UNIVERSITY OF TRENTO

DOCTORAL THESIS

Cognitive Modeling of high-level cognition through Discrete State Dynamic processes

Author:
Marco D'ALESSANDRO

Supervisor:
Dr. Luigi LOMBARDI

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Department of Psychology and Cognitive Science

January 11, 2021

Declaration of Authorship

I, Marco D'ALESSANDRO, declare that this thesis titled, Cognitive Modeling of high-level cognition through Discrete State Dynamic processes and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

UNIVERSITY OF TRENTO

Abstract

Department of Psychology and Cognitive Science

Doctor of Philosophy

Cognitive Modeling of high-level cognition through Discrete State Dynamic processes

by Marco D'ALESSANDRO

Modeling complex cognitive phenomena is a challenging task, especially when it is required to account for the functioning of a cognitive system interacting with an uncertain and changing environment. Psychometrics offers an heterogeneous corpus of computational tools to infer latent cognitive constructs from the observation of behavioural outcomes. However, there is not an explicit consensus regarding the optimal way to properly take into account the intrinsic dynamic properties of the environment, as well as the dynamic nature of cognitive states. In the present dissertation, we explore the potentials of relying on discrete state dynamic models to formally account for the unfolding of cognitive sub-processes in changing task environments. In particular, we propose Probabilistic Graphical Models (PGMs) as an ideal and unifying mathematical language to represent cognitive dynamics as structured graphs codifying (causal) relationships between cognitive sub-components which unfolds in discrete time. We propose several works demonstrating the advantage and the representational power of such a modeling framework, by providing dynamic models of cognition specified according ti different levels of abstraction.

Acknowledgements

First of all, I would like to thank my supervisor, Prof. Luigi Lombardi, for his patience and guidance, for working hard to fix all my bad habits in scientific writing and thinking, and for always being available to discuss about topics beyond the scope of research and PhD career. I also want to thank my special friends and colleagues Antonino Greco, Stefan Radev, and Giuseppe Gallitto, which proven to be a continuous source of insight for the most diverse scientific subjects. My final thanks goes to my fiancee, Laura, which was crucial for the entire doctoral journey, for putting up with me being sat at my desk for hours and hours, and for supporting me every time I was about to give up.

Contents

Chapter 1

Introduction

Computational cognitive modeling allows to test and design detailed, process-based mathematical models of how cognitive agents (e.g. human beings) internally represent and manipulate information, and produce behaviour (Sun, 2008; Farrell and Lewandowsky, 2018). The main advantage of such a formal approach to the study of cognition is that hypothesis about the interactions between cognitive processes and environmental contingencies can be precisely instantiated as computational models, which serve as a basis for simulating behaviour and investigating how meaningful cognitive parameters might affect response patterns. In general, computational modeling comes into play whenever a researcher is interested in capturing regularities in the data using parameters that represent separate statistical or psychological processes. This can be thought of as a method which overcomes the limitations of relying on summary statistics of individual behavioural outcomes to make inference about cognitive functioning.

However, as more complex cognitive phenomena are taken into account, more elaborate mathematical models, as well as general mathematical frameworks, are needed. This is often the case of high-level cognitive functions, such as those entailing learning, executive processes and decision-making, which are required in changing (or dynamic) and uncertain environments. In these contexts, cognitive agents are demanded to continuously adapt as new information are gathered from the (internal or external) environment, and distinct psychological components might contribute to reach an optimal behaviour.

Many experimental, or clinical, (neuro)psychological settings widely employed in cognitive research entail such a dynamic component. In particular, dynamic tasks aimed to investigate high-level cognition, such as the Wisconsin Card Sorting Test (Heaton, 1981; Berg, 1948), the Iowa Gambling Task (Bechara et al., 2001; Bechara and Damasio, 2002), the Balloon Analogue Risk Task (Lejuez et al., 2002; Lejuez et al., 2003), mouse-tracking tasks (Freeman and Ambady, 2010), or multi-alternative forced choice tasks (Krajbich and Rangel, 2011) among others, can be thought of as tools which capture dynamics in behavioural outcomes based on both (internal or external) feedback received and the configuration of environmental states which changes at each trial.

In this work, we explore the potentials of relying on discrete state dynamic modeling to formally account for the unfolding of psychological sub-processes in changing task environments, by assuming that evolving response patterns

which can be observed result from an underlying cognitive (or brain) state process which unfolds in time trialwise. We can formalize this concept by endorsing the view of cognitive agents as dynamical systems (Van Gelder, 1998), where such systems are governed by rules specifying the relationship between the current (cognitive) state, the trial-by-trial unfolding of the task, and the transitions between current states and new states. Furthermore, since we want to deal with uncertain environments, and more generally with cognitive agents processing partial and uncertain information, we assume that such a relationship is expressed in probabilistic terms.

