\big(U_p = t\big)p(U_a > t)dt\,. \quad (8)$$

Note that the conditional probability (given no inhibition failure) of a late antisaccade is defined as

$$p(anti.) = 1 - p(late\ pro.). \quad (9)$$

We were also interested in the expected hit times of the late units, defined as

$$E[late\ pro.hit\ time] = \frac{1}{p(late\ pro.)}\int_0^\infty t\,p\big(U_p = t\big)p(U_a > t)dt \quad (10)$$

and analogously so for antisaccades. This quantity is the expected hit time of the late prosaccade unit, conditioned on the antisaccade unit arriving at a later point. We report this statistic, as it conveys an interpretable quantity that can be readily compared to experimental data. The derivation for these terms can be found in Aponte et al.

(2017). Finally, we computed the probability of a saccade at time t to be an early response:

$$p(early|T = t, A = pro) = \frac{p(early\ pro, T=t)}{p(early\ pro, T=t)+p(late\ pro, T=t)}. \quad (11)$$

Results

The results are based on all trials of the experiment, after roughly 2.2% of the trials were discarded due to blinks or other artefacts, as explained in detail in Chapter 7. We first report a classical statistical analysis of ER and RT. Only correct trials entered the analysis of RT. Then, we proceed to the model-based findings.

Switch costs: Error rate

Mean ERs (Fig. 5A-C) were submitted to a binomial regression model. On prosaccade trials, there was a significant effect of both PP ($F(2,138) = 36.4, p < 10^{-3}$) and LT ($F(1,138) = 178.7, p < 10^{-3}$). On average, switch prosaccade trials showed 15% more errors than repeat trials. In the case of antisaccades, the main effects of PP ($F(2,138) = 17.1, p < 10^{-3}$) and LT ($F(1,138) = 147.7, p < 10^{-3}$) as well as the interaction between PP and LT ($F(2,138) = 4.3, p = 0.01$) were significant. Switch trials led to 18% more errors on antisaccades trials. When we included all trials in a single model, there was a significant effect of TT (anti. ER – pro. ER=9%; $F(1,276) = 112.4, p < 10^{-3}$) and a significant interaction between TT and LT ($F(1,276) = 4.8, p = 0.02$), indicating a significant difference in switch cost between anti- and prosaccade trials ($\text{anti. switch cost} - \text{pro. switch cost} = 2.8\%$).

Reaction time

To evaluate the effect of previous trials, we submitted the mean RT of correct trials (Fig. 5D-F), to a GLME that included PP, LT, and SUBJECT as independent variables. There was a significant effect of LT on both pro, and antisaccades trials ($\text{pro. switch cost} = 14ms, F(1,138) = 37.1, p < 10^{-3}$; $\text{anti. switch cost} = 12ms, F(1,138) = 32.7, p < 10^{-3}$). The effect of PP was significant on antisaccade ($F(1,138) = 6.6, p < 10^{-3}$) but not on prosaccade trials ($F(1,138) = 0.9, p = 0.39$). In both cases, the interaction between PP and LT was not significant. We then considered pro- and antisaccades in a single model with factors PP, LT, TT and all possible interactions. Our main interest was the interaction between the factors LT and TT, which would indicate a significant difference in the switch cost between pro- and antisaccades. This interaction was not significant ($F(2,276) = 0.2, p = 0.64$).

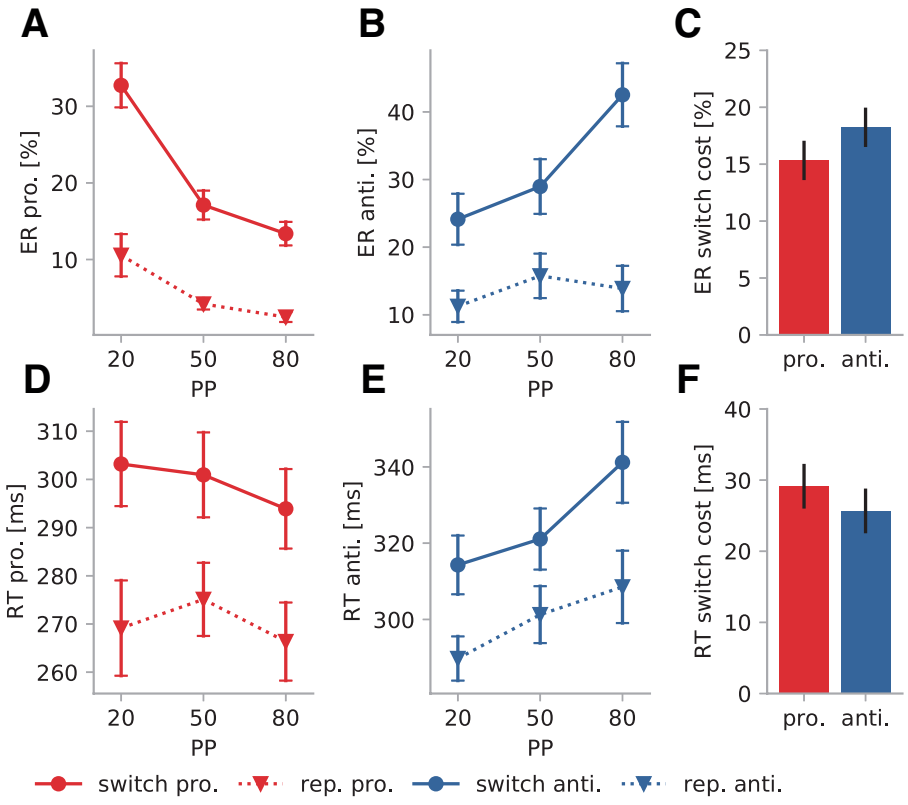


Figure 5: **A.** Mean ER on prosaccade trials. **B.** Mean ER on antisaccade trials. **C.** ER switch costs. **D.** Mean RT on prosaccade trials. **E.** Mean RT on antisaccade trials. **E.** RT switch cost. Error bars display the s.e.m.. PP: prosaccade probability

SERIA results – model comparison

All models were initially evaluated according to their log evidence or marginal likelihood, which corresponds to the accuracy or expected log likelihood adjusted by the complexity of the model (Stephan et al., 2009). Table 2 (top) reports the evidence of all models in log units. The model with the highest evidence (m_6, LME = -17467.6, ΔLME>8.0 log units) allowed for differences in the late and inhibitory units across all conditions. In addition, the accuracy or expected log likelihood is reported in Table 2 (bottom). In general, models in which switch costs depended on the last trial type, as opposed to the last action or error, obtained higher evidence.

Table 2 illustrates the penalization for complexity (or number of parameters) in terms of the difference between log model evidence (top) and expected log likelihood (bottom) in each model. For example, in the case of the *no-switch:late+inhib.* model (m_3) the penalty was 817.8 log

units. Similarly, the penalty for *switch:inhib* (m_5) was 857.7 units, whereas for the best model (*switch:late+inhib*; m_6) the penalty was 985.5 log units. This demonstrates that the *switch:late+inhib* model provided the most parsimonious explanation of the data as the increase in expected log likelihood outweighed the increased complexity.

Table 5

	Log model evidence		
	Late	Inhib.	Late+inhib.
No switch	-17596.0	-19571.0	-17591.4
Switch trial	-17481.1	-17475.8	**-17467.6**
Last action	-17599.1	-17577.1	-17634.8
Last error	-17657.5	-17623.0	-17706.1
	Expected log-likelihood		
No switch	-16842.8	-18806.0	-16773.6
Last trial type	-16532.2	-16618.1	**-16482.1**
Last action	-16661.9	-16797.4	-16598.0
Last error	-16751.1	-16804.4	-16715.5

Model comparison. The model with the highest evidence is high lightened in bold face. A difference of 3 log units is typically regarded as strong evidence (Kass and Raftery, 1995). The expected log-likelihood is also displayed for comparison. Note that within all model families (rows) the difference in expected log-likelihood between the most complex and the other models is strongly reduced (for last trial type) or even inverted (last action or last error) in log model evidence due to the penalization for complexity.

To illustrate qualitative differences between models, the fits of m_4 (*switch:late*), m_5 (*switch:inhib.*) and m_6 (*switch:late+inhib.*) are displayed in Fig. 6 and 7. Fits represent the expected posterior predictive probability estimated from posterior samples. Visual inspection suggests that the *switch:inhib.* model failed to capture the distribution of late responses, and particularly so on prosaccade switch trials. The *switch:late* model made a better job regarding late saccades, but it did not seem to capture early responses on, for example, prosaccade repeat trials (Fig. 6 bottom row). Finally, the *switch:late+inhib.* model was able to accommodate most of the features of the behavioral data, with perhaps the exception of the early peak on prosaccade switch trials.

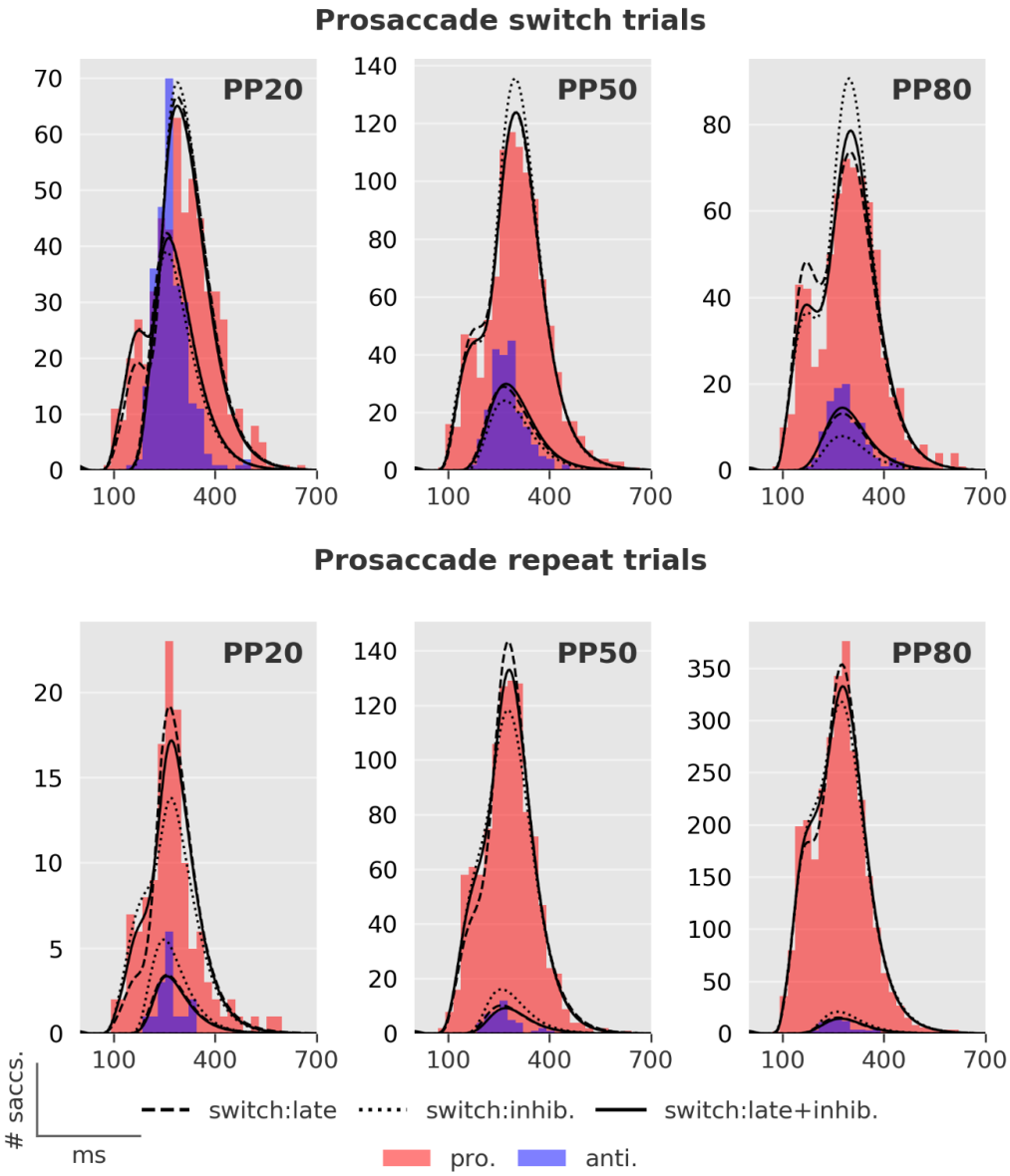


Figure 6: Histogram of prosaccade trials and model fits. Prosaccades are displayed in red, and antisaccades in blue. For comparison, the weighted posterior predictive distribution of models m_4 (*switch:late*), m_5 (*switch:inhib.*) and m_6 (*switch:late+inhib.*) computed from posterior samples are displayed. The *switch:late+inhib.* model fitted both the early and late peaks in the prosaccade RT distribution better.

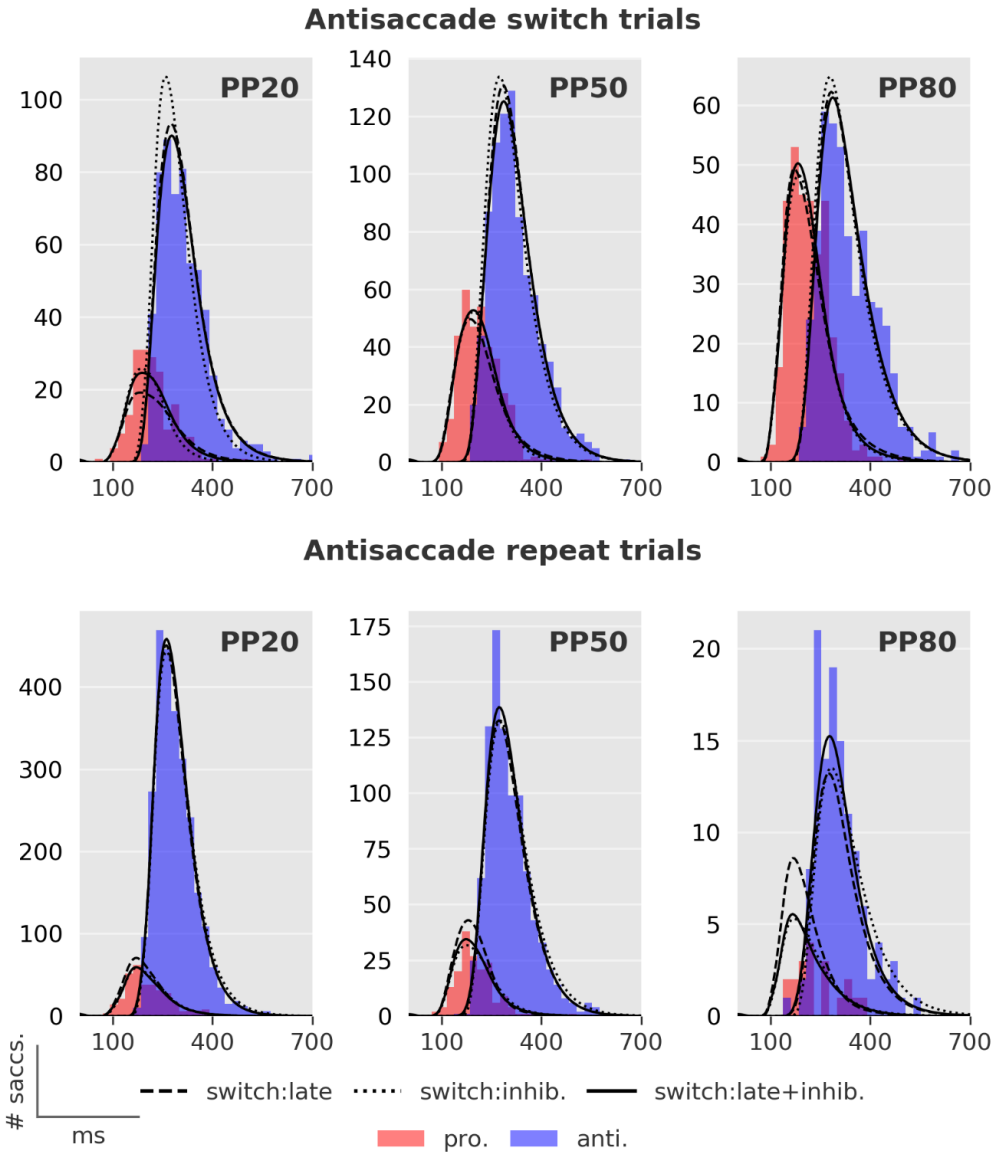


Figure 7: Histogram of antisaccade trials and model fits. For a detailed legend see Fig. 6. Note that all three models showed a relatively poor model fit for errors in the PP80 condition in repeat trials. This can be explained by the very low number of errors (N=19 across all subjects) in this condition.

SERIA results – parameter estimates

In the second part of the analysis, we investigated how model-based parameter estimates differed across conditions. For this, we used the parameter estimates from the model with the highest evidence, namely the *switch:late+inhib.* model. The focus of this analysis was the hit times of the late units and the early responses, as well as the expected

proportion of late errors, and the number of inhibition failures in switch and repeat trials.

Threshold hit times

The expected hit time of the late pro- (327ms, std. 35ms) and antisaccade (314ms, std. 33ms) units (Fig. 8A-B) were first submitted to two separate GLMs. The main effect of LT was significant for the two late units (pro. switch cost = 16ms, std. 19ms, $F(1,138) = 28.6, p < 10^{-3}$; anti. switch cost = 23ms, std. 21ms, $F(1,138) = 46.6, p < 10^{-3}$). When considered together, we found no significant interaction between LT and TT ($F(1,276) = 2.4, p = 0.11$). Finally, the interaction between PP and TT was significant ($F(1,276) = 5.0, p = 0.007$).

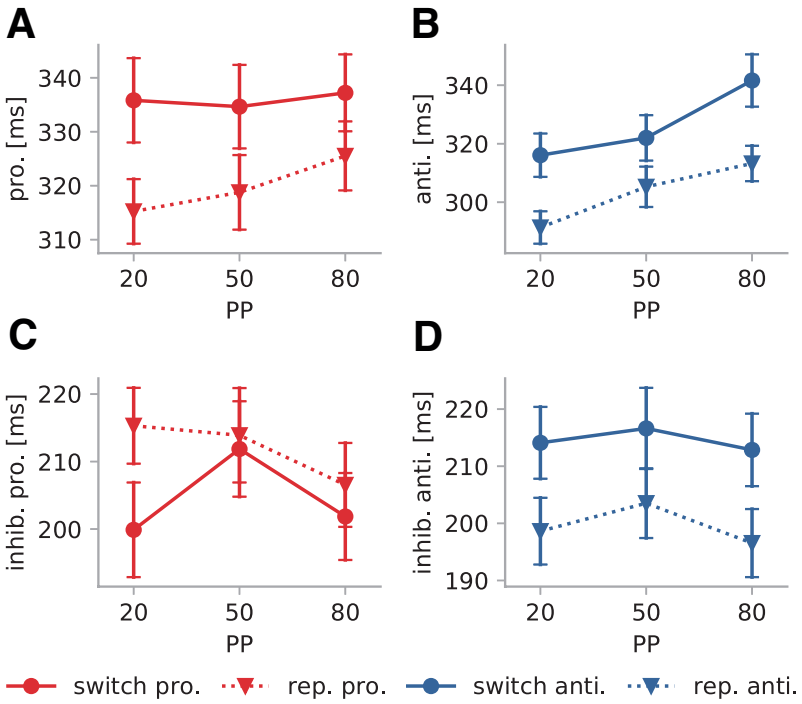


Figure 8: **A.** Hit times of the late prosaccade unit (see Eq. 10). **B.** Hit times of the antisaccade unit. **C.** RT of inhibition failures on prosaccade trials (see Eq. 7). **D.** RT of inhibition failures in antisaccade trials. Error bars depict the s.e.m..