The idea of modeling cognition by relying on probability models is certainly not new. Several pioneering works stressed the need to switch from symbolic rule-based and non-monotonic logical approaches to the study of cognition to structured probabilistic representations of cognitive processes (Chater et al., 2006; Griffiths et al., 2010; Tenenbaum et al., 2006; Ma, 2012). In recent years, such a paradigm shift has been motivated by two main scientific results: (1) the finding that most of the cognitive and motor processes cognitive scientists are interested in, implement probabilistic computations (Buckley et al., 2017; Yuille and Kersten, 2006; Pouget et al., 2013; Friston and Kiebel, 2009; Glaser et al., 2018); (2) the recent advances in mathematical and computer science techniques in information theory, stochastic processes and machine learning, which provided a basis for more exhaustive and realistic formal representations of complex cognitive systems (Yuille and Kersten, 2006; Friston, 2010; Friston et al., 2017a; Stoianov et al., 2016; Stoianov and Zorzi, 2012).

However, the probabilistic modeling approach proposed in this dissertation aims to embody the Cognitive Psychometrics (or model-based Psychometrics) perspective (Erdfelder et al., 2020; Batchelder, 2010), by allowing mathematical cognitive modeling and psychometric assessment to co-exist in a unified framework. In principle, this means balancing complexity and assessment capability of the computational cognitive model. In other words, we intend to find a reasonable compromise between models' neural (or physical) plausibility and models' parsimony and interpretability, while being consistent with a sophisticated probabilistic representation of cognition.

To do so, we rely on the mathematical framework of Probabilistic Graphical Models (PGM, (Koller and Friedman, 2009)), which are (possibly) hierarchically-organized probabilistic graphs expressing conditional dependencies between variables (e.g. psychological variables). Such a framework offers a perfect mathematical language to model cognition, since it allows to flexibly represent the causal (phenomenological) relationship between psychological variables and behaviour, as well as the relationship between psychological variables themselves at different time periods. The potentials of such a framework emerge, in particular, when a structured probabilistic graph is coupled with a set of interpretable cognitive parameters which directly shape the relationship between the variables in the system, allowing to adopt parameters as a proxy to assess cognitive performance, at an individual or a group-level of analysis.

In the following dissertation, we will often refer to the convenient concept of hidden (or latent) state process. In our context, an hidden state process can

be intended either as a set of variables indicating the true (possibly psychological) constructs underlying observations which are just a noisy measure of such constructs in precise statistical terms (e.g. (Zucchini et al., 2008; Bartolucci et al., 2012; Yousefi et al., 2019)), or as a (neuro)biologically plausible process assumed to play a role at higher levels of the hierarchy of a generative model which accounts for how a cognitive system behaves in a certain environment (e.g. (Friston et al., 2017a; Friston et al., 2017b; Mathys et al., 2014; Stoianov et al., 2016)).

Furthermore, a hidden state might be also conceived as a particular characteristic of the environment that a cognitive agent needs to infer to fulfill a task requirement. However, it is worth noticing that such a differentiation between environmental and cognitive hidden states strictly depends on the purpose of the modeling and the structure of the task being modeled, as will become clearer during the exposure of the works in the dissertation.

In the remainder of the introduction, we will give a brief overview of the computational modeling problems to be addressed in the coming chapters. The present dissertation encompasses a collection of diverse but related works; what makes such works blended together is the commitment to mathematical modeling of complex behaviour within a PGM framework, with a particular focus on discrete-time dynamic modeling.

1.1 Chapter Outline

In Chapter 2, a brief introduction to PGMs is provided, as well as a treatment of the main model's architectures which are used throughout the work. Although PGMs offer a consistent discrete state representation which is shared among the computational modeling studies, the proposed dynamic models of cognition are specified according to different levels of abstraction.

In Chapter 3 we will show the potentials of relying on dynamic latent class models designed for longitudinal-like data structure to improve the assessment of executive functions of individuals performing a set-shifting task. More precisely, scoring measures obtained from the analysis of the WCST responses are further processed in a dynamic computational framework. An intriguing feature of the WCST experimental protocols is that it allows the performance of an individual to change as the task unfolds. In this work, a Latent Markov Model is proposed to capture some dynamic aspects of observed response patterns in both healthy and substance dependent individuals. The main goal is to parameterize performance trends in terms of transition dynamics of latent (cognitive) states. The results highlighted how a dynamic modelling approach can considerably improve the amount of information a researcher, or a clinician, can obtain from the analysis of a set-shifting task. Here, cognitive functions are not modelled explicitly via structured graphs. Instead, latent states are thought to provide information on cognitive dynamics on a more abstract level, by entailing psychologically interpretable latent classes related to constructs such as the response strategy.

Differently, the PGM adopted in Chapter 4 directly deals with the problem of explicitly modeling the relationship between cognitive sub-components

which yield behavioural dynamics in a set-shifting framework. In particular, the work proposed here inherits the leading idea from the previous study, that is, the representation of the performance trend as a parameterized latent (cognitive) state dynamic process. Here, the main interest is exploring how cognitive (sub)processes taking place in a set-shifting context produce meaningful human response patterns. In this contribution, we propose a Dynamic Bayesian Network-inspired cognitive model to provide a formal account of the problem. We adopt a simulation approach to study how the behaviour of our cognitive model evolves during the dynamic unfolding of the task. Preliminary results show how the proposed model can nicely account for complex behavioural patterns as assessed by standard WCST scoring measures and can provide a way to explore how selective cognitive impairments affect observed response patterns. Practical and theoretical implications of our computational modelling approach for clinical and psychological sciences are finally discussed, as well as its possible future improvements.

In Chapter 5 we show how it is possible to instantiate a particular theory of brain functioning within a discrete state dynamic framework. In particular, the main purpose of the third work is to model adaptive behavior as an emerging factor derived from the interaction between cognitive agents and changing environmental demands. To do so, we use the WCST task environment, and we propose to model cognitive dynamics within the mathematical framework of Bayesian Brain Theory, according to which beliefs about the hidden environmental states (e.g. some rule of the game an agent has to infer) are dynamically updated following the logic of Bayesian inference. Our computational model maps distinct cognitive processes into separable, neurobiologically plausible, information-theoretic constructs underlying observed response patterns. We assess model identification and expressiveness in accounting for meaningful human performance through extensive simulation studies. We further apply the model to real behavioral data in order to highlight the utility of the proposed model in recovering cognitive dynamics at an individual level. Practical and theoretical implications of our computational modelling approach for clinical and cognitive neuroscience research are finally discussed.

Finally, the fourth work presented in Chapter 6 will show a PGM approach to the modeling of decision-making under risk, together with a method to analyze both neural and behavioural data. Understanding dependencies between brain functioning and cognition can be, indeed, a challenging task which might require more than applying standard statistical models to neural and behavioural measures. Recent developments in computational modelling have demonstrated the advantage to formally account for reciprocal relations between mathematical models of cognition and brain functional, or structural, characteristics to relate neural and cognitive parameters on a model-based perspective. This would allow to account for both neural and behavioural data simultaneously by providing a joint probabilistic model for the two sources of information. In the present work we proposed a PGM architecture for jointly modelling the reciprocal relation between behavioural and neural information. More precisely, we offered a way to relate Diffusion Tensor Imaging data to cognitive parameters of a computational dynamic model

accounting for behavioural outcomes in the popular Balloon Analogue Risk Task (BART). Results show that the proposed architecture has the potential to account for individual differences in task performances and brain structural features by letting individual-level parameters to be modelled by a joint distribution connecting both sources of information.

The final chapters will be dedicated to the discussion of the implication of discrete state dynamic modeling in cognitive research (Chapter 7)..

Chapter 2

Theoretical background

In this section, we provide an essential description of the main features of a PGM, and the intuitions behind its potential to represent (psychological) events which unfolds in a discrete state space and in discrete time.

Probabilistic graphical models consist in a powerful formalism which allows to encode relationships between multiple variables. In particular, they combine the mathematical rigor of probability theory with the flexibility of graph theory. Here, we focus on one of the most common type of PGM, that is, the one consisting in a probabilistic causal model represented by a directed acyclic graph (DAG, (Thulasiraman and Swamy, 2011; Koller and Friedman, 2009)). A DAG is specified by a graph $G(V, E)$, where V and E are the set of nodes (vertices) and edges, where each edge is directed from one vertex to another, and it is assumed that starting from a given node, there is no path crossing other nodes which eventually loops back to the starting node again (Murphy and Russell, 2002). Such a structure is also commonly referred to as a Bayesian Network. In this framework, nodes represent random variables which can be either hidden or observed. Edges encode the direct dependencies between the random variables. The whole structure represents the conditional independence assumptions which determine a given factorization of the joint probability distribution of the system. The joint distribution defined by a graph is given by the product, over all the nodes in the graph, of a conditional distribution for each node conditioned on the variables corresponding to the parents of that node (Bishop, 2006). Thus, for a graph with K nodes, the joint distribution is given by:

$$P(x_1, \ldots, x_K) = \prod_{k=1}^{K} P\big(x_k | \mathrm{pa}(x_k)\big)$$

where $\mathrm{pa}(x_k)$ denotes the set of parents of x_k. In the simplest case, one might think of the structuring of nodes and related parents on a graph both as an assumption of how observed variables influence each other, and as a way to organize knowledge about a phenomenon to state whether information on a given variable can, or can not, be obtained conditioned on the knowledge about other variables in the system. The way in which we factorize the joint probability distribution allows to obtain several configuration of knowledge and causal principles. As an example, consider a trivial case in which we want to structure knowledge about the relationship between the