In addition to the late units, we investigated the latency of inhibition failures (i.e., non-stopped early reactions; Fig. 8C-D). In SERIA, these are always prosaccades and occur on both pro- and antisaccade trials when the early unit hits threshold before all other units. Note that inhibition failures are correct in prosaccade trials, but errors in antisaccade trials. When pro- and antisaccade trials were analyzed together, we found a significant effect of PP ($F(2,276) = 16.2, p <$

10^{-3}), LT ($F(1,276) = 4.0, p = 0.045$), and the interaction between TT and LT was significant, as well ($F(1,276) = 46.9, p < 10^{-3}$). As shown in Fig. 8D, inhibition failures had a higher mean latency on antisaccade switch trials (214ms, std. 32ms) than on repeat trials (199ms, std. 29ms), whereas on prosaccade trials (Fig. 8D), the opposite pattern occurred (switch 204ms, std. 33ms; rep. 211ms, std. 31ms). Thus, according to our model fits, inhibition failures following an antisaccade trial were faster than inhibition failures following a prosaccade trial.

To illustrate the above effect, we compared the mean RT of errors on switch (227ms, std. 36ms) and repeat antisaccade trials (221ms, std. 48). We have shown previously that most of the errors in this paradigm are due to inhibition failures (Aponte et al., 2017), and we therefore hypothesized that errors on antisaccade trials that follow prosaccade trials should have a higher latency than errors that follow antisaccade trials. In agreement with our prediction, we found a positive switch cost (4ms, std42) with respect to errors on antisaccade trials. Note that the scant number of errors on antisaccade repeat trials in the PP80 condition (N=19, see Fig. 7, bottom right panel) prevented us from conducting a statistical analysis. However, if the PP80 block was excluded, the switch cost was still positive (11ms; std. 30ms) and significant ($F(1,86) = 5.1, p = 0.025$).

Error types

In order to characterize different sources of errors, we investigated early and late errors separately (Fig. 9). First, we submitted the probability of late errors (Eq. 8 and 9) on pro- (mean 19%, std. 14) and antisaccade (mean 4%, std.4) trials to a single GLME. This revealed a positive switch cost as demonstrated by a significant effect of LT ($F(1,276) = 55.9, p < 10^{-3}$). Moreover, there was a significant interaction between LT and TT ($F(1,276) = 14.7, p < 10^{-3}$). The mean switch cost for late prosaccades was 11% (std. 11), whereas for antisaccades, it was 1% (std. 4.8). When late antisaccades were analyzed separately, the effect of the previous trials was still significant ($F(1,138) = 9.87, p = 0.002$).

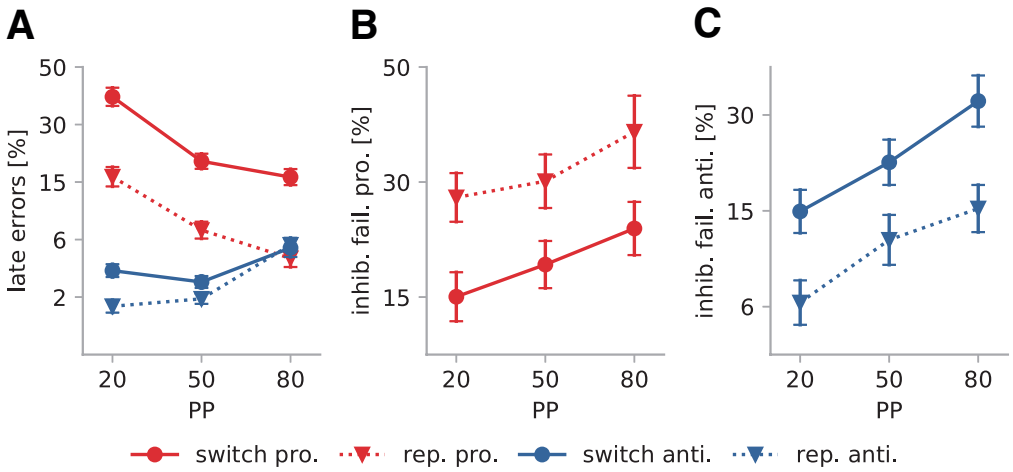


Figure 9: A. Percentage of late errors (see. Eq. 8 and 9). **B.** Percentage of inhibition failures in prosaccade trials (cf. Eq. 6). **C.** Percentage of inhibition failures in antisaccade trials. Percent errors are presented in the probit scale and summary statistics (mean and s.e.m.) were computed from probit transformed probabilities. Error bars represent the s.e.m..

We then investigated the probability of an inhibition failure (see Eq. 6), defined as the probability that the early unit hits threshold before all other units. According to our model, on prosaccade trials 28% (std. 19) of saccades were inhibition failures, whereas this number was lower on antisaccade trials (mean 21%, std. 18). The effect of LT on pro- ($F(1,138) = 105.5, p < 10^{-3}$) and antisaccades trials ($F(1,276) = 149.2, p < 10^{-3}$) was significant. When considered together, we found a significant interaction between the factors TT and LT ($F(1,276) = 219.8, p < 10^{-3}$). Concretely, prosaccade trials induced more inhibition failures on the next trial regardless of trial type (pro. switch cost=-12%, std. 10; anti. switch cost=11%, std. 9). Finally, while we found a weak but significant interaction between PP and TT ($F(1,276) = 3.2, p = 0.04$), the three-way interaction between PP, TT and LT was not significant ($F(2,276) = 1.3, p = 0.25$).

In Fig. 10, the effects of switch and repeat trials on early responses are exemplified. We depict the histogram of all responses in the PP50 condition sorted into trial types, and switch and repeat trials. We restrict this illustration to the PP50 case because the number of repeat and switch trials was similar (pro. switch: 2479, pro. repeat: 2150, anti. switch: 2435, anti. repeat 2150). The predicted histograms of early responses are overlaid on the empirical distributions. The former were

computed by weighting each saccade's RT by its probability of being an early response (Eq. 11). Clearly, on prosaccade trials, the early peak of repeat trials (orange) is larger than on switch trials (green), as emphasized by the predicted histograms. The opposite pattern is observable on antisaccade trials, in which switch trials have a larger early peak than repeat trials.

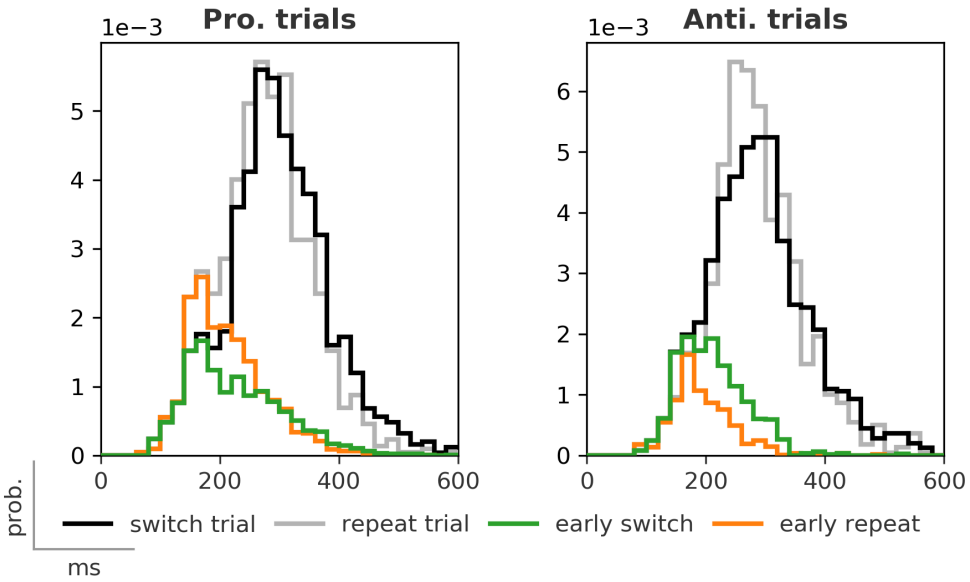


Figure 10: Normalized histograms of all saccades in the PP50 condition and predicted early responses. The histograms of all responses on pro- and antisaccade trials are sorted into switch and repeat trials. Histograms were normalized to have unit area. Overlaid are the predicted histogram of early responses, computed by weighting each saccade by its probability of being an early response (Eq. 11). In prosaccade trials, there were more early saccades in repeated trials than on switch trials, while the opposite pattern could be observed on antisaccade trials. Note that the black and gray histograms represent the empirical distributions, while orange and green histograms show modeling results.

Fig. 11 summarizes the effect of switching on the RT of early and late reactions, late ER, and the percentage of inhibition failures. Our findings indicate a strong positive switch cost in RT and ER for late responses, with the exception of late errors on antisaccade trials. Regarding inhibition failures, antisaccade trials decreased the frequency and latency of early reactions on the subsequent trial compared to prosaccade trials. This observation is compatible with the claim that an antisaccade trial leads to enhanced inhibitory control on the following trial. The reason is that in a race model, when the distribution of the hit times of the stop unit is shifted to the right, more early reactions with

higher RT are possible. In other words, faster inhibitory control leads to *fewer but faster* errors.

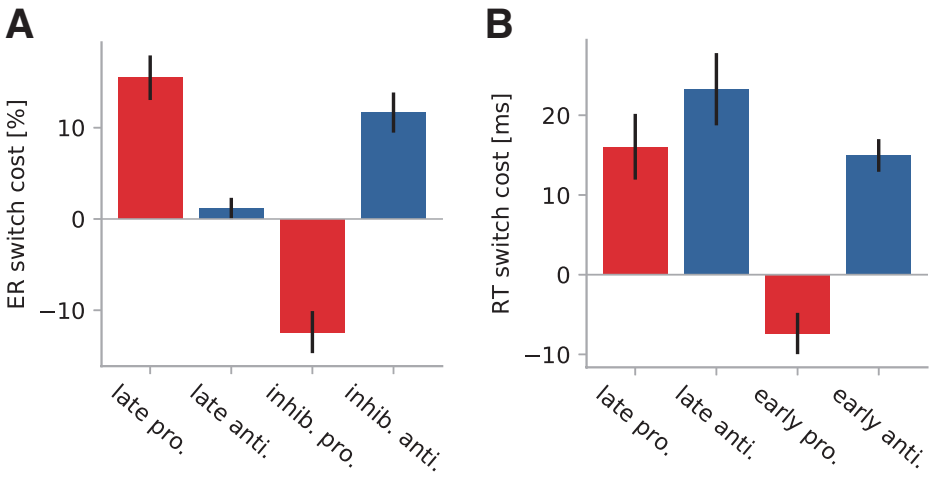


Figure 11: Summary of switch costs. A. Error rates: Task switching increase the probability of late errors in pro- and antisaccade trials. This cost was higher on prosaccade trials when compared to antisaccade trials. In contrast, we found that there was a negative switch cost in the number of inhibition failures on prosaccade trials, while we observed the opposite effect on antisaccade trials. Thus, prosaccade trials lead to more inhibition failures on the subsequent trial irrespective of trial type. **B. Hit times.** There was a positive and significant switch cost for late pro- and antisaccades. Early reactions that followed a prosaccade trial showed a higher latency than those following an antisaccade trial. Error bars display the s.e.m..

Discussion

In the present study, we investigated switch costs in a mixed pro- and antisaccade task with the SERIA model. This allowed us to quantify to what extent switch costs affected the inhibition of habitual reactions (early prosaccades) and goal-directed (rule-guided) behavior (late pro- and antisaccades), respectively. Modeling revealed two distinguishable effects: on the one hand, switch trials engendered RT costs in goal-directed behavior. On the other hand, antisaccade trials enhanced inhibitory control compared to prosaccade trials, as demonstrated by fewer inhibition failures with lower latency following antisaccade trials. In the following, we discuss these findings.

Behavioral findings

Two types of switch costs have been commonly reported for the antisaccade task. Early studies (e.g., Barton et al., 2002; Cherkasova et al., 2002; Manoach et al., 2002; Fecteau et al., 2004) reported positive prosaccade RT switch costs, negative antisaccade RT switch costs, as well as an increase in ER following switch trials in both type of trials. More recently, Heath and Weiler (e.g. Weiler and Heath, 2012b; Weiler et al., 2015) have reported a positive switch cost on prosaccade trials, and no switch cost on antisaccade trials.

Our behavioral analysis based on standard statistics supports positive switch costs in pro– and antisaccades, both in terms of ER and RT. Although our study seems at odds with previous reports, we believe that our findings complement rather than contradict them, because the design used here differs in key aspect from the tasks used by other researchers. Notably, in our paradigm (inspired by Sato and Schall, 2003), the peripheral spatial cue also signaled the action to be performed, whereas in most studies the trial type cue precedes the spatial cue by hundreds of milliseconds. We refer to these two paradigms (Chapter 7) as the synchronous cues (SC) and asynchronous cues (AC) designs, respectively. Importantly, it has been shown (e.g., Weiler and Heath, 2014) that displaying the task cue before the peripheral saccadic target has a dramatic effect on ER and RT when compared to the SC task. Supporting the idea that different designs can trigger different switch costs, Barton et al. (2006a) showed that when the task cue was presented only 200ms before the peripheral cue, there was a positive switch cost in both pro- and antisaccades. However, if the peripheral cue

was delayed by 1800ms, a negative antisaccade cost could be observed. Here, we replicated the observation that a very short cue-to-peripheral-stimulus delay (0ms in our case) causes switch costs regardless of trial type.

Weiler and Heath (2014) have also noted methodological differences across studies. For example, in studies reporting a negative antisaccade switch cost (e.g., Cherkasova et al., 2002), the saccadic targets (in our case, the two peripheral red points) were displayed during the entire experiment. This group has attributed the difference in findings to this manipulation (Barton et al., 2006a).

One possible explanation for the contrast between AC and SC designs is that if the trial type cue is presented early enough relative to the peripheral cue, the decision processes that triggers voluntary actions can be consolidated to the point that contextual factors, such as the previous trial or the probability of a task cue, become ineffective (Weiler and Heath, 2014). Thus, the AC might abolish task switching effects in voluntary action generation by allowing participants to overwrite contextual effects during the cue-to-stimulus interval.

We speculate that seemingly contradictory reports could be explained by a single computational mechanism. In other words, a unified explanation of behavioral findings might be attained through the lens of quantitative modeling. Thus, we proceed to discuss our computational results.

Computational Modelling

The main goal of our study was to account for switch costs in the antisaccade task using the SERIA model. The conclusions presented here are based on quantitative Bayesian model comparison, as well as on qualitative posterior predictive fits (Fig. 6 and 7). Our results indicate that the bulk of effects of the last trial can be accounted for by alternations of trial type, and not by the previous actions. Note, however, that we cannot rule out interactions between trial type and actions that have been previously reported (Tatler and Hutton, 2007; DeSimone et al., 2014).

Qualitative inspection of the predictive fits (Fig. 6 and 7) clearly indicates that accounting for switch costs in both habitual and goal-directed actions improved the fit of the RT distributions. The second best

model ($\Delta LME = 8.2$) accounted for switch costs only through changes in inhibitory control (*switch:inhib.*). Examination of the predictive fits clearly shows why this model was inferior: The RT distribution of late responses could not be properly fitted.

The analysis of parameter estimates led to two conclusions. First, there was a positive switch cost in late pro- and antisaccades in terms of RT and ER. However, the switch cost associated with late errors on antisaccade trials was small (1%). Second, antisaccade trials led to fewer and faster habitual responses on the following trial. Therefore, we conclude that goal-directed behavior is facilitated by the application of the same rule over trials. Moreover, the enhanced inhibitory control induced by antisaccade trials led to more efficient inhibition on the following trial.

It is not straightforward to compare our model-based findings to predictions from conceptual models, because these do not always translate into precise empirical predictions. In particular, our results support two theories that are considered opposites (Weiler and Heath, 2014). On the one hand, the switch cost observed in goal-directed behavior supports the task-set inertia hypothesis (Weiler et al., 2014), according to which the activation of a cue-action mapping facilitates the activation of the same rule on the next trial, while interfering with other mappings.

On the other hand, our results also support the oculomotor inhibition hypothesis (Allport et al., 1994), according to which the type of inhibitory control required to execute an antisaccade leads to enhanced inhibition on the following trial, as reflected by a fewer inhibition failures following an antisaccade trial, as well as faster early prosaccades. The latter is a natural consequence of more efficient inhibitory control, which allows only inhibition failures with very short latencies. Note that the existence of inter-trial effects on inhibitory control is supported by evidence in the countermanding saccade task (e.g., Barton et al., 2006a; Weiler and Heath, 2014). In summary, our model supports both theories, depending on whether the interpretation focusses on the inhibition of reactive responses or on the generation of voluntary actions.

Finally, we note that our modeling approach cannot distinguish whether antisaccade trials enhance inhibitory control, or whether prosaccade

trials disengage it. Here, we have opted to assert that antisaccade trials enhance inhibition; however, in the present context, both options are undistinguishable.

Other models of inter-trial effects

To our knowledge, this is the first computational, trial-by-trial model used to investigate switch costs in the antisaccade task. Although SERIA seems to accommodate the most salient features of our data, one conceptual limitation of our approach is its phenomenological or descriptive nature. For instance, SERIA does not postulate any specific neurophysiological mechanism to explain the different types of switch costs. This role could be better filled by biologically inspired models (Emeric et al., 2007; Pouget et al., 2011), that can be compared to behavioral data as well as to physiological recordings.

A further limitation is that learning effects were not modeled explicitly. Because no information about the trial type probability was provided to participants, the PP effects reported here and by others (Cutsuridis et al., 2007; Heinzle et al., 2007; Wiecki and Frank, 2013; Lo and Wang, 2016) can only be explained by adaptation to statistical regularities in the stimuli. Here, we made the assumption that subjects quickly reach a steady state such that dynamic effects can be ignored. However, since the pioneering work of Carpenter and Williams (1995), it has been hypothesized that changes in saccadic RT can be explained by the principles of optimal statistical inference (assuming, for example, that humans act like ideal Bayesian observers; Brodersen et al., 2008). More recently, Vossel et al. (2014) (but see also Anderson and Carpenter, 2006; Brodersen et al., 2008) used a Bayesian model (Mathys et al., 2011) to demonstrate trial-by-trial variations in saccadic RT in a modified Posner paradigm, in which the average validity of the cue changed over time. In future work we plan to account for learning effects, in addition to the last trial effects modeled here.

Summary

Our quantitative modeling suggests that conceptual theories of switch costs in the antisaccade task can profit from a more precise formulation in computational terms, as seemingly contradictory statements can be correct at the same time. This is possible because of the non-trivial interactions between a habitual response mechanism that can be subject to fast inhibitory control and the ability to generate goal-directed,

context sensitive behavior. Concretely, our analysis indicates that alternating between goal-directed behaviors engenders sizeable switch costs, whereas increased inhibitory demands on one trial enhance inhibitory control on the following trial.

References

Allport A, Styles EA, Hsieh S (1994) Shifting Intentional Set: Exploring the Dynamic Control of Tasks. In: Attention and performance series. Attention and performance 15: Conscious and nonconscious information processing (Umilta C, Moscovitch M, eds), pp 266–290. Cambridge, MA, US: The MIT press.

Anderson AJ, Carpenter RH (2006) Changes in expectation consequent on experience, modeled by a simple, forgetful neural circuit. J Vis 6:822–835.

Ansari TL, Derakshan N, Richards A (2008) Effects of anxiety on task switching: evidence from the mixed antisaccade task. Cogn Affect Behav Neurosci 8:229–238.

Aponte EA, Schobi D, Stephan KE, Heinzle J (2017) The Stochastic Early Reaction, Inhibition, and late Action (SERIA) model for antisaccades. PLoS Comput Biol 13:e1005692.

Barton JJ, Cherkasova MV, Lindgren K, Goff DC, Intriligator JM, Manoach DS (2002) Antisaccades and task switching: studies of control processes in saccadic function in normal subjects and schizophrenic patients. Ann N Y Acad Sci 956:250–263.

Barton JJ, Greenzang C, Hefter R, Edelman J, Manoach DS (2006a) Switching, plasticity, and prediction in a saccadic task-switch paradigm. Exp Brain Res 168:76–87.