psychometric variables *intelligence* (x_1) and *extraversion* (x_2), the *graduation grade* (x_3), and the *job salary* (x_4). Given a finite set of variables, or nodes, $V \in \{x_1, x_2, x_3, x_4\}$, and connections, or edges E, between them, several ways of structuring knowledge can be conceived. A realistic scenario consists in modeling relationships between variables by assuming that both intelligence and extraversion contribute to graduation grade, which, in turn, exerts an effect on the job salary. The structure of the model is depicted in Figure 2.1a, and can be simply factorized as follows:

$$P(x_1, x_2, x_3, x_4) = P(x_1)P(x_2)P(x_3|x_1, x_2)P(x_4|x_3)$$

and reflects our assumptions about how the information propagates from the nodes at the top of the graph to some target node at lower levels. Specifying a factorization of the joint distribution is also a fundamental step to assess which information might be gathered by knowing the status of another variable in the system. For instance, given that the graduation grade is observed, it is possible to take advantage of the (in)dependence assumption to gain information about intelligence by having knowledge about extraversion (Figure 2.1b). Thus, independent causes (e.g. intelligence and extraversion) are made dependent by conditioning on a common effect (Pearl, 2014).

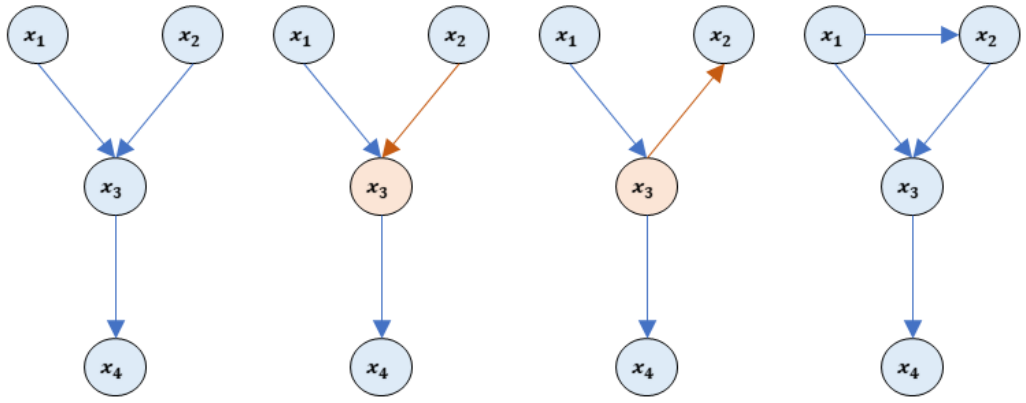


FIGURE 2.1: Examples of Bayesian Networks with different conditional (in)dependence assumptions.

However, this property does not hold if we take into account a different factorization of the joint probability, that is, different assumptions in structuring knowledge and relationships between variables. The graph in Figure 2.1c does not assume a common effect of intelligence and extraversion on graduation grade. Instead, it depicts a forward information propagation from intelligence to extraversion, through graduation grade:

$$P(x_1, x_2, x_3, x_4) = P(x_1)P(x_2|x_3)P(x_3|x_1)P(x_4|x_3).$$

In this scenario, observing graduation grade, makes intelligence and extraversion independent. That is, the path from intelligence to extraversion is blocked by conditioning on the grade graduation, a phenomenon known as *d-separation* (Koller and Friedman, 2009). It is worth noticing that, although probabilistic relationships are specified over the same set of random variables, the implications of conditional dependence assumptions may lead to different interpretation and results. Therefore, more complex configurations can be considered, by, for instance, relaxing the independence assumption between intelligence and extraversion, as shown in Figure 2.1d, as follows:

$$P(x_1, x_2, x_3, x_4) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)P(x_4|x_3).$$

In general, there is no limitations in the complexity of relationships which can be defined, neither in the type of object which can be codified as a random variable in the DAG. Therefore, relationships between nodes can be flexibly modeled by parameterized mathematical expressions, as well as common probability density functions. Understanding the advantage of such a modeling framework to simulate cognitive phenomena and behaviour is straightforward, since the information propagation assumptions in the graph can be mapped to hierarchically organized information propagation in a cognitive system (Alexander and Brown, 2018; Badcock et al., 2019).

In the present dissertation, we will put the emphasis on the usage of dynamic models (e.g. dynamic version of DAGs) to represent cognitive functioning. In particular, we refer to Bayesian Networks replicated through time (Barber, 2012), yielding what is commonly referred to as Dynamic Bayesian Networks (Murphy and Russell, 2002; Ghahramani, 1997). In this case, nodes in the DAG still codify random variables, whilst edges represent probabilistic dependencies between variables across time. A key assumption is that the probability distributions, or mathematical laws, describing the temporal dependencies between the nodes are time invariant.