Barton JJ, Raoof M, Jameel O, Manoach DS (2006b) Task-switching with antisaccades versus no-go trials: a comparison of inter-trial effects. Exp Brain Res 172:114–119.

Bojko A, Kramer AF, Peterson MS (2004) Age equivalence in switch costs for prosaccade and antisaccade tasks. Psychol Aging 19:226–234.

Brodersen KH, Penny WD, Harrison LM, Daunizeau J, Ruff CC, Duzel E, Friston KJ, Stephan KE (2008) Integrated Bayesian models of learning and decision making for saccadic eye movements. Neural Netw 21:1247–1260.

Camalier CR, Gotler A, Murthy A, Thompson KG, Logan GD, Palmeri TJ, Schall JD (2007) Dynamics of saccade target selection: race model analysis of double step and search step saccade production in human and macaque. Vision Res 47:2187–2211.

Carpenter RH, Williams ML (1995) Neural computation of log likelihood in control of saccadic eye movements. Nature 377:59–62.

Chan JL, Koval MJ, Johnston K, Everling S (2017) Neural correlates for task switching in the macaque superior colliculus. J Neurophysiol 118:2156–2170.

Cherkasova MV, Manoach DS, Intriligator JM, Barton JJ (2002) Antisaccades and task-switching: interactions in controlled processing. Exp Brain Res 144:528–537.

Cutsuridis V, Smyrnis N, Evdokimidis I, Perantonis S (2007) A neural network model of decision making in an antisaccade task by the superior colliculus. Neural Networks 20:690–704.

DeSimone JC, Weiler J, Aber GS, Heath M (2014) The unidirectional prosaccade switch-cost: correct and error antisaccades differentially influence the planning times for subsequent prosaccades. Vision Res 96:17–24.

Emeric EE, Brown JW, Boucher L, Carpenter RH, Hanes DP, Harris R, Logan GD, Mashru RN, Pare M, Pouget P, Stuphorn V, Taylor TL, Schall JD (2007) Influence of history on saccade countermanding performance in humans and macaque monkeys. Vision Res 47:35–49.

Fecteau JH, Au C, Armstrong IT, Munoz DP (2004) Sensory biases produce alternation advantage found in sequential saccadic eye movement tasks. Exp Brain Res 159:84–91.

Greenwood TA et al. (2007) Initial heritability analyses of endophenotypic measures for schizophrenia: the consortium on the genetics of schizophrenia. Arch Gen Psychiatry 64:1242–1250.

Hallett PE (1978) Primary and secondary saccades to goals defined by instructions. Vision Res 18:1279–1296.

Heath M, Gillen C, Samani A (2016) Alternating between pro- and antisaccades: switch-costs manifest via decoupling the spatial relations between stimulus and response. Exp Brain Res 234:853–865.

Heath M, Starrs F, Macpherson E, Weiler J (2015) Task-switching effects for visual and auditory pro- and antisaccades: evidence for a task-set inertia. J Mot Behav 47:319–327.

Heinzle J, Aponte EA, Stephan KE (2016) Computational models of eye movements and their application to schizophrenia. Current Opinion in Behavioral Sciences 11:21–29.

Heinzle J, Hepp K, Martin KA (2007) A microcircuit model of the frontal eye fields. J Neurosci 27:9341–9353.

Isoda M, Hikosaka O (2008) Role for subthalamic nucleus neurons in switching from automatic to controlled eye movement. J Neurosci 28:7209–7218.

Karayanidis F, Jamadar S, Ruge H, Phillips N, Heathcote A, Forstmann BU (2010) Advance preparation in task-switching: converging evidence from

behavioral, brain activation, and model-based approaches. Front Psychol 1:25.

Kass RE, Raftery AE (1995) Bayes factors. Journal of the american statistical association 90:773–795.

Lee AK, Hamalainen MS, Dyckman KA, Barton JJ, Manoach DS (2011) Saccadic preparation in the frontal eye field is modulated by distinct trial history effects as revealed by magnetoencephalography. Cereb Cortex 21:245–253.

Lo CC, Wang XJ (2016) Conflict Resolution as Near-Threshold Decision-Making: A Spiking Neural Circuit Model with Two-Stage Competition for Antisaccadic Task. PLoS Comput Biol 12:e1005081.

Logan GD, Cowan WB, Davis KA (1984) On the ability to inhibit simple and choice reaction time responses: a model and a method. J Exp Psychol Hum Percept Perform 10:276–291.

Manoach DS, Lindgren KA, Cherkasova MV, Goff DC, Halpern EF, Intriligator J, Barton JJ (2002) Schizophrenic subjects show deficient inhibition but intact task switching on saccadic tasks. Biol Psychiatry 51:816–826.

Mathys C, Daunizeau J, Friston KJ, Stephan KE (2011) A bayesian foundation for individual learning under uncertainty. Front Hum Neurosci 5:39.

Monsell S (2003) Task switching. Trends Cogn Sci (Regul Ed) 7:134–140.

Mueller SC, Swainson R, Jackson GM (2009) ERP indices of persisting and current inhibitory control: a study of saccadic task switching. Neuroimage 45:191–197.

Munoz DP, Everling S (2004) Look away: the anti-saccade task and the voluntary control of eye movement. Nat Rev Neurosci 5:218–228.

Myles JB, Rossell SL, Phillipou A, Thomas E, Gurvich C (2017) Insights to the schizophrenia continuum: A systematic review of saccadic eye movements in schizotypy and biological relatives of schizophrenia patients. Neurosci Biobehav Rev 72:278–300.

Noorani I, Carpenter RH (2013) Antisaccades as decisions: LATER model predicts latency distributions and error responses. Eur J Neurosci 37:330–338.

Peirce JW (2007) PsychoPy--Psychophysics software in Python. J Neurosci Methods 162:8–13.

Peirce JW (2008) Generating Stimuli for Neuroscience Using PsychoPy. Front Neuroinform 2:10.

Pierce JE, McCardel JB, McDowell JE (2015) Trial-type probability and task-switching effects on behavioral response characteristics in a mixed saccade task. Exp Brain Res 233:959–969.

Pouget P, Logan GD, Palmeri TJ, Boucher L, Pare M, Schall JD (2011) Neural basis of adaptive response time adjustment during saccade countermanding. J Neurosci 31:12604–12612.

Radant AD et al. (2015) Robust differences in antisaccade performance exist between COGS schizophrenia cases and controls regardless of recruitment strategies. Schizophr Res 163:47–52.

Reilly JL, Frankovich K, Hill S, Gershon ES, Keefe RS, Keshavan MS, Pearlson GD, Tamminga CA, Sweeney JA (2014) Elevated antisaccade error rate as an intermediate phenotype for psychosis across diagnostic categories. Schizophr Bull 40:1011–1021.

Sato TR, Schall JD (2003) Effects of stimulus-response compatibility on neural selection in frontal eye field. Neuron 38:637–648.

Stampe D (1993) Heuristic filtering and reliable calibration methods for video-based pupil-tracking systems. Behavior Research Methods, Instruments, \& Computers 25:137–142.

Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009) Bayesian model selection for group studies. Neuroimage 46:1004–1017.

Tatler BW, Hutton SB (2007) Trial by trial effects in the antisaccade task. Exp Brain Res 179:387–396.

Vossel S, Mathys C, Daunizeau J, Bauer M, Driver J, Friston KJ, Stephan KE (2014) Spatial attention, precision, and Bayesian inference: a study of saccadic response speed. Cereb Cortex 24:1436–1450.

Weiler J, Hassall CD, Krigolson OE, Heath M (2015) The unidirectional prosaccade switch-cost: electroencephalographic evidence of task-set inertia in oculomotor control. Behav Brain Res 278:323–329.

Weiler J, Heath M (2012a) The prior-antisaccade effect influences the planning and online control of prosaccades. Exp Brain Res 216:545–552.

Weiler J, Heath M (2012b) Task-switching in oculomotor control: unidirectional switch-cost when alternating between pro- and antisaccades. Neurosci Lett 530:150–154.

Weiler J, Heath M (2014) Oculomotor task switching: alternating from a nonstandard to a standard response yields the unidirectional prosaccade switch-cost. J Neurophysiol 112:2176–2184.

Weiler J, Mitchell T, Heath M (2014) Response suppression delays the planning of subsequent stimulus-driven saccades. PLoS ONE 9:e86408.

Wiecki TV, Frank MJ (2013) A computational model of inhibitory control in frontal cortex and basal ganglia. Psychol Rev 120:329–355.

Chapter 9

In the previous chapters, the SERIA model for the antisaccade task was developed, and several experimental findings were reported. For example, we showed that corrective antisaccades that follow an error are likely to be delayed late responses. In this chapter, we use SERIA to answer the following question: How do pro-cholinergic and pro-dopaminergic drugs affect the voluntary control of eye movements in the antisaccade task? These two types of drugs have been investigated in some detail in the literature, although the particular compounds used here (galantamine and levodopa) have either not being investigated before, or only in very different experimental setups (as explained later in this chapter). Their importance lies in that both compounds affect neurotransmitters that are directly or indirectly affected by antipsychotic medication. Moreover, while it has been suggested that galantamine, a cognitive enhancer approved for palliative treatment of Alzheimer's disease, could be of use in treating cognitive deficits in schizophrenia, recent clinical studies have not shown any significant therapeutic value, as discussed in detail later on. Thus, it is of interest to understand the effects of both types of compounds in the antisaccade task, and their relationship to the deficits observed in schizophrenia.

Here, we present two experiments that follow a well-powered, double-blind, placebo-controlled, within subject design. Our results indicate that the compounds used here have opposite effects on voluntary eye movements. This is shown in both summary statistics of RTs and ERs, as well as in the parameter estimates of SERIA. Moreover, SERIA offers an interpretable explanation of our empirical findings, and points to possible mechanistic explanations.

This last chapter incorporates the methodological ideas developed in Part I and II of this dissertation and offers a first product of the research agenda laid out in Chapter 4. In the final section of this dissertation we offer an outlook of the open questions and problems for future research.

Dopaminergic and cholinergic effects on goal-directed behavior

Eduardo A. Aponte[1], *Dario Schöbi*[1], *Klaas E. Stephan*[1,2], *Jakob Heinzle*[1]

[1] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich. Wilfriedstrasse 6, 8004, Zurich, Switzerland.

[2] Wellcome Centre for Neuroimaging, University College London. 12 Queen Square, London WC1N 3BG, UK.

[*]**Corresponding authors:**

aponte@biomed.ee.ethz.ch, heinzle@biomed.ee.ethz.ch

Abstract

The ability to stop habitual responses and to simultaneously initiate rule guided actions is pivotal to adaptive behavioral control. To study how dopamine (DA) and acetylcholine (ACh) affect this ability under environmental uncertainty, we investigated two agonist compounds (galantamine 8mg and levodopa 200mg) in humans (N=90) in the context of the antisaccade task. Using a recent computational model of antisaccade latencies, we found that while ACh reduced the latency of voluntary actions, higher systemic DA had the opposite effect. Although inhibitory control was affected by environmental uncertainty, neither of these compounds had an impact on inhibition, nor did they interact with the effect of uncertainty. Crucially, a classifier was able to predict the compound (DA vs. Ach) with an accuracy of 70%. Our results contribute to further the current understanding of the role of DA and ACh in decision making.

Introduction

Higher-order cognitive control often requires the interplay of different, competitive decision processes. For example, when confronted with situations that demand a rapid but thoughtful change of plans, it is not enough to simply stop our initial course of action, it is also imperative to select a secondary plan that copes with changing circumstances. Interestingly, these two decision processes might recruit different neuronal circuits (Hikosaka and Isoda, 2010; Isoda and Hikosaka, 2011; 2008) that are differently affected by the neuromodulators acetylcholine (ACh) and dopamine (DA). The dysregulation of these neurotransmitters is at the heart of disorders such as Parkinson's disease and schizophrenia.

An experimental paradigm that has been extensively used to investigate the interplay of these mechanisms is the antisaccade task (Hallett, 1978), in which subjects are instructed to saccade in the opposite direction of a cued location. Because the most common response to a suddenly presented stimulus is a ballistic eye movement towards it (a prosaccade), the execution of a saccade in the opposite direction (an antisaccade) requires both the inhibition of an automatic, habitual response, and the start of a secondary, voluntary action (Everling and Johnston, 2013; Reuter et al., 2007). In addition to inhibitory control and conflict resolution in human and non-human primates (Munoz and Everling, 2004), the antisaccade task has been extensively used to study clinical populations (Hutton and Ettinger, 2006; Myles et al., 2017). Arguably, the most remarkable and consistent finding in this domain is that patients diagnosed with schizophrenia as well as their first order unaffected relatives display high error rates (ER) and reaction times (RT) in this paradigm (Myles et al., 2017; Radant et al., 2015; Reilly et al., 2014). These findings indicate that deficits in the antisaccade task constitute a genuine biomarker of schizophrenia.

It has also been proposed that errors caused by different decision processes could discriminate different psychiatric and neurological diseases (Coe and Munoz, 2017; Heinzle et al., 2016). Unfortunately, only few studies (Reuter et al., 2007; 2005) have considered the question whether significant changes in ER and RT can be related to one or both putative decision processes that are thought to underlie the execution of this task.

Arguably, one reason why this question has not been systematically addressed is the lack of computational models that could disentangle the decision processes involved in the generation of antisaccades (Heinzle et al., 2016). Such models have only emerged in recent years (Aponte et al., 2017; Camalier et al., 2007; Cutsuridis et al., 2007; Heinzle et al., 2007; Lo and X. J. Wang, 2016; Noorani and Carpenter, 2013; Wiecki and Frank, 2013) and with them more precise predictions of the effect of different changes in neuromodulation associated with psychiatric diseases have been made. For example, (Wiecki and Frank, 2013) suggested that increased tonic DA in the basal ganglia impairs inhibitory control in the antisaccade task. This hypothesis is based on the observation that schizophrenia is associated with the dysregulation of DA as well as with higher ERs in the antisaccade task (Matsuda et al., 2004). Moreover, deficits could be partly explained by abnormalities in the inhibitory pathway from the basal ganglia to the motor layers of the superior colliculus, which is critically involved in saccadic eye movements (Hikosaka et al., 2000; Isoda and Hikosaka, 2008). Similarly, a large body of evidence points to a role of this structure in inhibitory control across different tasks (Aron and Poldrack, 2006; Schmidt and Berke, 2017). However, the location of the circuits responsible for the inhibition of prosaccades is still debated (although see Stuphorn and Schall, 2006). It has also been reported that lesions in the basal ganglia do not affect antisaccade performance (Condy et al., 2004), whereas prefrontal lesions are known to critically impair it (Guitton et al., 1985; Pierrot-Deseilligny et al., 1991). In this direction, Everling and Johnston (2013) have challenged the view that the elevated error rate in schizophrenia is caused by compromised inhibitory control, instead hypothesizing that cortical cue-action mapping might be affected in this condition.

In contrast to the hypothesized deleterious effect of high tonic DA, pro-cholinergic compounds targeting the $\alpha 7$ and $\alpha 4\beta 2$ nicotine receptors have been postulated as a possible treatment to the negative symptoms and cognitive impairments associated with schizophrenia (Arango et al., 2013; Buchanan et al., 2008; Lieberman et al., 2013), although recent studies did not reveal any or only limited therapeutic effects of these compounds (Conley et al., 2009; Lindenmayer and Khan, 2011; Umbricht et al., 2014; Walling et al., 2016). In a similar spirit, several studies (Bowling and Donnelly, 2010; Depatie et al., 2002; Ettinger et

al., 2009; Larrison et al., 2004; Larrison-Faucher et al., 2004; Petrovsky et al., 2012; 2013; Powell et al., 2002; Roos et al., 1995; Rycroft et al., 2006; 2007) have investigated whether nicotine has a positive effect on the antisaccade task. These reports indicate that that nicotine decreases antisaccade RT (Bowling and Donnelly, 2010; Ettinger et al., 2009; Petrovsky et al., 2013; Rycroft et al., 2007; 2006), in line with a meta-analysis of the effect of nicotine on a wide variety of cognitive paradigms (Heishman et al., 2010). However, not all studies have replicated this result (for example Ettinger et al., 2017). Similarly, several studies have found that nicotine reduces ER (Depatie et al., 2002; Petrovsky et al., 2013; 2012; Rycroft et al., 2006; Schmechtig et al., 2013) but others have not found a significant effect (Rycroft et al., 2007). Thus, it is not clear whether pro-cholinergic and pro-dopaminergic drugs have opposite effects in the antisaccade task, and, if so, whether these effects are reflected in either inhibitory control or the initiation of voluntary actions involved in this paradigm.

Here, we used a recent computational model of the antisaccade task (Aponte et al., 2017) to investigate the effects of a pro-cholinergic and a pro-dopaminergic compound in healthy volunteers in two well powered (Ettinger et al., 2017; Heishman et al., 2010), comparable experiments (Experiment 1: DA, N=46, Experiment 2: ACh, N=44). Subjects were instructed to saccade either to a cued location (prosaccade trial), or in the opposite direction (antisaccade trial) depending on the orientation of a visual, peripheral cue (Fig. 1A). Both trial types were presented in three mixed blocks with different prosaccade trial probability (20, 50, or 80%). In Exp. 1, participants received either placebo or levodopa (200mg), and in Exp. 2 placebo or galantamine (8mg) in two different sessions, according to a double-blind, placebo-controlled, crossover protocol. Levodopa (l-dopa) is a precursor of DA that crosses the blood-brain barrier and increases the systemic availability of this neurotransmitter. Galantamine is a weak inhibitor of acetylcholinesterase, an enzyme involved in the hydrolysis of ACh at the synaptic clef, as well as an allosteric potentiating ligand of the α7 (Texido et al., 2005) and α4β2 ACh nicotinic receptors (nAChR; Maelicke et al., 2001; Samochocki et al., 2003; Santos et al., 2002).

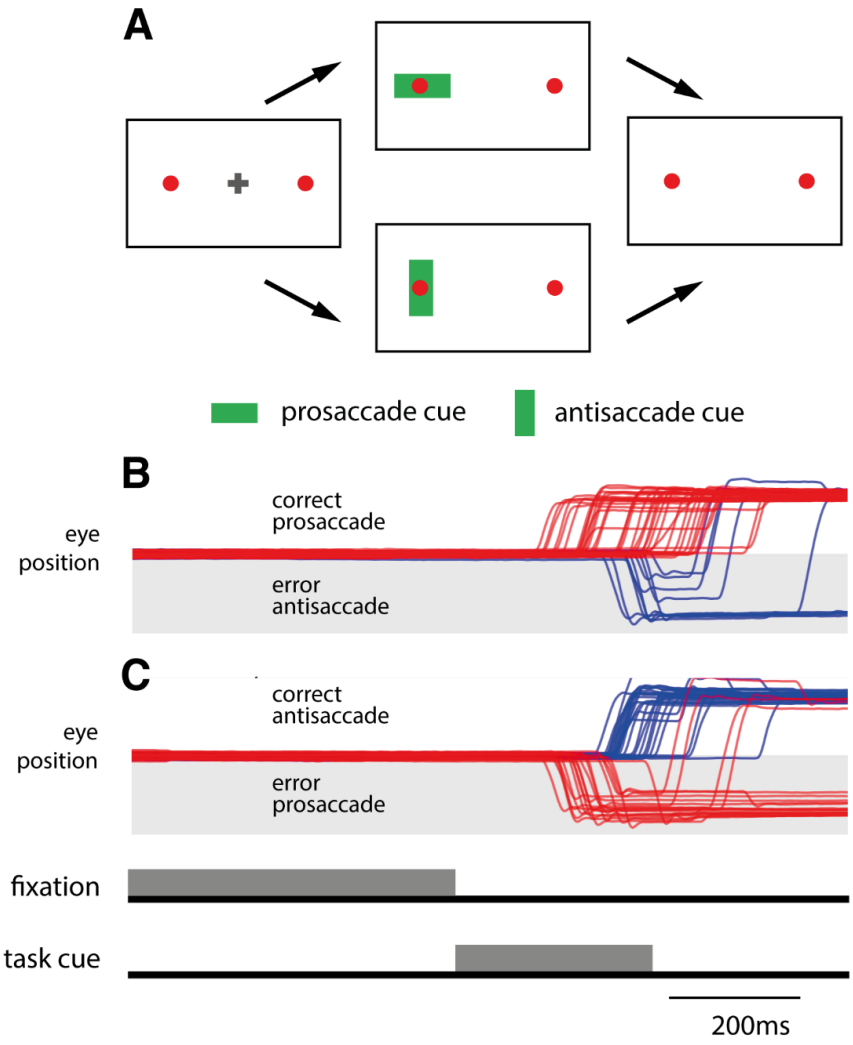


Figure 1: A. Task design. Subjects were required to initially fixate to a centrally presented cross for 500 to 1000ms. Two red dots were constantly displayed at ±12° indicating the possible locations of the cue. After the initial fixation period, the central cross was removed and a cue (green bar 3.5°) was presented either in horizontal or vertical orientation for 500ms. Subjects were instructed to saccade in the direction cued by horizontal cue (prosaccade trial) or the opposite direction in case of a vertical cue (antisaccade trial). **B** Exemplary eye position trace in the horizontal plane in prosaccade trials. Correct prosaccades distribution extended from as early as 200ms to as much as 450ms after cue presentation. Importantly, several errors (antisaccades) can be observed. **C.** Exemplary horizontal eye position in antisaccade trials. Eye position in correct antisaccades is displayed in blue, whereas incorrect prosaccades are displayed in red. Errors latency was as short as 200ms from cue presentation.

In this study, our main interests were two: First, we aimed at investigating the possibly opposing effects of levodopa and galantamine

on either ER and RT, and whether these effects could be imputed to any of the decision processes involved in the antisaccade task. For this, we first analyzed average ER and RT and then proceeded to apply a computational model to these data. Second, we used a classification algorithm to test whether eye movement data could be used to disambiguate the type of drug given to the participants.

Results

Participants

In Exp. 1, 46 male subjects (23.6 mean years of age, std. 2.9, range $19-33$) were included in the final analysis. Four participants were excluded because their data was incomplete. No subject demonstrated an overt negative reaction to levodopa or placebo. Participants could not guess the substance administered significantly better than chance after the first ($\chi^2(1, N = 44) = 3.2, p = 0.07$) or second session ($\chi^2(1, N = 42) = 0.9, p = 0.75$). In total, 52993 trials were analyzed, from which 1303 trials were excluded (Supp. Table S1).

In Exp. 2, 44 subjects were included in the final analysis (mean age 22.4, std. 2.3, range $18-29$). Four subjects reported a negative reaction to the substance administered and were excluded from the study. Two subjects were excluded because their data was incomplete. Again, participants were not significantly better than chance at discriminating galantamine from placebo (first session $\chi^2(1, N = 44) = 0.8, p = 0.54$; second session $\chi^2(1, N = 44) = 0.8, p = 0.36$). A total of 50688 trials were analyzed, from which 1416 were excluded, mostly due to blinks (Supp. Table S1).

Error rate and reaction times

All variables of interest were analyzed with a generalized, mixed-effects linear model (GLME). Both experiments were first included into a single GLME, and then each experiment was considered separately for the factor DRUG (see below) and its interactions. The fixed effects entered into the model were the categorical variables EXPERIMENT (EXP) with levels *Exp. 1* and *2*, DRUG with levels *drug* and *placebo*, SESSION, *prosaccade trial probability* (PP; levels PP20, PP50 and PP80), and the continuous factor DOSE (mg/kg). Moreover, we divided trials into switch and repeat trials, to account for switch costs as shown in Chapter 8. In addition, we included several interactions as detailed in the methods section. Our main interest were the interactions between the factors DRUG and EXP, DRUG and PP, and DRUG and DOSE. The factor SUBJECT was entered as a random effect.

Error rate

The mean ER in pro- and antisaccade trials is displayed in Fig. 2 and Supp. Table S2. Higher congruent trial type probability was associated

with fewer errors in pro- ($\chi^2(2, N = 1080) = 871.7, p < 10^{-5}$), and antisaccade trials ($\chi^2(2, N = 1080) = 901.2, p < 10^{-5}$). Subjects made fewer errors in the second session (pro.: $\chi^2(1, N = 1080) = 13.9, p = 10^{-3}$, anti.: $\chi^2(1, N = 1080) = 7.6, p = 0.005$). There were less errors in prosaccade trials in Exp. 1 compared to Exp. 2 ($\chi^2(2, N = 1080) = 6.1, p = 0.0129$).

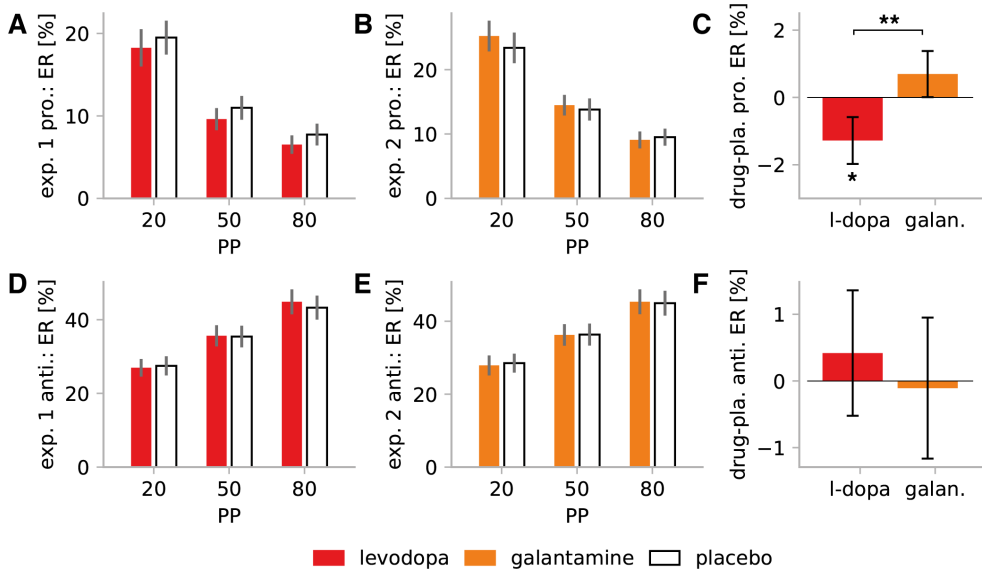


Figure 2: A. Mean prosaccade ER in Exp. 1 (levodopa vs. placebo). **B.** Mean antisaccade ER in Exp. 2 (galantamine vs. placebo). **C.** Difference in ER between the drug and placebo condition in Exp. 1 and 2. There was a significant difference between the effect of levodopa and galantamine ($\chi^2(1, N = 1080) = 6.8, p = 0.008$). When analyzed separately, levodopa significantly reduced prosaccade ER ($\chi^2(1, N = 552) = 8.5, p = 0.010$). **D.** Mean prosaccade ER in Exp. 1. **E.** Mean antisaccade ER in Exp. 2. **F.** Difference in ER in Exp. 1 and 2. No significant effect was found. Error bar depicts the sem.. PP: prosaccade trial probability; $*: p < 0.05$; $**: p < 0.01$.

Error rate: Drug effects

Regarding the effects of DRUG and its interactions, in prosaccade trials (Fig. 3A) we found a significant interaction between DRUG and EXP ($\chi^2(1, N = 1080) = 6.8, p = 0.008$). We proceeded to investigate Exp. 1 and 2 independently for drug related effects. In prosaccade trials in Exp. 1, we found a significant main effect of DRUG ($\chi^2(1, N = 552) = 8.5, p = 0.010$) that demonstrated that levodopa reduced the number of error in prosaccade trials, and this effect was DOSE dependent ($\chi^2(1, N = 552) = 8.0, p = 0.004$). In antisaccade trials, we found no significant

effect of drug. Regarding Exp. 2, we found a significant interaction between DRUG and DOSE (Fig. 3B; $\chi^2(1, N = 528) = 29.6, p = 10^{-5}$) in antisaccade trials. Galantamine had no significant effect on ER in prosaccade trials. Qualitative inspection of the ER in Exp. 2 (Fig. 3B), revealed that galantamine increased the number of errors at a high dose, while it reduced them at more moderate levels.

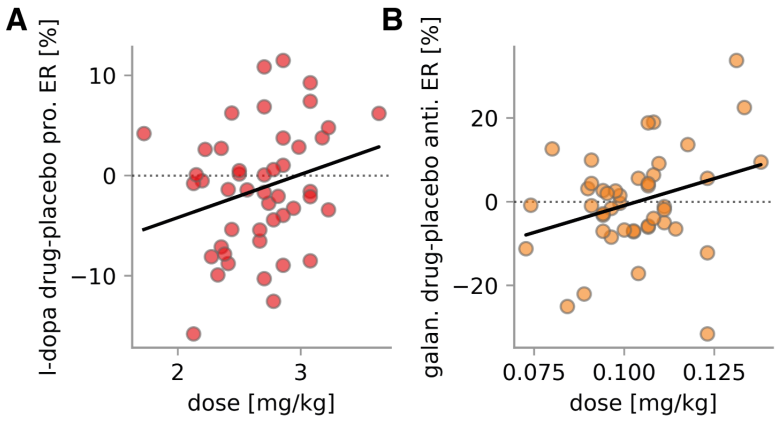


Figure 3:. **A.** Difference in ER between the drug and placebo conditions in prosaccade trials and the linear fit to dose in Exp. 1 ($\chi^2(1, N = 552) = 8.0, p = 0.004$). **B.** Difference in ER between the drug and placebo conditions in antisaccade trials and the linear fit to dose in Exp. 2 ($\chi^2(1, N = 528) = 29.6, p = 10^{-5}$). Linear fits are displayed for interpretability and do not correspond to the estimated effects according to a GLME.

Reaction time

Reaction times (RT) in correct trials (Fig. 4 and Sup. Table S3) were analyzed similarly to ER. Higher trial type probability led to lower RT in both pro- ($F_{2,972} = 30.3, p < 10^{-5}$) and antisaccade trials ($F_{2,968} = 35.1, p < 10^{-5}$). RTs were lower in the second session (pro: $F_{1,972} = 31.5, p < 10^{-5}$, anti: $F_{1,968} = 75.0, p < 10^{-5}$). Moreover, RT in Exp. 2 were significantly lower than in Exp. 1 (pro: $F_{1,86} = 8.4, p = 0.004$; anti: $F_{1,86} = 12.5, p < 10^{-3}$).

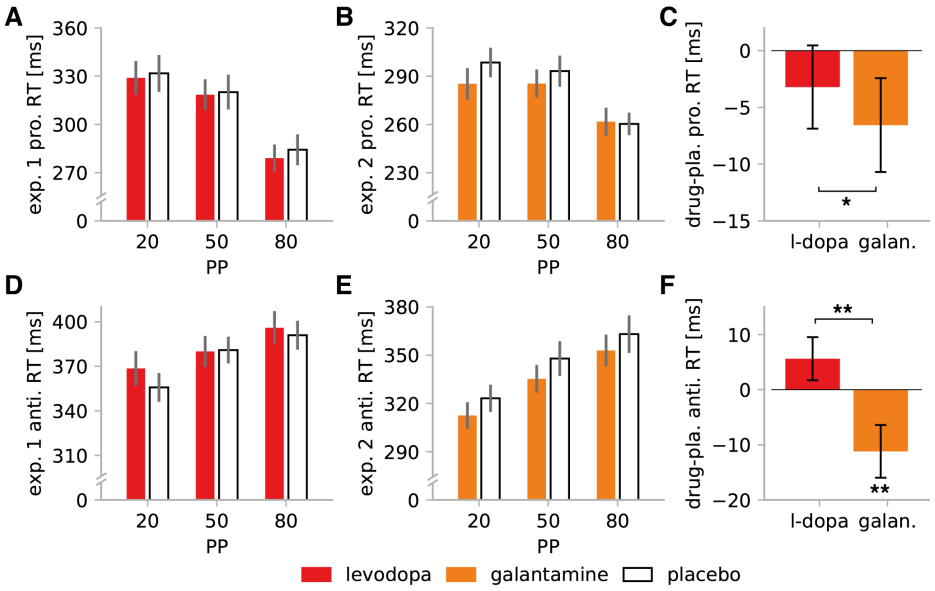


Figure 4: Only RT of correct trials are shown. **A.** Mean prosaccade RT in Exp. 1 (levodopa vs. placebo). **B.** Mean antisaccade RT in Exp. 1. **C.** Difference in RT between the drug and placebo conditions on prosaccade trials. The main effect of DRUG was significant ($F_{2,972} = 4.0, p = 0.022$). **D.** Mean antisaccade RT in Exp. 2 (galantamine vs. placebo). **E.** Mean antisaccade RT in Exp. 2. **F.** Difference in RT on antisaccade trials. Levodopa increased the RT of antisaccades when compared to galantamine ($F_{2,968} = 12.2, p = 10^{-3}$). Error bars depict the sem. PP: prosaccade trial probability; $*: p < 0.05$; $**: p < 0.01$.

Reaction times: Drug effects

Regarding the DRUG factor, we found a significant interaction between DRUG and EXP ($F_{2,968} = 12.2, p = 10^{-3}$) in antisaccades trials (Fig. 4F) and an effect of DRUG in prosaccade trials (Fig. 4C; $F_{2,972} = 4.0, p = 0.022$). When the two experiments were analyzed independently for an effect of DRUG, we found that galantamine decreased antisaccade RT ($F_{2,250} = 11.1, p = 10^{-3}$).

Modeling

The *Stochastic Early Response, Inhibition and late Action* (SERIA) model (Aponte et al., 2017) was used to fit actions (pro- and antisaccades) and RTs. SERIA models two different race processes involved in the generation of an antisaccade: An initial GO/NO-GO competition between an early and an inhibitory unit (U_e and U_i), and a second race between a late pro- and an antisaccade unit (U_p and U_a). Accordingly, two types of errors in an antisaccade trial are postulated by the model:

inhibition failures that occur when an early prosaccade reaction is not stopped, and late errors that occur when the late prosaccade unit hits threshold before the antisaccade unit. SERIA explains the commonly observed bimodality of prosaccade RTs by the same mechanism: early prosaccades are actions that are not inhibited (inhibition failures), whereas voluntary or late prosaccades are generated by the same decision process responsible for antisaccades. The model accounted for differences across pro- and antisaccade trials, and switch and repeat trials in the same manner as in Chapter 8 and as elaborated in the Methods section.

We were interested in several parameters derived from the model (see Methods). First, we considered the expected hit time of the early, inhibitory, late pro-, and antisaccade units. Furthermore, we investigated the probability of an inhibition failure, i.e., the probability that the early unit hit threshold before all other units, and the probability late errors in pro- and antisaccade trials.

The estimated posterior distribution of RT and actions of exemplary subjects are display in Supp. Fig. S4. The fits of the entire data set (aggregated across subjects) are displayed in Fig. 5. Posterior fits demonstrate that SERIA captures the data well. Concretely, it reproduces well the shape of the distribution of RT including the clear bimodality of prosaccade responses.

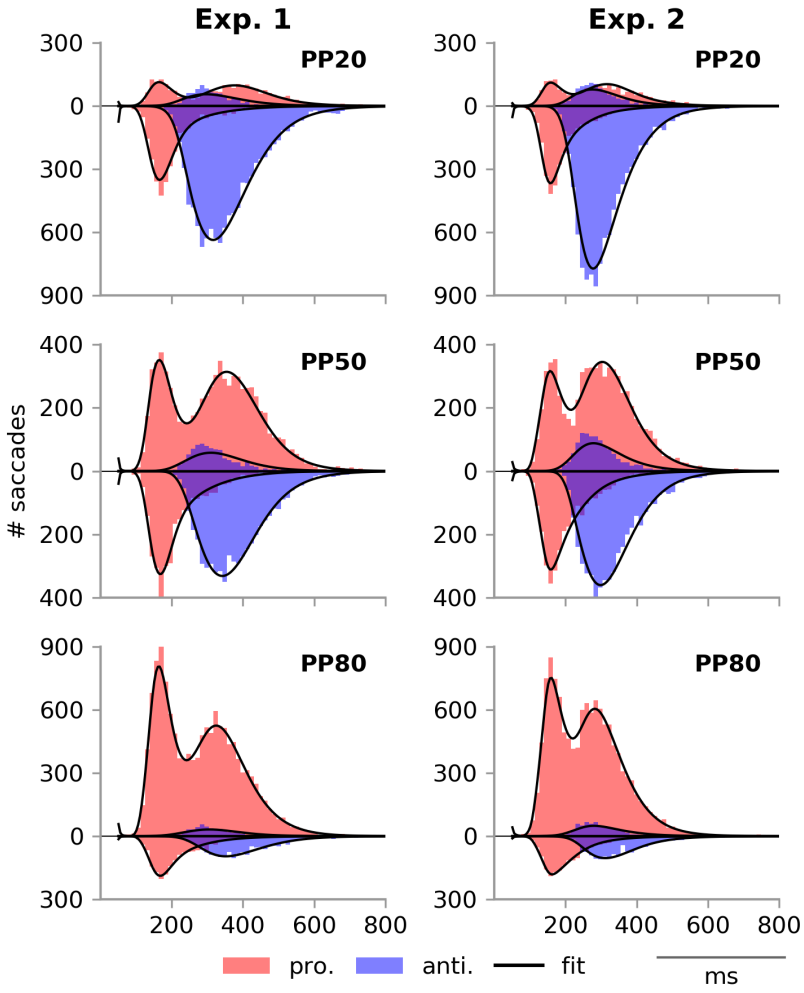


Figure 5: Mean fits of the model in the three PP conditions. The histogram of the prosaccades and antisaccades is displayed in red and blue respectively. Prosaccade trials are shown in upper (positive) part of the plane; antisaccades in the bottom (negative) part. To display the model fits, the predictive likelihood of each trial type was computed for each subject and normalized according to the number of trials per condition.

We proceeded to investigate the metrics derived from the model fits. Our main interest was to determine whether the arrival times of any of the units or the probability of an uninhibited early saccade and the probability of a late antisaccade changed when levodopa or galantamine were administered.

Threshold hit times

The threshold hit times of the early, the inhibitory, and late pro- and antisaccades units were analyzed as in the previous section. For the late

units, we only report the hit times in correct trials. In the case of the inhibitory unit, pro- and antisaccade trials were analyzed together by including the factor TRIAL-TYPE (TT). In the case of the early unit, we previously established through model comparison (see Chapter 8) that its parameters can be fixed across trial types and switch and repeat trials. Thus, we did not include the TT and SWITCH factors.

Prosaccade trial probability (PP) had a significant effect on all units (see Supp. Tables S5-8). In agreement with Chapter 7 and 8, we found that the arrival time of the inhibitory unit increased with higher prosaccade trial type probability ($F_{1,2047} = 242.7, p < 10^{-5}$), indicating reduced inhibition of early responses. On average, hit times were lower in the second session (Supp. Table S5-8). Furthermore, we found a main effect of the factor EXP in the response time of the late units (late pro.: $F_{1,86} = 18.5, p < 10^{-3}$, anti: $F_{1,86} = 15.4, p < 10^{-3}$). However, we found no effect of EXP in the early ($F_{1,86} = 0.5, p = 0.461$) or inhibitory units ($F_{1,86} = 0.0, p = 0.771$).

Response times: Drug effects

The interaction between DRUG and EXP. (Fig. 6, left and center) was significant in the case of the late pro- ($F_{1,972} = 19.4, p < 10^{-3}$) and antisaccade units ($F_{1,972} = 15.1, p < 10^{-3}$). Levodopa increased response times, while galantamine had the opposite effect. The interactions between DRUG and PP, or DRUG, PP and EXP were not significant.