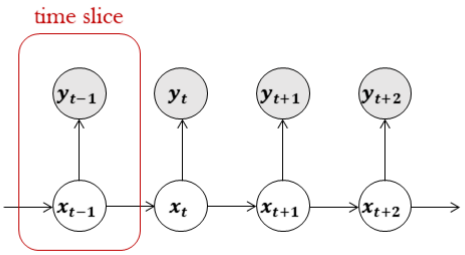
Given a set of K nodes at each time-slice t, namely, $\mathbf{x}_t$, we can define the joint probability model of $\mathbf{x}_{1:t}$ as follows:

$$P(\mathbf{x}_1, \ldots, \mathbf{x}_T) = \prod_{t=1}^{T} \prod_{k=1}^{K} P\big(x_{k,t} | \text{pa}(x_{k,t})\big)$$

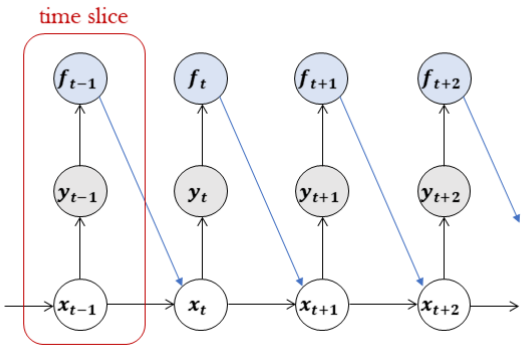
where $\text{pa}(x_{k,t})$ denotes the set of parents of x_k at time t. In a first-order DBN, the set of parental variables of $x_{k,t}$ is taken from the previous time-slice or from the current time-slice (Barber, 2012). Due to their properties, DBNs are useful to formalize an evolving system that changes in time, by describing how variables influence each other within and across time periods. In general, a DBN can be seen as an umbrella term which encompasses a variety of stochastic models for time series, and longitudinal data, such as those related to the analysis of fully observed sequential data (e.g. autoregressive models) and those involving an hidden state space (e.g. Hidden Markov Models and State-Space Models) (Dagum et al., 1991). However, irrespective of the

particular instance considered, a DBN modeling procedure must fulfill a requirement to allow mathematical tractability. The temporal structure of the event needs to be simplified in order to discretize the timeline into a set of time slices by ensuring that relevant attributes of the system state can be captured by the snapshot of a given time slice. This requires both evaluating whether the evolving nature of the phenomenon can be thought of as partitioned in meaningful time slices, and determining the equally spaced intervals to define time granularity. The latter requirement depends on both the phenomenon under investigation and the specific modeling choice (e.g. how finely the state of the system has to be monitored in time). When dealing with these models, we will also rely on further simplification, that is, the Markov assumption (Dobrow, 2016). It states that variables in $\mathbf{x}_{t+1}$ cannot depend directly on variables in $\mathbf{x}_{t^*}$, where $t^* < t$. From a strict DBN perspective, the Markov assumption corresponds to a graph structure in which there are no edges into $\mathbf{x}_{t+1}$ coming from variables in time slices $t-1$ or earlier. Such an assumption can be considered as a sufficiently reasonable approximation to the dependencies in our joint probability model (Koller and Friedman, 2009).

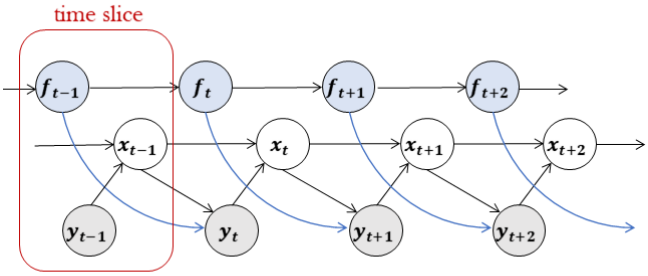
A dynamic model can now be represented by an initial state distribution, describing the (possible) state of the system at the beginning of the process, a transition probability model between the (hidden) states, and an eventual observation model describing how the time series of observations emerges from an underlying (possibly latent) state process. A variety of dynamic model's architectures can be conceived based on such ingredients (Figure 2.2).



(A)



(B)



(C)

FIGURE 2.2: Examples of dynamic model's architectures with different conditional (in)dependence assumptions.

During the dissertation, we will treat several PGM architectures, each endowed with different dynamical properties, and built upon different knowledge representation strategies. The discrete time approximation which yields the temporal evolution of the system's state will consist in treating single trials as the basic building blocks of a time slice. Variables in the system are rephrased in order to account for psychological constructs, or interacting cognitive sub-components. Thus, we recommend the reading of the following chapters to get an exhaustive overview of the different architectures, as well as the problem-specific joint probability factorizations, employed.