We proceeded to investigate each experiment independently for DRUG effects. In Exp. 1, the increase in the hit time of late pro- ($F_{1,494} = 8.2, p = 0.004$) and antisaccades ($F_{1,494} = 6.4, p = 0.011$) in the DRUG condition were significant. On average, the response time of late reactions was 5ms higher in the DRUG condition as compared to the placebo condition. In Exp. 2, we found that galantamine reduced the response time of late pro- ($F_{1,472} = 15.0, p < 10^{-3}$) and antisaccades ($F_{1,472} = 12.0, p < 10^{-3}$). On average, the response times were 6ms faster in the galantamine condition when compared to the placebo condition.

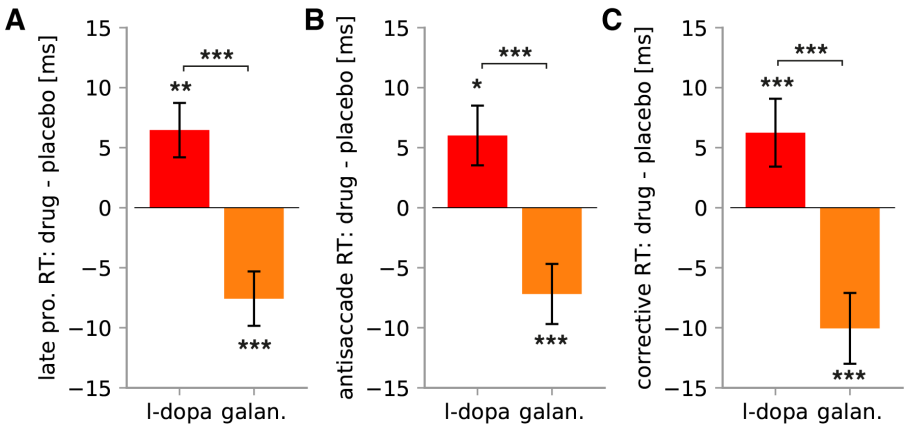


Figure 6: A. Difference in the RT of late prosaccades between drug and placebo conditions ($F_{1,972} = 19.4, p < 10^{-3}$). **B.** Difference in RT of antisaccades ($F_{1,972} = 15.1, p < 10^{-3}$). **C.** Difference in RT of corrective antisaccades ($F_{1,10620} = 24.4, p < 10^{-3}$). In each case, there was a significant difference between the effects of levodopa and galantamine. Error bars represent the sem.. $*: p < 0.05$; $**: p < 0.01$, $***: p < 0.001$.

Corrective antisaccades

In Chapters 7 and 8, we demonstrated that the latency of corrective antisaccades after a prosaccade error follow the distribution of late responses up to a fixed delay. Because the results above point to opposite effects of levodopa and galantamine on the latency of voluntary saccades, a prediction of SERIA is that corrective antisaccades should display the same drug effects as antisaccades, i.e., higher latency of corrective antisaccades in the levodopa condition, and lower latency in the galantamine condition.

We analyzed 5696 corrective saccades in Exp. 1 (levodopa: 2736, placebo 2960) and 4996 in Exp. 2 (galantamine: 2479, placebo: 2517). Because the frequency of corrective antisaccades varied widely over subjects and conditions, we accounted for the inhomogeneous number of trials by investigating trial-by-trial RT as opposed to mean RT. In general, the model accounted for a large fraction of the variance in the data (R^2=0.30). When analyzed together, the interaction between DRUG and EXP ($F_{1,10620} = 24.4, p < 10^{-3}$, Fig. 6 right) was significant. Considered independently, the effect of DRUG was significant in Exp. 1 ($F_{1,5654} = 10.8, p < 10^{-3}$) and 2 ($F_{1,4960} = 14.6, p < 10^{-3}$). Supporting

our hypothesis, levodopa increased the RT of corrective antisaccades ($\Delta RT = 8ms$), whereas galantamine had the opposite effect ($\Delta RT = -10ms$).

Inhibition failures and late errors

We proceeded to investigate the probability of late errors and inhibition failures, i.e., the probability that the early unit hit threshold before all other units (Supp. Table S9-11). PP had a significant effect on late pro- ($\chi^2(2, N = 1080) = 1634.6, p < 10^{-5}$) and antisaccade errors ($\chi^2(2, N = 1080) = 201.7, p < 10^{-5}$) and inhibition failures ($\chi^2(2, N = 2160) = 294.7, p < 10^{-5}$).

Drug effects: Inhibition failures and late errors

In the case of late errors in prosaccade trials, we found a significant interaction between DRUG and EXP ($\chi^2(1, N = 1080) = 24.0, p < 10^{-5}$; Fig. 7A). Regarding late errors in antisaccade trials (Fig. 7B), there was a significant main effect of DRUG ($\chi^2(2, N = 1080) = 19.6, p < 10^{-5}$). The main effect of DRUG ($\chi^2(2, N = 2160) = 4.8, p = 0.028$) was also significant in the case of inhibition failures (not displayed). Again, no interaction between DRUG and PP or DRUG, PP, and EXP was significant.

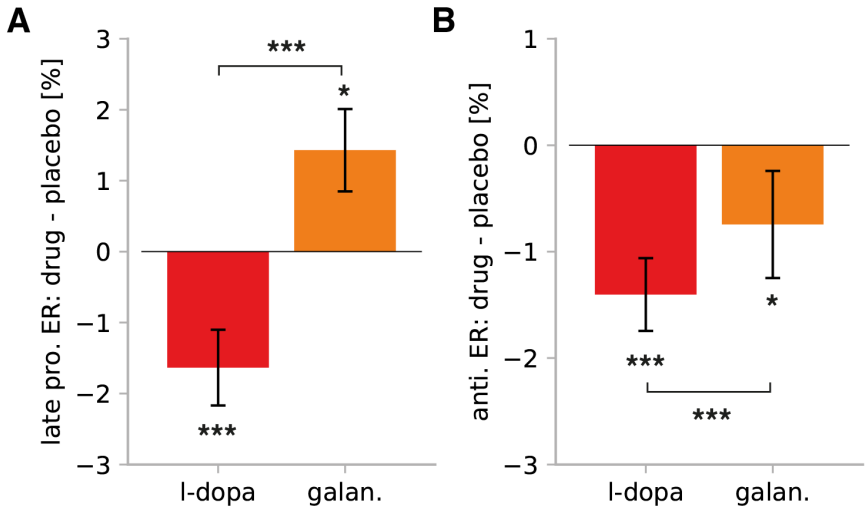


Figure 7: A. Difference in ER in late prosaccades between the drug and placebo conditions ($\chi^2(1, N = 1080) = 24.0, p < 10^{-5}$). **B.** Difference in error rate in antisaccade trials ($\chi^2(2, N = 1080) = 19.6, p < 10^{-3}$). Error bars represent the s.e.m. $*: p < 0.05$; $**: p < 0.01$, $***: p < 0.001$.

When we examined Exp. 1, we found a significant effect of DRUG ($\chi^2(1, N = 552) = 14.9, p < 10^{-3}$) in the probability of late errors in prosaccade trials indicating fewer late errors in the levodopa condition. In the case of late errors in antisaccade trials, we found a comparable effect of DRUG ($\chi^2(1, N = 552) = 17.8, p < 10^{-3}$). There was a small but significant effect of DRUG in the probability of an inhibition failure $\chi^2(1, N = 1104) = 3.9, p = 0.046$).

In Exp. 2, we found a main effect of DRUG in late errors in both pro ($\chi^2(2, N = 528) = 5.4, p = 0.019$) and antisaccade trials ($\chi^2(2, N = 528) = 4.19, p = 0.040$). Galantamine increased the probability of a late error in prosaccade trials and decreased them in antisaccade trials. There was no significant effect on the number of inhibition failures ($\chi^2(2, N = 1056) = 0.5, p = 0.457$).

Dose dependent effects

In addition to the main effect of levodopa and galantamine, we investigated dose dependent effects. In Exp. 1, we found a significant interaction between the factors DRUG and DOSE in the probability of a late error in prosaccade trials (Fig. 8A; pro. trials $\chi^2(1, N = 552) = 28.1, p < 10^{-5}$). At a low dose, levodopa decreased the probability of late antisaccades, whereas at high dose it had the opposite effect.

On average, galantamine did not affect the arrival time of the inhibitory unit, nor did it affect the probability of an inhibition failure. However, there was a highly dose dependent effect on both metrics (Fig. 8B-C; RT: $F_{1,997} = 64.1, p < 10^{-5}$, probability $\chi^2(2, N = 1056) = 53.2, p < 10^{-5}$). At a low dose, galantamine reduced the hit time of the inhibitory unit and the probability of an inhibition failure, and this effect was reversed at higher dose levels.

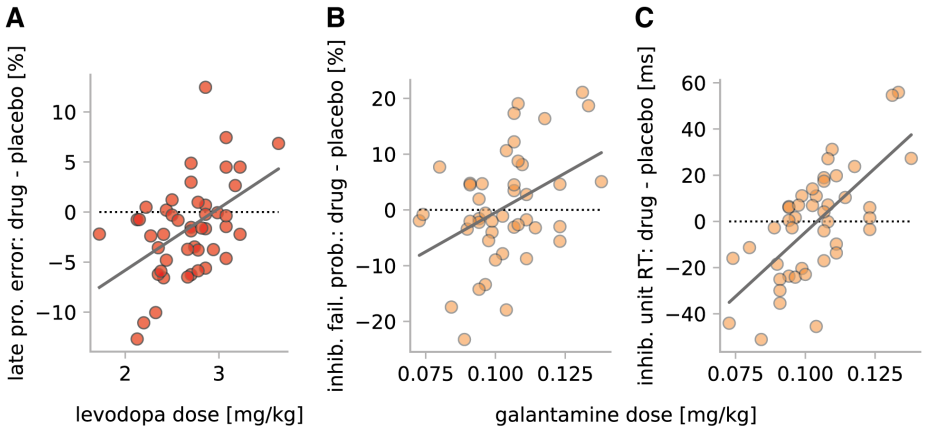


Figure 8: Dose dependent drug effects. A. Difference (drug-placebo) in late errors in prosaccade trials as a function of dose in Exp. 1 ($\chi^2(1, N = 552) = 28.1, p < 10^{-5}$). At a high dose, levodopa increased the number of errors. **B.** Difference (drug-placebo) in the percentage of inhibition failures averaged across all conditions in Exp. 2 ($\chi^2(2, N = 1056) = 53.2, p < 10^{-5}$). Galantamine increased the number of inhibition failures as a function of dose. **C.** Difference (drug-placebo) in the RT of the inhibitory unit averaged across all condition in Exp. 2 $F_{1,997} = 64.1, p < 10^{-3}$. Galantamine increased the latency of the inhibition unit as a function of dose.

Classification of drug effects

The main goal of the present study was to contrast the effects of a pro-dopaminergic and a pro-cholinergic compound. To demonstrate that these effects have predictive validity, we aimed to classify the compound administered to subjects using an approach that we refer to as *hierarchical generative embedding*. In generative embedding (Brodersen et al., 2011), a generative model is used to infer upon features of the data that are assumed to be relevant for predictive classification. The latter task can be performed by a second algorithm such as a support vector machine or a decision tree classifier. Here, we extend this

approach by relying on a hierarchical model to extract the features used to train a classifier, hence *hierarchical embedding*.

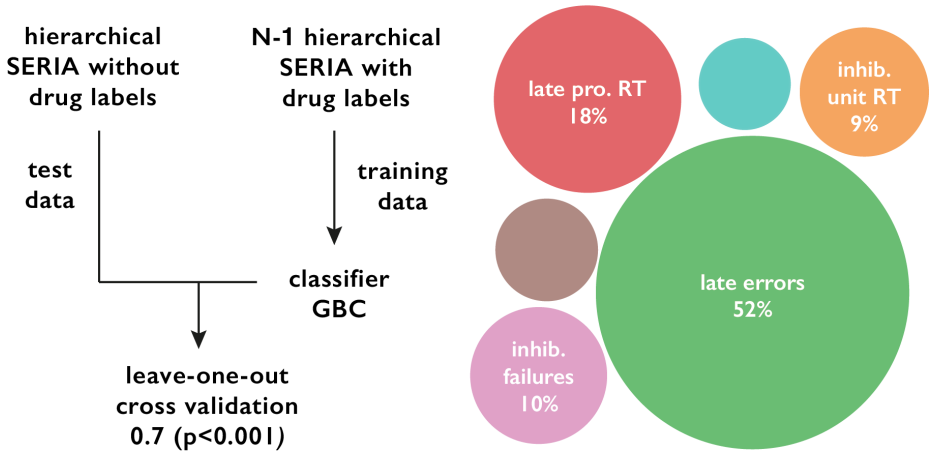


Figure 9: Generative embedding for drug classification. We used the hierarchical SERIA model to extract the features used to train a *gradient boosting classifier* (GBC). Train and test features were extracted independently. In particular, the training features were estimated by fitting the hierarchical SERIA model *N* times including information about the drug or experiment but excluding the test subject. Thus, no information about the test data was used to generate the training features. The test features were estimated using the totality of the data, without entering any information related to the drug/placebo or experiment labels.

Classification was performed on the difference between the DRUG and PLACEBO condition of the SERIA statistics described in previous sections. The goal was to predict whether a subject received levodopa or galantamine. Leave-one-out cross-validation resulted in 0.7 predictive accuracy (99% CI [0.57, 0.82]). A permutation test in which the DRUG and PLACEBO labels were randomly swapped revealed that the probability of this predictive accuracy under the null hypothesis was below 0.001. Because drug/placebo labels (but not experiment labels) were permuted, this test rules out that the accuracy of the classifier depends on the main effects of the experiment.

Finally, since a GBC assigns interpretable weights to the features used for classification, we could examine the relative importance of each of the features extracted by SERIA. As expected from the univariate regression analysis, the probability of a late error combined with hit time of late prosaccades accounted for 71% of the decision tree weights.

Discussion

The present study investigated the effects of a pro-dopaminergic and a pro-cholinergic compound in the antisaccade task. Pro- and antisaccade trials were mixed in three blocks with different probability in two experiments. Qualitatively, both experiments were comparable, although Exp. 2 was characterized by lower RTs, likely caused by faster late responses. Despite this difference, we replicated the findings in Chapter 7 and 8. For example, both RT and ER were strongly influenced by trial type probability. Thus, qualitatively both experiments followed identical patterns, evident in terms of summary statistics (Fig. 2 and 4). In the following, we discuss our experimental findings with a focus on drug effects.

Effects of levodopa

Levodopa is a precursor of DA that increases its systemic availability. As such, its influence on the central nervous system is not specific and cannot be univocally associated with specific receptor types, for example D1 or D2. Even though it is a widely-used compound in humans, the effect of levodopa on performance in the antisaccade task has not been investigated systematically. We are aware of only two studies that have used it in the context of this paradigm (Duka and Lupp, 1997; Hood et al., 2007). Unfortunately, these studies cannot be directly compared to our experiment because Hood et al. (2007) examined Parkinson's disease patients on and off medication, whereas Duka and Lupp (1997) did not follow a placebo-controlled design. In the following, we summarize our findings and present possible interpretations, emphasizing the methodological challenges and conflicting evidence available in the literature.

First, we did not find an effect of levodopa on the number of errors in antisaccade trials, which was also reflected by the absence of a significant effect on the probability of an inhibition failure in our modeling results. This negative finding is compatible with previous studies in the stop-signal task that have failed to demonstrate higher error rate or changes in the stop-signal response time (SSRT) with levodopa (Obeso et al., 2011; Overtoom et al., 2003). One possible explanation is that inhibitory control of early responses in the antisaccade task is not controlled uniquely by the basal ganglia (BG; Condy et al., 2004). This possibility is supported by the sparsity of fMRI

human studies showing increased activation of the BG when comparing pro- and antisaccades (Jamadar et al., 2013; McDowell et al., 2008). Note that although several studies have reported increased activation of the BG in antisaccade trials when compared to fixation (Ettinger et al., 2008; Matsuda et al., 2004), a quantitative meta-analysis (Jamadar et al., 2013) failed to establish differences between trial types in this region. However, more recent imaging studies (Pierce and McDowell, 2017; 2016) using liberal thresholds (Eklund et al., 2016), reported increased activation in the BG, in line with the activations in the caudate nucleus reported by Ford et al. (2009) in the macaque brain.

Models that relate DA mediated activity in the basal ganglia to inhibitory control in the antisaccade task (Wiecki and Frank, 2013) suggest that global inhibition should not be affected by changes in tonic DA levels in the BG. Indeed, inhibitory control in our version of the antisaccade is not affected by the cue presented in a particular trial (Overtoom et al., 2003) and is thus global. This reconciles our findings with the modeling predictions by Wiecki and Frank (2013).

Prominently, we found that levodopa increased the RT of late pro- and antisaccades, and, on average, reduced the number of errors in prosaccade trials. According to our model (Aponte et al., 2017), an error in a prosaccade trial can only occur when the competition between late responses is won by the unit representing an antisaccade. Arguably, these errors can be interpreted as failures to map the visual cue with the correct behavioral response. Similarly, Lo and Wang (2016) explained late errors as the result of the winner-take-all competition between two attractors (representing voluntary pro- and antisaccades). Based on this physiological model, one possible explanation of our results is that at low doses D1R mediated inhibition induces higher network stability and reduces error rates, while at higher doses excessive inhibition leads to loss in accuracy (Vijayraghavan et al., 2007). In both cases, higher inhibition would lead to increased reaction time in late responses but not necessarily in early reactions. We found that levodopa increased the mean hit time of both late pro- and antisaccades, while also reducing the number of late errors.

Although this interpretation aligns with several theoretical accounts (Lo and Wang, 2016) and previous findings such as Funahashi et al. (1993), a recent study (Vijayraghavan et al., 2016) seems to provide evidence

against it. This group reported that iontophoretic application of a D1R agonist in the dorsolateral prefrontal cortex of macaque monkeys increased antisaccade RT and ER. However, the latter effect was specific to D1R neuromodulation, as it was not elicited by the application of a D2R agonist. Moreover, application of the D1R agonist reduced the rule-encoding selectivity of neurons that preferentially fired for pro- or antisaccades cues, a somewhat surprising result when compared to findings in the delayed memory task (Vijayraghavan et al., 2007). Thus, an alternative explanation of our results is that the effect of levodopa was not connected to activation of the D1R in prefrontal areas. In this direction, Watanabe and Munoz (2013) showed that subthreshold electrical stimulation of the caudate can bias voluntary responses in the antisaccade task, suggesting that the effects observed here could still be mediated by DA neuromodulation in the BG.

Effects of galantamine

To our knowledge, no study has investigated the effect of galantamine or any other ACh allosteric potentiating drugs in the antisaccade task, although a recent study investigated the effect of galantamine in a modified version of the Posner task (Vossel et al., 2014a; 2014b). In general, our results are compatible with the common finding that nicotine decreases RT in a variety of paradigms (Heishman et al., 2010). In the antisaccade task, our results replicate previous findings that nicotine leads to reduced antisaccade RT (Bowling and Donnelly, 2010; Ettinger et al., 2009; Larrison et al., 2004; Rycroft et al., 2006; (Petrovsky et al., 2013) although for negative evidence see Ettinger et al. (2017).

Interestingly besides the effect on late responses, galantamine also affected the number of early errors in dose dependent manner. Such an interaction was reported in a previous study (Vossel et al., 2014), and agrees with in vitro studies (Samochocki et al., 2003) as well as an in vivo study in rodents (Woodruff-Pak et al., 2001). In humans, galantamine at high doses (32mg/day) exhibits a deleterious effect in inhibitory control and working memory (Dyer et al., 2008), despite its good tolerability. Model based analysis revealed that galantamine affected ER mostly through its influence on response inhibition. At high dose, galantamine had a deleterious effect, whereas at more moderate levels, it improved performance.

Prior uncertainty: Levodopa and galantamine

Here, we manipulated the ratio between pro- and antisaccade trials across different blocks. Replicating the findings in Chapter 6 to 8, this factor had a profound impact on RT and ER. Because of its contextual nature, trial type probability can be considered a top-down effect, in contrast to stimulus driven, bottom-up effects. Our experiment allowed us to investigate whether levodopa or galantamine also had an effect on prior expectations that could lead to behavioral changes, as it has been proposed before (Feldman and Friston, 2010; Iglesias et al., 2013; Vossel et al., 2014a; Yu and Dayan, 2005). For example, Yu and Dayan (2005) have suggested that ACh signaling is related to prior uncertainty, which in our experiment is highest in the PP50 condition but equal in the PP80 and PP20 conditions.