Chapter 3

A Dynamic latent state approach to model set-shifting performances

The content of the chapter has been in part published as: D'Alessandro, M., & Lombardi, L. (2019). A dynamic framework for modelling set-shifting performances. *Behavioral Sciences*, 9, 79.

3.1 Introduction

In recent years there has been an increasing interest in modelling behavioural data from experimental tasks aimed at investigating higher-level cognitive functioning (Dehaene and Changeux, 1991; Busemeyer and Stout, 2002; Yechiam et al., 2006; Hull et al., 2008; Bartolucci and Solis-Trapala, 2010; Bishara et al., 2010). Generally, higher-order cognitive functions can be seen as a class of cognitive processes which are crucial in situations requiring to flexibly adjust behaviour in order to correspond to changing environmental demands (Zelazo et al., 2003), and to integrate previous experience and feedback in order to maximize optimal choices (Busemeyer et al., 2003). Deficits at this level of cognitive functioning can be observed in rather heterogeneous clinical populations (Bechara and Damasio, 2002; Braff et al., 1991), each characterized, ideally, by a different pattern of impaired psychological sub-processes (see for example (Yechiam et al., 2006)). Cognitive evaluation of these functions mostly relies on neuropsychological tasks that endorse a dynamic, or longitudinal, aspect requiring a person to change hisher actions over the course of the task. Consider, for example, a general context in which participants may learn to pay attention to the correct stimulus while ignoring irrelevant stimuli as a function of experimental feedback. Here, negative feedback should allow participants to conceive a given stimulus as irrelevant, modifying their responses accordingly. In this context, observed response patterns could consist of the occurrences of casual errors, feedback-related errors, and perseverations on shifting tasks, to name a few (e.g., (Heaton, 1981)). The basic idea is that these response patterns reflect the presence (or the absence) of a cognitive impairment, either at a functional or neural level (Buchsbaum et al., 2005).

In this work we suggest a latent variable approach to model cognitive performances on a standard set-shifting task from a group level perspective. The method is applied to data from the Wisconsin Card Sorting Test (WCST;

(Heaton, 1981; Berg, 1948)), which offers a renowned tool to measure set-shifting, defective inhibitory processes on the basis of environmental feedback in cognitive settings (Demakis, 2003). In general, the test consists of a target and a series of stimulus cards with geometric figures that differ according to three perceptual characteristics. The task demands that participants recognize the correct classification principle by means of trials and error, and the feedback of the examiner. An intriguing, and underestimated, aspect of such experimental protocol regards how individual performances change as the task unfolds, plausibly due to "learning to learn" capacity (Tarter, 1973) or shifting cognitive strategies (Berg, 1948). According to our view, a formal analysis of this *performance trend* (see, for example, (Tarter, 1973)) can provide a novel interesting metric for the cognitive assessment of test outcomes. Therefore, several dynamic models regarding decision-making (Dai et al., 2018), learning (Gershman, 2015), risky behaviour (Wallsten et al., 2005) and categorization (Kruschke, 2008), have proven to be able in uncovering characteristics of cognitive functioning which could not be detected with a standard (static) analytic approach based on collecting summary measures of individuals' responses.

In order to formally account for task dynamics, we adopted a Latent Markov Model (LMM; (Wiggins, 1973; Bartolucci et al., 2012)) perspective to assess the evolution of a latent states process underlying the observed behaviour. The basic assumption is that participants may evolve in their latent characteristic/states during the unfolding of the task. Thus, rather than simply analysing how the observed responses configuration evolves during the unfolding of the task, our target becomes to model the entire evolution of the latent states underlying these responses. The idea that observed behaviour is the final result arising from two or more latent data-generating states is clearly not new (e.g., (Smallwood and Schooler, 2015; Hawkins et al., 2017)). Here, the basic intuition consists in the fact that human cognition can be influenced not only by external task demands, which are usually known and observable, but also by unknown latent mental processes and brain states that dynamically change with time (Taghia et al., 2018). In our context, the observation of a changing pattern in the response of the participants (e.g. an increase in the frequency of persevering errors at a given stage of the task) will indicate the fact that there has been a change in the latent process. The Latent Markov Model captures this insight beautifully by considering the observed responses as a measure of the underlying latent states, and subsequently by providing a coherent account of the dynamics of the latent states process. The reader is referred to (Bishara et al., 2010) and (Speekenbrink et al., 2010) for different model-based approaches to model cognitive phenomena related to that analysed in the present work.

The chapter is organized as follows. The next section provides the basic features of the LMM framework. In the third section, a model application is presented on a real data set collected using the WCST in the context of substance addiction. Finally, the fourth section presents a brief discussion and some conclusions.

3.2 Materials and Methods

3.2.1 The formal framework

Latent Markov models (LMM) can be seen as a generalization of latent class models, which were developed for the analysis of longitudinal data, in particular categorical response variables (Wiggins, 1973; Pennoni, 2014). To do so, these models make use of time-specific latent variables which are assumed to be discrete. Below, we briefly outline the main properties of this modelling approach.

To begin with, consider the directed graph depicted in Figure 3.1 illustrating the logic behind a basic LMM which evolves across T discrete time steps (e.g., (Wiggins, 1973; Bartolucci et al., 2012)).

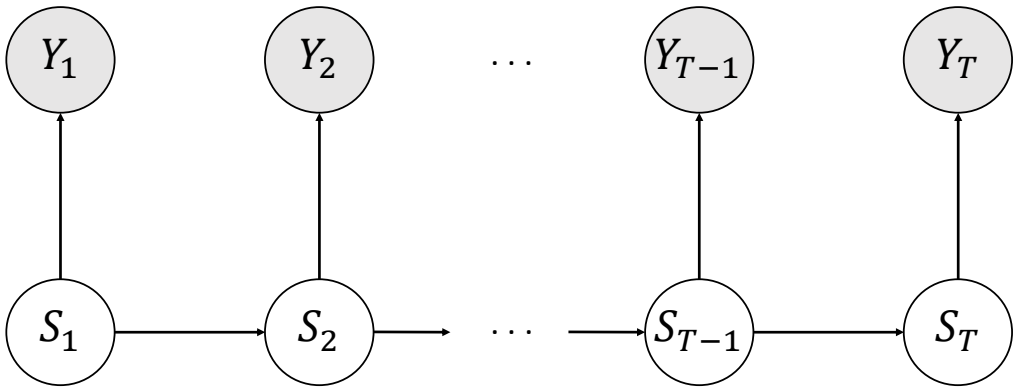


FIGURE 3.1: Conditional (in)dependencies structure of a basic LMM. Shaded nodes represent observed variables. White nodes represent unobserved (latent) variables.

In a LMM, it is assumed that a sequence of observed response variables, $Y_1, Y_2, \ldots, Y_T$, are conditionally independent given a corresponding pairwise sequence of latent variables, $S_1, S_2, \ldots, S_T$, called states. More formally:

$$P(Y_1, Y_2, \ldots, Y_T | S_1, S_2, \ldots, S_T) = P(Y_1|S_1)P(Y_2|S_2) \cdots P(Y_T|S_T). \quad (3.1)$$

The lack of directional connections (directed arrows) between observed variable nodes reflects the idea that only the latent states dynamics are responsible for the response pattern observed across the entire task. In other words, the evolution of the observed responses in time can be (phenomenologically) considered as the result of transition dynamics between latent states. In particular, the latent process follows a first-order Markov chain in that the latent variable S_t at step t only depends on the outcome of the former step, S_{t-1},

(with $t = 2, \ldots, T$), thus yielding a memoryless process:

$$P(S_1, S_2, \ldots, S_T) = P(S_1) \prod_{t=2}^{T} P(S_t|S_{t-1}). \quad (3.2)$$

There are at least three key explanations why we find this modeling technique to be effective in the representation of cognitive behaviors observed in set-shifting tasks: (1) the latent states formally described by an LMM can be placed in relation with a certain observed behavioural outcome (Visser, 2011) (2) it is a method to shape and analyse the unfolding of behaviour in time and its relation with the evolving dynamic of some aspects of cognition (3) it has a clear probabilistic framework to examine how different intervening factors (e.g., observed covariates) might affect evolving behavioural outcomes. Understanding the advantages of such a general framework for modeling complex cognitive processes may be of great importance, as discrete latent states may well be correlated with other brain, cognitive, or abstract states that we believe may affect observed response patterns.

3.2.2 Model application

In this section, we present the proposed modelling approach to analyse participants' performances in the WCST and show how the LMM framework can account for differences between dynamic patterns in different experimental groups. To this purpose, we apply the model to the analysis of an already published dataset (see (Bishara et al., 2010) and (Bechara and Damasio, 2002)) which represents an ideal case study to investigate set-shifting performances.

Participants

In our study, we analysed responses of 38 substance dependent individuals (SDI) and 44 healthy individuals in the Wisconsin Card Sorting Test (*ibidem*). Control participants had no history of mental retardation, substance abuse, or any systemic central nervous system disease. Regarding the SDI, the Structured Clinical Interview for DSM-IV (First, 1997) was used to determine a diagnosis of substance dependence. All participants in the study were adults (¿18 years old) and gave their informed consent for inclusion which was approved by the appropriate human subject committee at the University of Iowa (see (Bechara and Damasio, 2002) for details).