Although, in our study prior expectations and drug manipulations had significant effects, we did not find any significant interaction between them. Our findings suggest that none of the compounds used here significantly affected how prior expectations were represented or leveraged by our participants. One possible explanation for this is that DA and ACh might play a role in prior belief updates (Iglesias et al., 2013; Vossel et al., 2014a; Yu and Dayan, 2005), but not in the actual *beliefs*. Our study did not aim at capturing inter-trial differences in behavior, based on the assumption that subjects quickly adapted to the different trial type probabilities.

Summary

Here we investigated the effect of a pro-cholinergic and a pro-dopaminergic compound on inhibitory control and voluntary, goal-directed actions in the antisaccade task. Through a computational model we were able to show that both substances had opposite effects on goal-directed behavior but not in inhibitory control. Although top-down effects had a large effect on the probability of an inhibition failure, contextual effects did not interact with DA or ACh. Our study helps to further our understanding of the role of this neuromodulator in cognitive control.

Methods

Experiment and Apparatus

All the procedures described here were approved by the local ethics committee (KEK-ZH-Nr.2014-0246). Part of the data presented here was reported in a previous study (Aponte et al., 2017).

Participants

Prospective subjects were contacted through the University Registration Center for Study Participants of the University of Zurich. Subjects received first an electronic mail describing the experiment. Those who answered to the initial contact were pre-screened on the telephone, and then invited to our facilities for a complete screening. During this visit, and after inform consent was obtained in written form, medical and demographic information was collected. Besides these data, subjects underwent an electrocardiogram, a visual acuity test. Only male participants were recruited in this experiment, as hormonal changes in females can interact with medication affecting the dopaminergic system (Fernandez et al. 2003; Caldu and Dreher 2007). Exclusion criteria were: age outside the range 18-40, pulse lower than 55bps or higher than 100bps or any abnormality detected through electrocardiography, regular smoking, recreational drug consumption in the past 6 months, serious mental or neurological illness, current use of psychiatric medication, use of medication that could interact with levodopa, benserazide or galantamine, lactose intolerance, or if the medical doctor supervising the experiment considered the participant not apt.

Subjects that fulfilled all the inclusion criteria were invited to two experimental sessions separated by at least one week but no more than 8 weeks apart from the screening session. At the end of the experiment, participants received monetary compensation.

Experimental procedure

In each session, participants were first asked for alcohol consumption in the previous day and consumption of recreational drugs since the screening session. Immediately afterwards, subjects were administered a capsule containing either Madopar® DR 250g (200mg levodopa, 50mg benserazide), or lactose (Exp. 1), or Reminyl® (8mg galantamine), or lactose (Exp. 2). These were color and shape matched by a pharmacist and experimenters and participants were unaware of the drug-session

labeling but were informed in which experiment they participated. Once the substance was administered, subjects were brought to the experimental room where they received written instructions regarding the experiment. This was immediately followed by a training session scheduled to last between 20 and 30 minutes. Once the training was completed, subjects were offered a pause until the next phase of the session.

The main experiment started 70 minutes after subjects were administered the capsule. After the end of each experimental session, subjects were required to fill a debriefing questionnaire. In the first experimental session, participants were asked whether they have been administered placebo or the active substance. After the second session, they were asked in which session they thought they have received the active substance. In both sessions, subjects were also asked about the difficulty of the task and their subjective feeling of tiredness.

Procedure and Apparatus

All the experiments were conducted in a dimly illuminated room. Participants sat $60cm$ in front of a computer screen (Philips 20B40 CRT, $30cm \times 41.5cm$) working at a $75Hz$ refreshing frequency. Their head was stabilized with a chin rest. Eye movements were measured with a remote infrared camera (Eyelink I, SR Research, Ottawa, Canada), and stored at a $500Hz$ sampling rate. Before each block the eyetracker was calibrated with a five-point procedure.

Fig. 1 displays schematically the task design. During the totality of the experiment two red circles subtending $0.25dva$ were presented at a horizontal eccentricity of $\pm 12dva$. At the beginning of each trial, a fixation cross was presented at the center of the screen. Once participants fixated the cross, it remained on the screen for a random interval between 500 to 1000 ms, after which it was removed. Simultaneously a green bar ($3.48 \times 0.8dva$) centered at either the right or left red circle was displayed. The bar was presented in either vertical or horizontal orientation for $500ms$. Subjects were instructed to saccade to the red dot in the ipsilateral location of the bar if the bar was oriented horizontally (prosaccade trials), and to saccade to the contralateral red dot if the bar was oriented vertically (antisaccade trials). Once the bar was removed from the screen and after $1000ms$ the next trial started.

The training block comprised of 50 prosaccade trials followed by 50 antisaccade trials. Each trial was followed by automatic feedback provided in form of a green "Correct" sign on the screen when a trial was correctly performed, a red "Wrong direction" sign when a directional error was committed, or "Wrong. No saccade" when no saccade was detected after $500ms$. This was introduced to encouraged participants to react quickly to the peripheral visual cue.

The main experiment consisted of three blocks of 192 trials each. Every block contained randomly interleaved pro- and antisaccade trials. Each block comprised of either 20%, 50% or 80% of prosaccade trials (horizontal bar). Thereby, both experiments followed a two by three factorial design with factors $DRUG$ and prosaccade trial probability (PP). The order of presentation of the blocks was kept constant across the two sessions of a participant, but pseudo-randomized across subjects. The same random sequence of trials (which was different for each block) was used in all subjects. The location of the cue (right or left) was created randomly once and used in all blocks and participants. A short pause was offered to the participants between blocks.

Data preprocessing

Saccades were detected using the algorithm provided by the eye tracker manufacturer (SR-Research, Ottawa, Canada). According to it, a saccade is defined by a velocity threshold of $22dva/s$ and an acceleration threshold of $3800dva/s^2$ (Stampe, 1993). Saccades with a magnitude lower than $2dva$ were ignored.

Data was parsed using in-house software written in the Python programming language (2.7.11) using the numpy (1.10.0) and scipy (0.15.0) libraries. A trial was excluded i) if a blink was detected between the cue presentation (CP) and the main saccade, ii) if data from the trial was missing, iii) if the trial was aborted by the experimenter, iv) if the subject fixation was not maintained until the cue presentation, v) if a saccade was detected only after 800ms from CP, vi) if it had a latency lower than 50ms and in the case of antisaccades, if its latency was lower than 110ms. Corrective antisaccades were defined as contralateral saccades that followed an antisaccade error. These were required to have an absolute magnitude of less than 15dva and more than 4dva and to have a latency below 900ms.

Modeling

The SERIA model formalizes and extends the work of Noorani and Carpenter (2013). Our main assumption is that saccades are the result of four independent racing processes or units (Logan et al., 1984): an early response unit u_e associated with fast prosaccades, an inhibitory unit u_i whose function is to stop early actions, and two late response units that represent the commands to perform either a late prosaccade (u_p) or an antisaccade (u_a). These processes can be conceptualized as racing to a threshold s with an increase rate r. The threshold-hit-time t is given by:

$$rt = s, \tag{77}$$

$$t = \frac{s}{r}. \tag{78}$$

Fundamentally, the RT and response type on a trial are a function of the order and hit time of each of the units. Based on this simple assumption, we define the joint likelihood of a reaction time $T \in [0, \infty[$ and response action $A \in \{pro, anti\}$. This term is a function of the hit times of each of the units: $U_e,\ U_i,\ U_p, U_a \in]0, \infty[$, which we assume are randomly distributed in each trial.

According to the model, a prosaccade is generated at time t in three different scenarios. First, a prosaccade is produced at time t if the early unit hits threshold at time t before the other units:

$$(U_e = t) \cap (t < U_i) \cap \left(t < U_p\right) \cap (t < U_a). \tag{79}$$

Second, a prosaccade is generate also if the late prosaccade unit arrives to threshold at time t and all other units arrive at some later point

$$\left(U_p = t\right) \cap (t < U_i) \cap (t < U_e) \cap (t < U_a), \tag{80}$$

or, third, if the inhibitory unit arrives before the early until, and the late prosaccade unit hits threshold at time t before the antisaccade unit

$$\left(U_p = t\right) \cap (U_i < t) \cap (U_i < U_e) \cap (t < U_a). \tag{81}$$

In the following we assume that the probability that a unit U reaches threshold at time t is given by the density $p(U=t)$. Thereby, the probability of a prosaccade at time t can be written down as the sum of the probability of each of these three scenarios:

$$p(A = pro, T = t) = p(U_e = t)p(t < U_i)p(t < U_a)p\left(t < U_p\right) +$$
$$p\left(U_p = t\right)p(t < U_a)p(t < U_e)p(t < U_i) +$$

$$p\big(U_p = t\big)p(t < U_a)\int_0^t p(U_i = \tau)p(\tau < U_e)d\tau\,. \tag{82}$$

The probability of an antisaccade is given by a similar equation, except that antisaccades can only be triggered by the antisaccade unit u_a:

$$p(A = anti, T = t) = p(U_a = t)p\big(t < U_p\big)p(t < U_e)p(t < U_i) +$$
$$p(U_a = t)p\big(t < U_p\big)\int_0^t p(U_i = \tau)p(\tau < U_e)d\tau \tag{83}$$

We were interested in several statistics that could be derived from the model. First, we considered the expected hit time of the early and inhibitory unis. The expected hit times of the early and inhibitory units U_e and U_i are given by:

$$E[U_e] = \int_0^\infty tp(U_e = t)dt, \tag{84}$$

$$E[U_i] = \int_0^\infty tp(U_i = t)dt. \tag{85}$$

In addition, we computed the probability of inhibition failures that occurs when the early unit hits threshold before all other units:

$$p(inhib.fail) = \int_0^\infty p(U_e = t)p(U_i > t)p\big(U_p > t\big)p(U_a > t)\,dt, \tag{86}$$

and their expected response time

$$E[time\ inhib.fail.] = \frac{\int_0^\infty t\,p(U_e = t)p(U_i > t)p\big(U_p > t\big)p(U_a > t)dt}{p(inhib.fail.)}. \tag{87}$$

Finally, we considered the conditional probability of making a late antisaccade, which happens when the antisaccade unit U_a hits threshold before the late prosaccade unit U_p:

$$p(anti.) = \int_0^\infty p(U_a = t)p\big(U_p > t\big)\,dt, \tag{88}$$

$$= 1 - \int_0^\infty p\big(U_p = t\big)p(U_a > t)\,dt, \tag{89}$$

$$= 1 - p(late\ pro.) \tag{90}$$

and the mean response time of late pro and antisaccades

$$E[hit\ time\ anti.] = \frac{1}{p(anti.)}\int_0^\infty t\,p(U_a = t)p\big(U_p > t\big)\,dt, \tag{91}$$

$$E[hit\ time\ late\ pro.] = \frac{1}{p(late\ pro.)} \int_0^\infty t\, p(U_p = t) p(U_a > t)\, dt. \quad (92)$$

As shown in previous chapters, the distribution of the early and inhibitory units are well described by the inverse Gamma distribution. However, the distribution of the arrival times of the late pro- and antisaccade units is better described by the Gamma distribution.

Using the same approach as in Chapter 8, we divided trials in repeat and switch pro- and antisaccades. Note that in Chapter 8, we showed that in the model with the highest Bayes factor, the parameters of the early unit should be fixed across all trial types, but the parameters of the inhibitory and late units could vary across them. We used the same model here, which requires 29 parameters per block per subject. Although this number of parameters might seem elevated, Bayesian model comparison with other nested models demonstrated that less complex models could not account adequately for the switch costs in the mixed antisaccade task. In order to regularize the parameter estimates, we used a hierarchical model in which information from all subjects was pooled together, as explained below.

Statistical inference

Model estimation and fitting

We used Bayesian inference to fit all models. The goal in this approach is to compute the posterior distribution of the parameters θ given the observed data y and a model m. This distribution is given by Bayes theorem:

$$\overbrace{p(\theta|y,m)}^{posterior} = \frac{\overbrace{p(y|\theta,m)}^{likelihood}\overbrace{p(\theta|m)}^{prior}}{\underbrace{p(y|m)}_{evidence}}. \quad (93)$$

Because the posterior distribution of the models evaluated here lacks a closed form, we resorted to the Metropolis Hasting algorithm to obtain samples from the posterior distribution of the parameters. Details can be found in Aponte et al. (2017, 2016).

We assumed that in each block and session, each trial was independently and identically distributed conditioned a set of parameters $\theta_{b,s,j,e,t,l}$, where the index $b \in \{PP20, PP50, PP80\}$ codes the block, $s \in \{1,2\}$ codes the session, $j \in \{1, \dots, N\}$ codes the subject, $e \in \{madopar, reminyl\}$

codes the experiment, $t \in \{pro, anti\}$ codes the trial type, and $l \in \{switch, repeat\}$ codes for switch trials.

For regularization, we used a hierarchical model in which the prior distribution of the parameters in a session depended on the population distribution, which was estimated using the totality of the data. To account for the effects of our experimental manipulations and intersubjective variability, the prior mean of each set of parameters was defined as the linear combination of a design matrix coding for the factors SUBJECT, SESSION, PP, and DRUG, DRUG*DOSE and EXP. and a set of coefficients β, estimated from the data. A further level of the model was defined by estimating the mean of the subject specific parameters, or equivalently, by estimating the population mean from the subject specific coefficients. Mathematical details can be found in the supplementary materials of Chapter 7.

This approach provides an empirically motivated prior that regularizes the parameter estimates and also models experimental manipulations. This is important, as non-hierarchical models that disregard all the manipulations used in an experiment are less sensitive than parametric hierarchical models (Bae et al., 2016; Cannon et al., 2001).

Univariate regression analysis

Statistical analyses were performed using mixed effect generalized linear models (GLME) implemented in R 3.4.3 (packages *lmer* and *glmmadmb*). Subjects were always entered as a random effect, whereas the factors SWITCH, PP, SESSION, DRUG, EXP., and DOSE were treated as fixed effects. In addition, we considered the following interaction PP*SESSION, PP*SWITCH, PP*DRUG, DOSE*DRUG. When both experiments were analyzed together, we also included the interactions DRUG*EXP., DRUG*DOSE*EXP and DRUG*PP*EXP. Error rates were analyzed using a binomial regression model (*glmer* from the package *lmer*), whereas probabilities estimated from the model were estimated using a Beta regression model (*glmmadmb*). To verify the results, we also analyzed probabilities with a GLME in which these were probit transform, obtaining almost identical results. Inferential statistic on RT were performed using *F*-tests with the degrees of freedom computed using the Sattherthwait approximation. For error rates and probabilities, Wald tests using the χ^2 statistic were used. Statistical significance was asserted at $\alpha = 0.05$.

Clustering

The main goal of clustering was to determine whether the features extracted using the SERIA model were truly predictive of the type of drug administered to a subject. Thus, we were interested into determining whether the *differences* between the drug and placebo condition could be classified as either *levodopa* or *galantamine*.

We generated *N* training data-sets by excluding one of the subjects from the hierarchical model, which included information about drug labels and experiment. The corresponding leave-one-out test dataset was generated by fitting a hierarchical model in which all subjects were included, but no information about the labeling was present. We followed this strategy, which we refer to as *hierarchical generative embedding*, to implement robust regularization on the parameter estimates, without introducing any bias into the algorithm. We aim to classify whether a subject received levodopa or galantamine based on the *difference* between model parameters in the drug and placebo condition.

Clusturing was performed using *gradient boosting* (Friedman, 2001) as implemented in the package *xgboost* (Chen and Guestrin, 2016). This algorithm uses simple decision trees, which are sequentially improved on the subsets of the data in which the classifier performed poorly. Since it uses simple decision trees, it is possible to directly extract the features that were used by the classifier.

To evaluate the clustering accuracy, we used leave-one-out cross validation. Because in principle the algorithm could use general differences between Exp. 1 and 2 to classify subjects, as opposed to differences between drug and placebo condition, we implemented a simple permutation test in which the drug and placebo labels were randomly swapped 50000 times. This represents a null hypothesis in which the drug but not the experiment label was modified.

References

Aponte, E.A., Schobi, D., Stephan, K.E., Heinzle, J., 2017. The Stochastic Early Reaction, Inhibition, and Late Action (SERIA) Model for Antisaccades SERIA - A model for errors and reaction times in the antisaccade task. bioRxiv. doi:10.1101/109090

Bae, H., Monti, S., Montano, M., Steinberg, M.H., Perls, T.T., Sebastiani, P., 2016. Learning Bayesian Networks from Correlated Data. Sci Rep 6, 25156.

Brodersen, K.H., Schofield, T.M., Leff, A.P., Ong, C.S., Lomakina, E.I., Buhmann, J.M., Stephan, K.E., 2011. Generative embedding for model-based classification of fMRI data. PLoS Comput Biol 7, e1002079.

Cannon, M.J., Warner, L., Taddei, J.A., Kleinbaum, D.G., 2001. What can go wrong when you assume that correlated data are independent: an illustration from the evaluation of a childhood health intervention in Brazil. Stat Med 20, 1461–1467.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in:. Presented at the Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794.

Condy, C., Rivaud-Pechoux, S., Ostendorf, F., Ploner, C.J., Gaymard, B., 2004. Neural substrate of antisaccades: role of subcortical structures. Neurology 63, 1571–1578.

Duka, T., Lupp, A., 1997. The effects of incentive on antisaccades: is a dopaminergic mechanism involved? Behav Pharmacol 8, 373–382.

Dyer, M.A., Freudenreich, O., Culhane, M.A., Pachas, G.N., Deckersbach, T., Murphy, E., Goff, D.C., Evins, A.E., 2008. High-dose galantamine augmentation inferior to placebo on attention, inhibitory control and working memory performance in nonsmokers with schizophrenia. Schizophr. Res. 102, 88–95.

Eklund, A., Nichols, T.E., Knutsson, H., 2016. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proc. Natl. Acad. Sci. U.S.A. 113, 7900–7905.

Ettinger, U., Faiola, E., Kasparbauer, A.M., Petrovsky, N., Chan, R.C., Liepelt, R., Kumari, V., 2017. Effects of nicotine on response inhibition and interference control. Psychopharmacology (Berl.) 234, 1093–1111.

Everling, S., Johnston, K., 2013. Control of the superior colliculus by the lateral prefrontal cortex. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 368, 20130068.

Ford, K.A., Gati, J.S., Menon, R.S., Everling, S., 2009. BOLD fMRI activation for anti-saccades in nonhuman primates. Neuroimage 45, 470–476.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics 1189–1232.

Funahashi, S., Chafee, M.V., Goldman-Rakic, P.S., 1993. Prefrontal neuronal activity in rhesus monkeys performing a delayed anti-saccade task. Nature 365, 753–756.

Hallett, P.E., 1978. Primary and secondary saccades to goals defined by instructions. Vision Res. 18, 1279–1296.

Heinzle, J., Aponte, E.A., Stephan, K.E., 2016. Computational models of eye movements and their application to schizophrenia. Current Opinion in Behavioral Sciences 11, 21–29. doi:http://dx.doi.org/10.1016/j.cobeha.2016.03.008

Heishman, S.J., Kleykamp, B.A., Singleton, E.G., 2010. Meta-analysis of the acute effects of nicotine and smoking on human performance. Psychopharmacology (Berl.) 210, 453–469.

Hikosaka, O., Isoda, M., 2010. Switching from automatic to controlled behavior: cortico-basal ganglia mechanisms. Trends Cogn. Sci. (Regul. Ed.) 14, 154–161.

Hood, A.J., Amador, S.C., Cain, A.E., Briand, K.A., Al-Refai, A.H., Schiess, M.C., Sereno, A.B., 2007. Levodopa slows prosaccades and improves antisaccades: an eye movement study in Parkinson's disease. J. Neurol. Neurosurg. Psychiatr. 78, 565–570.

Isoda, M., Hikosaka, O., 2011. Cortico-basal ganglia mechanisms for overcoming innate, habitual and motivational behaviors. Eur. J. Neurosci. 33, 2058–2069.

Isoda, M., Hikosaka, O., 2008. Role for subthalamic nucleus neurons in switching from automatic to controlled eye movement. J. Neurosci. 28, 7209–7218.

Jamadar, S.D., Fielding, J., Egan, G.F., 2013. Quantitative meta-analysis of fMRI and PET studies reveals consistent activation in fronto-striatal-parietal regions and cerebellum during antisaccades and prosaccades. Front Psychol 4, 749.

Lo, C.C., Wang, X.J., 2016. Conflict Resolution as Near-Threshold Decision-Making: A Spiking Neural Circuit Model with Two-Stage Competition for Antisaccadic Task. PLoS Comput Biol 12, e1005081.

Logan, G.D., Cowan, W.B., Davis, K.A., 1984. On the ability to inhibit simple and choice reaction time responses: a model and a method. J Exp Psychol Hum Percept Perform 10, 276–291.

Matsuda, T., Matsuura, M., Ohkubo, T., Ohkubo, H., Matsushima, E., Inoue, K., Taira, M., Kojima, T., 2004. Functional MRI mapping of brain activation during visually guided saccades and antisaccades: cortical and subcortical networks. Psychiatry Res 131, 147–155.

Munoz, D.P., Everling, S., 2004. Look away: the anti-saccade task and the voluntary control of eye movement. Nat. Rev. Neurosci. 5, 218–228.

Noorani, I., Carpenter, R.H., 2013. Antisaccades as decisions: LATER model predicts latency distributions and error responses. Eur. J. Neurosci. 37, 330–338.

Petrovsky, N., Ettinger, U., Quednow, B.B., Landsberg, M.W., Drees, J., Lennertz, L., Frommann, I., Heilmann, K., Strater, B., Kessler, H., Dahmen, N., Mossner, R., Maier, W., Wagner, M., 2013. Nicotine enhances antisaccade performance in schizophrenia patients and healthy controls. Int. J. Neuropsychopharmacol. 16, 1473–1481.

Reuter, B., Jager, M., Bottlender, R., Kathmann, N., 2007. Impaired action control in schizophrenia: the role of volitional saccade initiation. Neuropsychologia 45, 1840–1848.

Rycroft, N., Hutton, S.B., Clowry, O., Groomsbridge, C., Sierakowski, A., Rusted, J.M., 2007. Non-cholinergic modulation of antisaccade performance: a modafinil-nicotine comparison. Psychopharmacology (Berl.) 195, 245–253.

Samochocki, M., Hoffle, A., Fehrenbacher, A., Jostock, R., Ludwig, J., Christner, C., Radina, M., Zerlin, M., Ullmer, C., Pereira, E.F., Lubbert, H., Albuquerque, E.X., Maelicke, A., 2003. Galantamine is an allosterically potentiating ligand of neuronal nicotinic but not of muscarinic acetylcholine receptors. J. Pharmacol. Exp. Ther. 305, 1024–1036.

Stampe, D., 1993. Heuristic filtering and reliable calibration methods for video-based pupil-tracking systems. Behavior Research Methods, Instruments, \& Computers 25, 137–142. doi:10.3758/BF03204486

Stuphorn, V., Schall, J.D., 2006. Executive control of countermanding saccades by the supplementary eye field. Nat. Neurosci. 9, 925–931.

Texido, L., Ros, E., Martin-Satue, M., Lopez, S., Aleu, J., Marsal, J., Solsona, C., 2005. Effect of galantamine on the human alpha7 neuronal nicotinic acetylcholine receptor, the Torpedo nicotinic acetylcholine

receptor and spontaneous cholinergic synaptic activity. Br. J. Pharmacol. 145, 672–678.

Vijayraghavan, S., Major, A.J., Everling, S., 2016. Dopamine D1 and D2 Receptors Make Dissociable Contributions to Dorsolateral Prefrontal Cortical Regulation of Rule-Guided Oculomotor Behavior. Cell Rep 16, 805–816.

Vijayraghavan, S., Wang, M., Birnbaum, S.G., Williams, G.V., Arnsten, A.F., 2007. Inverted-U dopamine D1 receptor actions on prefrontal neurons engaged in working memory. Nat. Neurosci. 10, 376–384.

Vossel, S., Bauer, M., Mathys, C., Adams, R.A., Dolan, R.J., Stephan, K.E., Friston, K.J., 2014. Cholinergic stimulation enhances Bayesian belief updating in the deployment of spatial attention. J. Neurosci. 34, 15735–15742.

Watanabe, M., Munoz, D.P., 2013. Effects of caudate microstimulation on spontaneous and purposive saccades. J. Neurophysiol. 110, 334–343.

Wiecki, T.V., Frank, M.J., 2013. A computational model of inhibitory control in frontal cortex and basal ganglia. Psychol Rev 120, 329–355.

Woodruff-Pak, D.S., Vogel, R.W., Wenk, G.L., 2001. Galantamine: effect on nicotinic receptor binding, acetylcholinesterase inhibition, and learning. Proc. Natl. Acad. Sci. U.S.A. 98, 2089–2094.

Yu, A.J., Dayan, P., 2005. Uncertainty, neuromodulation, and attention. Neuron 46, 681–692.

Supplementary

Supplementary 1

	Mean # invalid trials							
	Experiment 1							
	Valid	Blink	Missing	Aborted	FE	Late	Early	NS
Placebo/	563(10)	4(5)	1(3)	1(2)	5(5)	0(0)	1(1)	0(0)
L-dopa	562(14)	5(7)	1(3)	1(2)	5(6)	0(0)	1(2)	0(0)
	Experiment 2							
	Valid	Blink	Missing	Aborted	FE	Late	Early	NS
Placebo/	558(24)	10(17)	3(6)	1(1)	3(7)	0(0)	1(1)	0(0)
Galan.	562(18)	9(14)	1(3)	1(2)	3(5)	0(0)	0(1)	0(0)

Table S1: Mean number of invalid trials and std. in brackets. FE = fixation errors, NS = No saccade.

Supplementary 2

	Error rate (%)					
	Experiment 1					
	Placebo			Levodopa		
	PP20	PP50	PP80	PP20	PP50	PP80
Anti.	23(17)	36(22)	52(20)	23(16)	36(21)	52(24)
Pro.	26(15)	11(8)	4(3)	25(18)	10(8)	3(4)
	Experiment 2					
	Placebo			Galantamine		
	PP20	PP50	PP80	PP20	PP50	PP80
Anti.	23(17)	36(22)	55(19)	22(17)	36(21)	53(22)
Pro.	31(17)	14(10)	5(3)	32(16)	15(9)	5(4)

Table S2: Mean error rate and std. in brackets.

Supplementary 3

Mean reaction times (ms)							
		Experiment 1					
		Placebo			Levodopa		
Trial	Action	PP20	PP50	PP80	PP20	PP50	PP80
Anti	Anti	355	381	391	367	380	396
		59	55	60	73	65	70
Anti	Pro	234	231	226	231	228	222
		51	48	31	56	43	39
Pro	Pro	332	320	285	329	319	279
		72	68	59	65	60	51
Pro	Anti	327	330	338	338	351	342
		68	45	56	70	82	65
		Experiment 2					
		Placebo			Galantamine		
Trial	Action	PP20	PP50	PP80	PP20	PP50	PP80
Anti	Anti	323	348	363	313	335	353
		50	66	72	49	52	60
Anti	Pro	218	225	215	220	220	211
		35	41	30	44	39	34
Pro	Pro	299	293	261	285	286	262
		55	58	41	59	52	52
Pro	Anti	300	307	305	293	301	314
		42	52	62	50	42	59

Table S3: Mean reaction time and std..

Supplementary 4

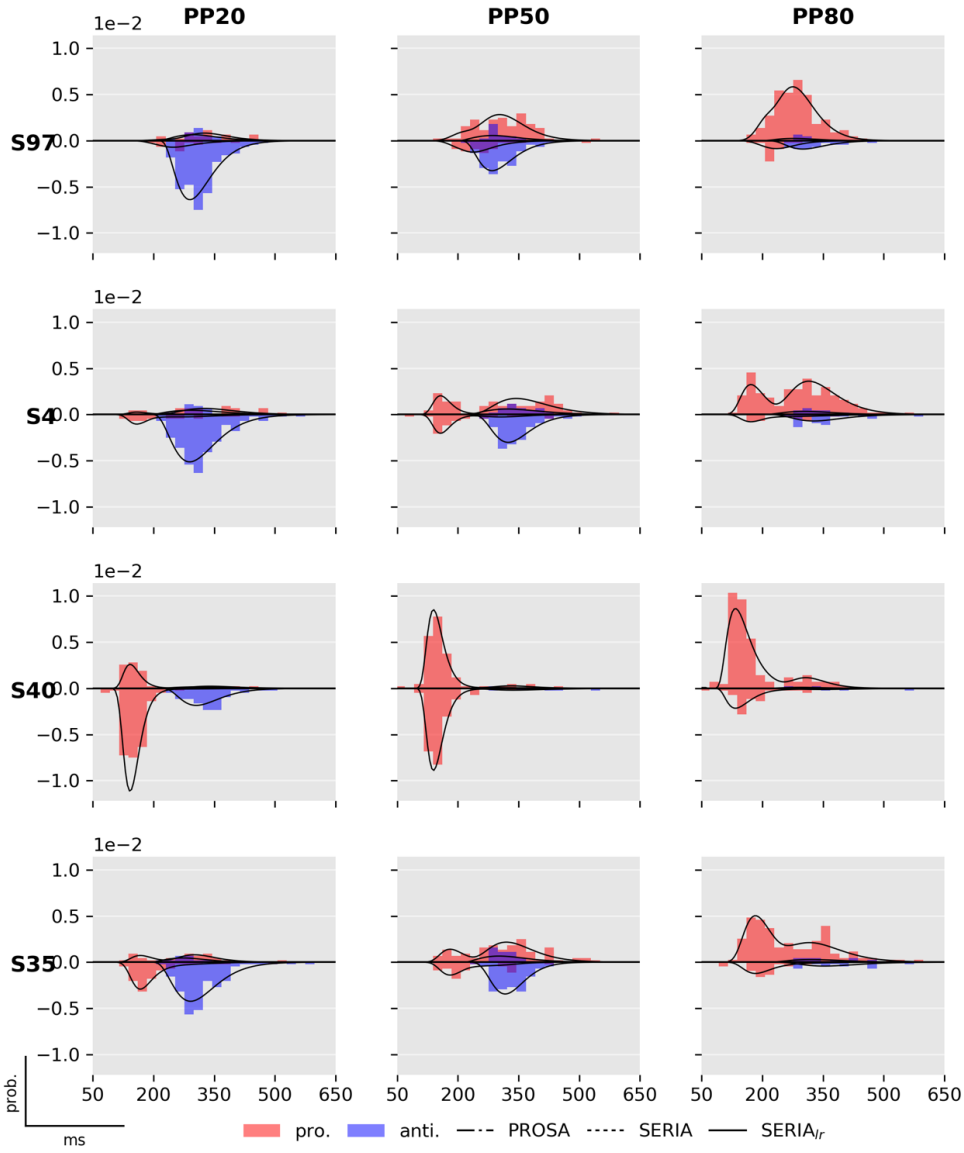


Figure S4: Subject specific fits for the three conditions. Displayed are the histogram of prosaccade and antisaccade trials of four representative subject. Antisaccade trials are displayed in the upper half-plane. Prosaccades are displayed in the bottom half plane.

Supplementary 5

The parameters of the model were analyzed using an ANOVA in the case of the hit times of the late prosaccade, antisaccade, inhibitory and early unit. The probability of a late error in pro- and antisaccade trials, and the probability of an inhibition failure were analyzed using the χ^2 statistic. The complete tests are reported in the following.

	Num. DF	Den. DF	F value	Pr(>F)	
experiment	1	86	18.5	4.37E-05	***
drug	1	972	0.3	0.572884	
session	1	972	86.42	2.20E-16	***
pp	2	972	23.2	1.35E-10	***
switch trial	1	972	517.4	2.20E-16	***
dose	1	86	1.2	0.2723037	
pp * switch trial	2	972	7.5	0.0005606	***
session * pp	2	972	0.7	0.4562863	
drug * pp	2	972	2.6	0.0729269	
experiment * drug	1	972	19.4	1.14E-05	***
experiment * pp	2	972	8.6	0.0001803	***
drug * dose	1	972	2.3	0.1274583	
experiment * dose	1	86	0.5	0.4520714	
experiment * drug * pp	2	972	1.0	0.3675007	
experiment * drug * dose	1	972	0.0	0.9209532	

Table S5. Late prosaccade unit hit time. **:p<0.05, **:p<0.01, ***:p<0.001*.

	Num. DF	Den. DF	F value	Pr(>F)	
experiment	1	86	15.4	0.00017	***
drug	1	972	0.3	0.548542	
session	1	972	137.8	2.20E-16	***
pp	2	972	64.0	2.20E-16	***
switch trial	1	972	211.2	2.20E-16	***
dose	1	86	0.5	0.4676415	
pp * switch trial	2	972	8.3	0.0002483	***
session * pp	2	972	1.3	0.2677091	
experiment * drug	1	972	15.1	1.04E-04	***
drug * dose	1	972	3.9	0.0467454	*
experiment * dose	1	86	0.4	0.5070893	
drug * pp	2	972	1.5	0.2077454	
experiment * pp	2	972	3.0	0.0489029	*
experiment * drug * dose	1	972	0.0	0.7724891	
experiment * drug * pp	2	972	2.1	0.111831	

Table S6. Antisaccade unit hit time. **:p<0.05, **:p<0.01, ***:p<0.001*.

	Num. DF	Den. DF	F value	Pr(>F)	
experiment	1	86	0.0	0.77121	
drug	1	2047	1.7	0.18731	
session	1	2047	141.7	2.20E-16	***
pp	2	2047	242.7	2.20E-16	***
trial type	1	2047	52.7	5.34E-13	***
switch trial	1	2047	414.4	2.20E-16	***
dose	1	86	2.7	0.10278	
pp * switch trial	2	2047	19.6	3.41E-09	***
session * pp	2	2047	2.9	0.05052	
pp * trial type	2	2047	59.2	2.20E-16	***
trial type * switch trial	1	2047	1913.5	2.20E-16	***
drug * trial type	1	2047	3.4	0.06378	
experiment * drug	1	2047	0.4	0.51493	
drug * dose	1	2047	44.7	2.95E-11	***
experiment * dose	1	86	3.2	0.07475	
drug * pp	2	2047	0.0	0.98194	
experiment * pp	2	2047	0.2	0.79608	
experiment * drug * dose	1	2047	31.7	2.02E-08	***
experiment * drug * pp	2	2047	0.3	0.70137	

Table S7. Inhibitory unit hit time. **:p<0.05, **:p<0.01, ***:p<0.001*.

	NumDF	DenDF	F value	Pr(>F)	
experiment	1	86	0.5	0.46012	
drug	1	435	1.6	0.19608	
session	1	435	37.9	1.65E-09	***
pp	2	435	46.2	2.20E-16	***
dose	1	86	1.8	0.17701	
session * pp	2	435	1.7	0.17962	
experiment * drug	1	435	0.0	0.875	
drug * dose	1	435	4.6	0.03173	*
experiment * dose	1	86	0.0	0.94148	
drug * pp	2	435	0.1	0.81887	
experiment * pp	2	435	0.2	0.74173	
experiment * drug * dose	1	435	4.0	0.04423	*
experiment * drug * pp	2	435	0.4	0.6511	

Table S8. Early unit hit time. *:p<0.05, **:p<0.01, ***:p<0.001.

	Df	Chisq	Pr(>Chisq)	
experiment	1	7.7	0.005442	**
drug	1	0.9	0.319472	
session	1	7.0	0.007872	**
pp	2	1634.6	2.20E-16	***
switch trial	1	3427.1	2.20E-16	***
dose	1	0	0.995112	
pp * switch trial	2	13.6	0.001112	**
session * pp	2	1.6	0.446444	
experiment * drug	1	24.0	9.31E-07	***
drug * dose	1	18.2	1.91E-05	***
experiment * dose	1	0.2	0.646652	
drug * pp	2	4.1	0.125329	
experiment * pp	2	2.1	0.349895	
experiment * drug * dose	1	7.3	0.006762	**
experiment * drug * pp	2	0.4	0.781459	

Table S9: Probability of a late error in a prosaccade trial. *:p<0.05, **:p<0.01, ***:p<0.001.

	Df	Chisq	Pr(>Chisq)	
experiment	1	0.3	0.5485	
drug	1	19.6	9.44E-06	***
session	1	0.6	0.42009	
pp	2	201.7	2.20E-16	***
switch trial	1	939.3	2.20E-16	***
dose	1	1.8	0.17895	
pp * switch trial	2	1.8	0.3938	
session * pp	2	20.1	4.15E-05	***
experiment * drug	1	1.1	0.2851	
drug * dose	1	3.6	0.05636	
experiment * dose	1	0.3	0.57927	
drug * pp	2	0.5	0.76736	
experiment * pp	2	6.0	0.04818	*
experiment * drug * dose	1	0.0	0.85902	
experiment * drug * pp	2	1.3	0.50174	

Table S10: Probability of a late error in an antisaccade trial. **:p<0.05, **:p<0.01, ***:p<0.001*.

	Df	Chisq	Pr(>Chisq)	
experiment	1	0.1	0.7052589	
drug	1	4.8	0.028009	*
session	1	0.3	0.5623175	
pp	2	294.7	2.20E-16	***
switch trial	1	92.4	2.20E-16	***
trial type	1	2.4	0.1158398	
dose	1	0.0	0.8623445	
pp * switch trial	2	1.4	0.4959548	
session * pp	2	4.7	0.0913679	
experiment * drug	1	0.1	0.7325691	
drug * trial type	1	0.0	0.8564022	
experiment * trial type	1	0.5	0.4776703	
drug * dose	1	14.0	0.0001826	***
experiment * dose	1	0.8	0.3562499	
drug * pp	2	1.6	0.4366927	
experiment * pp	2	1.7	0.4235288	
experiment * drug * trial type	1	0.2	0.6141654	
experiment * drug * dose	1	17.4	2.99E-05	***
experiment * drug * pp	2	0.8	0.6684671	

Table S11: Probability of an inhibition failure. **:p<0.05, **:p<0.01, ***:p<0.001*.

Outlook

This dissertation includes the following five contributions to the field of computational psychiatry and eye movement research. First, we provided a theoretical presentation of thermodynamic integration (TI), a sampling method for estimating the model evidence, and compared TI to other similar methods in the context of dynamic causal modeling (DCM) for fMRI. Second, we developed a novel estimator of the predictive likelihood based on TI and evaluated this estimator on an empirical data set. Third, we provide evidence for the involvement of the corollary discharge (CD) pathway in saccadic adaptation (SA), which we establish using a combination of fMRI and computational modeling. Fourth, we introduce a novel set of formal models of the antisaccade task. And finally, fifth, we apply these models to two versions of the antisaccade task, testing for both the effect of task switching and the influence of a pro-cholinergic and a pro-dopaminergic compound on the oculomotor system.

These ideas and empirical findings were sequentially organized, starting first with the epistemological question of how to evaluate competing models, followed by a more mathematical treatment of this question involving the implementation of the methods discussed, and finally the application of these methods and models to very concrete empirical questions. In the following, we discuss the next steps based on this progress, highlighting open question and ideas that we judge are worth considering in follow-up research.

Model selection

One of the main conclusions from Part I of this thesis is that population MCMC is more robust than gradient-based optimization methods when the likelihood landscape of a model is very complicated. This result should not be surprising, taking into account the history of these methods in other fields. For example, in physical chemistry, ‘temperate’ methods are the state of the art approach when simulating complex

systems (Sugita and Okamoto, 1999; Trebst et al., 2006; Sindhikara et al., 2010). This seems to be the case for other dynamical system as well (Calderhead and Girolami, 2009; Vousden et al., 2015; Ballnus et al., 2017). For these reasons, we hope to implement the same type of inference used for DCM for fMRI to DCM for EEG (Kiebel et al., 2009), in which the likelihood landscape may be even more complex than in DCM for fMRI. Note that Sengupta et al. (2015;2016) and Penny and Sengupta (2016) conducted work in this direction. Unfortunately, this work has not been followed up or applied to empirical problems.

An interesting question that was left unresolved in Chapter 1 is how to combine hierarchical models that pool information from a population of subjects with the goal of making single subject statements. The option that we presented in Chapter 1 (but see also Vehtari and Lampinen, 2002) is to use the predictive likelihood as a score on a subject-by-subject basis. An objection to this approach is that, if by hypothesis not all subjects should be 'assigned' to the same model, it seems artificial to assume first that all observations have been generated from a single model, and then use their predictive likelihood to make subject-by-subject model comparison. The state of the art method to deal with this problem is the random effects model proposed by Stephan et al. (2009b). According to it, subjects are 'assigned' to a model depending on their marginal likelihood and then the number of assignments is used to perform inference (Rigoux et al., 2014). Two are the main advantages of this approach: it explicitly models the assumed heterogeneity in a population of subjects, and it provides a way to reduce the influence of outliers. This is important when a small sample of subjects with many observations per subject is evaluated, as a single outlier can have a large influence in a 'fixed effect' analysis, in which observations are assumed to be independent (see Chapter 1). Despite these advantages, this model has also drawbacks: first, it does not pool information across subjects that are assigned to the same model. Rather, the prior distribution of each subject is assumed to be fixed. Second, it requires to precompute the model evidence of all the models and observations in advance.

One option to solve this problem when all the models in consideration share the same likelihood function is to extend the random effects model proposed by Stephan et al. (2009b) into a semi-supervised clustering algorithm. From this perspective, the method of Stephan and colleagues

can be seen as a version of K-means in which the centroids of the classifier are fixed. However, if this restriction is weakened to allow at least a subset of the dimensions to vary, it is possible to implement a version of the random effects model that pools information across subjects. This idea is a compromise between to extremes: on one hand, in unsupervised algorithms like K-means, no information is entered in the centroids, and the algorithm is left to discover structure in the data (Raman et al., 2016). Hence, these algorithms are required to find appropriate centroids in addition to assigning subjects to the correct cluster in the absence of any prior information. On the other hand, in the model proposed by Stephan et al. (2009b) the centroids are fixed and the algorithm only needs to assign observations to clusters. In a third option, clusters can be treated as models in which a set of parameters is 'softly' fixed, i.e., a shrinkage prior is applied to some of the parameters of the model.

We are currently developing this simple approach, which will avoid the costs of pre-computing the evidence of a large set of models. The main disadvantage of this idea is that it would be restricted to models that share the same likelihood function. Although this is an important limitation, model comparison is often performed in the context of nested models, as in the case of DCM.

The antisaccade task

Here, we investigated the antisaccade task with the help of the SERIA model from several perspectives. First, we provided extensive evidence that prosaccades are the result of both habitual as well as voluntary actions (Chapter 7). Moreover, we showed that switch costs affect mostly the generation of voluntary actions (Chapter 8). Finally, we demonstrated that dopaminergic (DA) and cholinergic (ACh) compounds have opposite effects on voluntary saccades (Chapter 9).

We envision four ways to continue the research program started here. The first future direction of research is to investigate the *paradoxical switching costs* (Cherkasova et al., 2002) that occur when the task demands cue is presented in advance of the peripheral target (cf. Chapter 7). We believe that this is mostly a result of inhibitory inter-trial effects that were documented in Chapter 8, although only careful modeling work might answer this question.

Second, in Chapter 8 we hypothesized that rule guided switch costs are masked in asynchronous task designs because the decision process between voluntary actions takes places in advance of the presentation of the peripheral cue. This hypothesis can be easily tested by presenting the task demands cue at different latencies before the peripheral cue, as routinely done in other domains (Kiesel et al., 2010).

Third, the neurocorrelates of the late race process could be investigated with the help of model-based EEG and fMRI. To our knowledge no previous study has taken into account the possibility that saccades are generated by habitual and voluntary responses. Because voluntary prosaccades could be triggered by a similar mechanism than antisaccades, a direct contrast between these two types of trials might have led to undeceive results in previous studies (review in Jamadar et al., 2013) .

Fourth, although not discussed in this dissertation, preparatory activity in the antisaccade task might be reflected in changes in pupil size (Wang et al., 2015). While we have been able to replicate the finding reported by Wang and Munoz, a model-based analysis that incorporates information about the physiology of pupil size changes might enrich the analysis of these data.

Finally, we aim to make the software used here available to the community in an accessible format. Although it is currently part of the tapas software package (https://www.tnu.ethz.ch/de/software/tapas.html), it lacks an accessible interface. Thus, future releases should make the SERIA model more user friendly.

Clinical applications

Despite these results, an important question remains open: what are the possible applications of this work in a clinical context? An immediate answer is that the tools developed here could be used to better address the problem of selecting a model from a set of competing hypotheses. Our analysis of empirical data in Chapter 1 clearly exemplifies that in the absence of carefully devised priors, the conclusions drawn from standard methods of Bayesian model comparison might depend on subjective beliefs. Although this might be acceptable when priors originate from expert judgement (Kass and Raftery, 1995), in the case

of the complex models routinely used in computational psychiatry, this type of prior knowledge is rare. In this direction, in Chapters 7 and 8 we showcased the application of the methods developed in Chapter 1 to 3 to test hypotheses operationalized as different hierarchical models.
Moreover, results in Chapter 3 suggest that population MCMC could decrease estimation variance when applied to complex dynamical systems as DCM. This observation has been confirmed by a large and thorough study by Ballnus et al. (2017) in which several dynamical systems were analyzed with Monte Carlo methods. The main finding of this study is that population MCMC performed better than comparable inversion methods at capturing the multimodality of these systems, and additionally, was more statistically efficient in terms of the correlation between samples, even when the increased costs of multi-chain simulations were considered. Hence, we believe that massively parallel architectures and methods that can fully leverage hardware advances will impact Bayesian estimation of complex biological models. Thus, this method could be helpful in clinical contexts.

A second path towards truly useful translational tools is laid down by the models of eye movement behavior proposed in this dissertation. These can be used to ask very concrete questions in psychiatry, such as, for example: is CD and thereby SA compromised in patients diagnosed with schizophrenia? What are the nature of the antisaccade deficits in this disorder? Are these primary due to lack of inhibitory control, attention, or impediments in the generation of voluntary actions? Are there different patient subgroups characterized by different deficiencies that could potentially be relevant for treatment or outcome prediction?

The last questions are particularly important in a translational context, as deficits in the antisaccade task have been demonstrated to constitute an endophenotype of schizophrenia (Radant et al., 2015). Unfortunately, in the absence of computational models, our understanding of this endophenotype is restricted to a purely phenomenological characterization that might mask differences among patients and hide the connections to other deficits in working memory or attention, for example.

In summary, the next step in the research agenda proposed in this dissertation is to characterize the deficits of diagnosed patients in terms of computational quantities such as the probability of early and late

errors. Hence, the two research lines that we consider most relevant are the following: (i) are the high number of errors in the antisaccade task in schizophrenia mostly due to poor inhibitory control *or* weak voluntary action initiation, and (ii) are there subgroups characterized by the probability of each type of error. Answering these questions could lead to longitudinal studies that examine the connection of this endophenotype with disease progression, allowing us to make clinically relevant predictions such as treatment response and relapse rate on a subject-specific level.

Bibliography

Adams RA, Bauer M, Pinotsis D, Friston KJ. Dynamic causal modelling of eye movements during pursuit: Confirming precision-encoding in V1 using MEG. *Neuroimage* 132: 175–189, 2016.

Adams RA, Stephan KE, Brown HR, Frith CD, Friston KJ. The computational anatomy of psychosis. *Front Psychiatry* 4: 47, 2013.

Allport A, Styles EA, Hsieh S. Shifting Intentional Set: Exploring the Dynamic Control of Tasks. In: *Attention and performance series. Attention and performance 15: Conscious and nonconscious information processing*, edited by Umilta C, Moscovitch M. Cambridge, MA, US: The MIT press, 1994, p. 266–290.

Bach DR, Daunizeau J, Friston KJ, Dolan RJ. Dynamic causal modelling of anticipatory skin conductance responses. *Biol Psychol* 85: 163–170, 2010.

Ballnus B, Hug S, Hatz K, Gorlitz L, Hasenauer J, Theis FJ. Comprehensive benchmarking of Markov chain Monte Carlo methods for dynamical systems. *BMC Syst Biol* 11: 63, 2017.

Barton JJ, Greenzang C, Hefter R, Edelman J, Manoach DS. Switching, plasticity, and prediction in a saccadic task-switch paradigm. *Exp Brain Res* 168: 76–87, 2006.

Ben Calderhead, Girolami MA. Estimating Bayes factors via thermodynamic integration and population MCMC. *Computational Statistics \& Data Analysis* 53: 4028–4045, 2009.

Bittencourt J, Velasques B, Teixeira S, Basile LF, Salles JI, Nardi AE, Budde H, Cagy M, Piedade R, Ribeiro P. Saccadic eye movement applications for psychiatric disorders. *Neuropsychiatr Dis Treat* 9: 1393–1409, 2013.

Brodersen KH, Schofield TM, Leff AP, Ong CS, Lomakina EI, Buhmann JM, Stephan KE. Generative embedding for model-based classification of fMRI data. *PLoS Comput Biol* 7: e1002079, 2011.

Cherkasova MV, Manoach DS, Intriligator JM, Barton JJ. Antisaccades and task-switching: interactions in controlled processing. *Exp Brain Res* 144: 528–537, 2002.

Cutsuridis V. Behavioural and computational varieties of response inhibition in eye movements. *Philos Trans R Soc Lond, B, Biol Sci* 372, 2017.

Deserno L, Schlagenhauf F, Heinz A. Striatal dopamine, reward, and decision making in schizophrenia. *Dialogues Clin Neurosci* 18: 77–89, 2016.

Frassle S, Yao Y, Schobi D, Aponte EA, Heinzle J, Stephan KE. Generative models for clinical applications in computational psychiatry. *Wiley Interdiscip Rev Cogn Sci*.

Friston K, Mattout J, Trujillo-Barreto N, Ashburner J, Penny W. Variational free energy and the Laplace approximation. *Neuroimage* 34: 220–234, 2007.

Friston KJ, Harrison L, Penny W. Dynamic causal modelling. *Neuroimage* 19: 1273–1302, 2003.

Friston KJ, Lawson R, Frith CD. On hyperpriors and hypopriors: comment on Pellicano and Burr. *Trends Cogn Sci (Regul Ed)* 17: 1, 2013.

Gaymard B, Rivaud S, Pierrot-Deseilligny C. Impairment of extraretinal eye position signals after central thalamic lesions in humans. *Exp Brain Res* 102: 1–9, 1994.

Haker H, Schneebeli M, Stephan KE. Can Bayesian Theories of Autism Spectrum Disorder Help Improve Clinical Practice? *Front Psychiatry* 7: 107, 2016.

Hallett PE. Primary and secondary saccades to goals defined by instructions. *Vision Res* 18: 1279–1296, 1978.

Hikosaka O, Isoda M. Switching from automatic to controlled behavior: cortico-basal ganglia mechanisms. *Trends Cogn Sci (Regul Ed)* 14: 154–161, 2010.

Hutton SB, Ettinger U. The antisaccade task as a research tool in psychopathology: a critical review. *Psychophysiology* 43: 302–313, 2006.

Isoda M, Hikosaka O. Cortico-basal ganglia mechanisms for overcoming innate, habitual and motivational behaviors. *Eur J Neurosci* 33: 2058–2069, 2011.

Jaafari N, Rigalleau F, Rachid F, Delamillieure P, Millet B, Olie JP, Gil R, Rotge JY, Vibert N. A critical review of the contribution of eye movement

recordings to the neuropsychology of obsessive compulsive disorder. *Acta Psychiatr Scand* 124: 87–101, 2011.

Jablensky A. The diagnostic concept of schizophrenia: its history, evolution, and future prospects. *Dialogues Clin Neurosci* 12: 271–287, 2010.

Jamadar SD, Fielding J, Egan GF. Quantitative meta-analysis of fMRI and PET studies reveals consistent activation in fronto-striatal-parietal regions and cerebellum during antisaccades and prosaccades. *Front Psychol* 4: 749, 2013.

Jaynes ET. Information theory and statistical mechanics. *Physical review* 106: 620, 1957.

Jersild AT. Mental set and shift. *Archives of psychology*.

Kass RE, Raftery AE. Bayes factors. *Journal of the american statistical association* 90: 773–795, 1995.

Kiebel SJ, Garrido MI, Moran R, Chen CC, Friston KJ. Dynamic causal modeling for EEG and MEG. *Hum Brain Mapp* 30: 1866–1876, 2009.

Kiesel A, Steinhauser M, Wendt M, Falkenstein M, Jost K, Philipp AM, Koch I. Control and interference in task switching--a review. *Psychol Bull* 136: 849–874, 2010.

Kirkwood JG. Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics* 3: 300–313, 1935.

Krystal JH, D'Souza DC, Mathalon D, Perry E, Belger A, Hoffman R. NMDA receptor antagonist effects, cortical glutamatergic function, and schizophrenia: toward a paradigm shift in medication development. *Psychopharmacology (Berl)* 169: 215–233, 2003.

Maes J, Van Gool AR. Misattribution of agency in schizophrenia: An exploration of historical first-person accounts. *Phenomenology and the Cognitive Sciences* 7: 191–202, 2008.

McLaughlin S. Parametric adjustment in saccadic eye movements. *Perception \& Psychophysics* 2: 359–362, 1967.

Monsell S. Task switching. *Trends Cogn Sci (Regul Ed)* 7: 134–140, 2003.

Montague PR, Dolan RJ, Friston KJ, Dayan P. Computational psychiatry. *Trends Cogn Sci (Regul Ed)* 16: 72–80, 2012.

Munoz DP, Everling S. Look away: the anti-saccade task and the voluntary control of eye movement. *Nat Rev Neurosci* 5: 218–228, 2004.

Myles JB, Rossell SL, Phillipou A, Thomas E, Gurvich C. Insights to the schizophrenia continuum: A systematic review of saccadic eye movements in schizotypy and biological relatives of schizophrenia patients. *Neurosci Biobehav Rev* 72: 278–300, 2017.

Pellicano E, Burr D. When the world becomes "too real": a Bayesian explanation of autistic perception. *Trends Cogn Sci (Regul Ed)* 16: 504–510, 2012.

Penny W, Sengupta B. Annealed Importance Sampling for Neural Mass Models. *PLoS Comput Biol* 12: e1004797, 2016.

Radant AD, Dobie DJ, Calkins ME, Olincy A, Braff DL, Cadenhead KS, Freedman R, Green MF, Greenwood TA, Gur RE, Gur RC, Light GA, Meichle SP, Millard SP, Mintz J, Nuechterlein KH, Schork NJ, Seidman LJ, Siever LJ, Silverman JM, Stone WS, Swerdlow NR, Tsuang MT, Turetsky BI, Tsuang DW. Antisaccade performance in schizophrenia patients, their first-degree biological relatives, and community comparison subjects: data from the COGS study. *Psychophysiology* 47: 846–856, 2010.

Radant AD, Millard SP, Braff DL, Calkins ME, Dobie DJ, Freedman R, Green MF, Greenwood TA, Gur RE, Gur RC, Lazzeroni LC, Light GA, Meichle SP, Nuechterlein KH, Olincy A, Seidman LJ, Siever LJ, Silverman JM, Stone WS, Swerdlow NR, Sugar CA, Tsuang MT, Turetsky BI, Tsuang DW. Robust differences in antisaccade performance exist between COGS schizophrenia cases and controls regardless of recruitment strategies. *Schizophr Res* 163: 47–52, 2015.

Raman S, Deserno L, Schlagenhauf F, Stephan KE. A hierarchical model for integrating unsupervised generative embedding and empirical Bayes. *J Neurosci Methods* 269: 6–20, 2016.

Reilly JL, Frankovich K, Hill S, Gershon ES, Keefe RS, Keshavan MS, Pearlson GD, Tamminga CA, Sweeney JA. Elevated antisaccade error rate as an intermediate phenotype for psychosis across diagnostic categories. *Schizophr Bull* 40: 1011–1021, 2014.

Reilly JL, Lencer R, Bishop JR, Keedy S, Sweeney JA. Pharmacological treatment effects on eye movement control. *Brain Cogn* 68: 415–435, 2008.

Rigoux L, Stephan KE, Friston KJ, Daunizeau J. Bayesian model selection for group studies - revisited. *Neuroimage* 84: 971–985, 2014.

Schreber DP. *Memoirs of my nervous illness*. New York Review of Books, 1955.

Sengupta B, Friston KJ, Penny WD. Gradient-free MCMC methods for dynamic causal modelling. *Neuroimage* 112: 375–381, 2015.

Sengupta B, Friston KJ, Penny WD. Gradient-based MCMC samplers for dynamic causal modelling. *Neuroimage* 125: 1107–1118, 2016.

Sindhikara DJ, Emerson DJ, Roitberg AE. Exchange often and properly in replica exchange molecular dynamics. *Journal of chemical theory and computation* 6: 2804–2808, 2010.

Sommer MA, Wurtz RH. A pathway in primate brain for internal monitoring of movements. *Science* 296: 1480–1482, 2002.

Stephan KE, Baldeweg T, Friston KJ. Synaptic plasticity and dysconnection in schizophrenia. *Biol Psychiatry* 59: 929–939, 2006.

Stephan KE, Friston KJ, Frith CD. Dysconnection in schizophrenia: from abnormal synaptic plasticity to failures of self-monitoring. *Schizophr Bull* 35: 509–527, 2009a.

Stephan KE, Iglesias S, Heinzle J, Diaconescu AO. Translational Perspectives for Computational Neuroimaging. *Neuron* 87: 716–732, 2015.

Stephan KE, Mathys C. Computational approaches to psychiatry. *Curr Opin Neurobiol* 25: 85–92, 2014.

Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. Bayesian model selection for group studies. *Neuroimage* 46: 1004–1017, 2009b.

Stephan KE, Schlagenhauf F, Huys QJ, Raman S, Aponte EA, Brodersen KH, Rigoux L, Moran RJ, Daunizeau J, Dolan RJ, Friston KJ, Heinz A. Computational neuroimaging strategies for single patient predictions. *Neuroimage* 145: 180–199, 2017.

Sugita Y, Okamoto Y. Replica-exchange molecular dynamics method for protein folding. *Chemical physics letters* 314: 141–151, 1999.

Terao Y, Fukuda H, Ugawa Y, Hikosaka O. New perspectives on the pathophysiology of Parkinson's disease as assessed by saccade performance: a clinical review. *Clin Neurophysiol* 124: 1491–1506, 2013.

Trebst S, Troyer M, Hansmann UH. Optimized parallel tempering simulations of proteins. *J Chem Phys* 124: 174903, 2006.

Vehtari A, Lampinen J. Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Comput* 14: 2439–2468, 2002.

Vousden WD, Farr WM, Mandel I. Dynamic temperature selection for parallel tempering in Markov chain Monte Carlo simulations. *Monthly Notices of the Royal Astronomical Society* 455: 1919–1937, 2015.

Wang CA, Brien DC, Munoz DP. Pupil size reveals preparatory processes in the generation of pro-saccades and anti-saccades. *Eur J Neurosci* 41: 1102–1110, 2015.

Wang XJ, Krystal JH. Computational psychiatry. *Neuron* 84: 638–654, 2014.

Woodward TS, Bub DN, Hunter MA. Task switching deficits associated with Parkinson's disease reflect depleted attentional resources. *Neuropsychologia* 40: 1948–1955, 2002.