Task procedure

In the common variant of the WCST, participants are presented a target card and a set of four stimulus cards. All the cards consist of geometric figures which differ in terms of three characteristics, namely, color (red, green, blue, yellow), shape (triangle, star, cross, circle) and number of objects (1, 2, 3 and 4). Figure 3.2 depicts an example of a standard WCST trial. For each trial, a participant is asked to sort the target card with one of the four stimulus

cards according to one of the three sorting rules. Each participant's response is followed by a feedback (either positive or negative) telling the individual whether the sorting is correct or incorrect. After a certain number of consecutive correct responses, the experimenter updates the sorting rule without any warning to the participant. Thus, for each trial, a correct response, a perseverative error, or a non-perseverative error can be observed (see (Heaton, 1981) for details). Undoubtedly, one of the most representative information attainable from the analysis of behavioural outcomes is that related to the occurrence of errors. Indeed, the error-related information has proven to be a reliable predictor of executive function deficits and frontal lobe dysfunctions in clinical population (e.g., (Nagahama et al., 2005; Miller et al., 2015)).

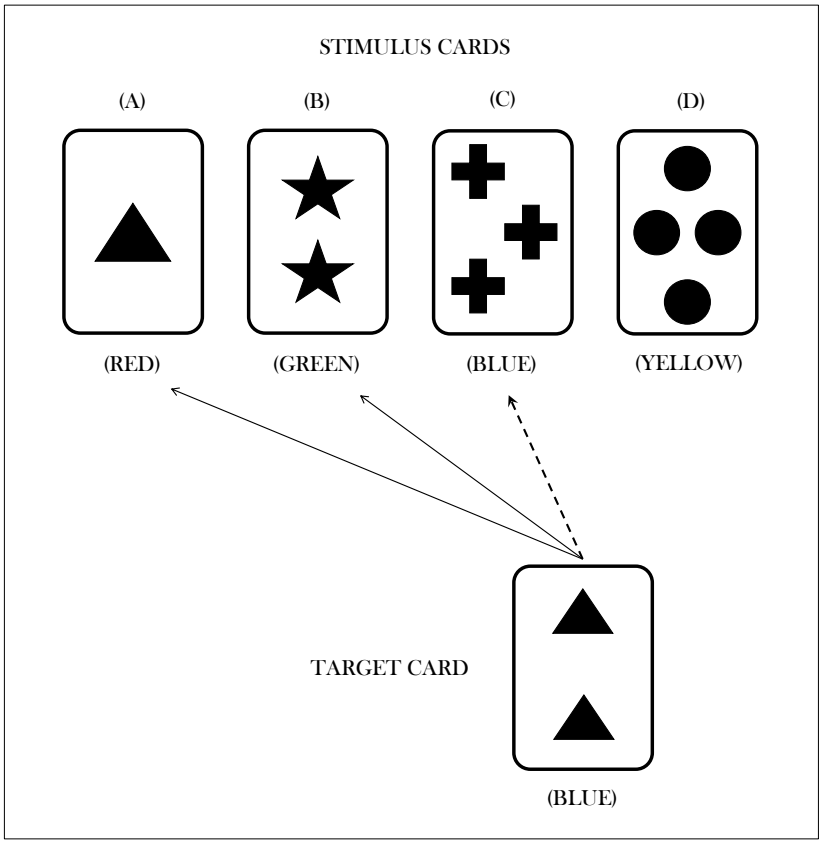


FIGURE 3.2: Example of a typical trial in the Wisconsin card sorting test. Arrows represent possible choices. In this example, the current sorting principle is color. Solid arrows, which sort the target card with stimulus cards (A) and (B), represent wrong matches. The dotted arrow, which sortes the target card with stimulus card (C), represents a right match.

Data modelling

In order to model the performance trends, we relied on the following data transformation procedure. First, we codified the observed sequence of participants' responses according to a popular neuropsychological (Flashman et al., 1991). In particular, we focused on three categories of responses: *correct responses* (C), *non-perseverative errors* (E), and *perseverative errors* (PE). As a further step, for each participant, we considered the entire response pattern as partitioned into a limited number of blocks, also defined as *windows of trials*. Our main purpose was to model the dynamics of participants' response patterns across these trial windows, rather than single trials. To this aim, in our application we considered for each participant five distinct windows which are thought to partition the entire task into (virtual) phases.

More precisely, let $Z^{(j)}$ be the vector of responses for the individual j, such that, $Z^{(j)} \in \{\mathrm{C}, \mathrm{E}, \mathrm{PE}\}$. The response vector is partitioned as follows:

$$Z^{(j)} = \left((z_1^{(j)}, \ldots, z_{n_j}^{(j)}), (z_{n_j+1}^{(j)}, \ldots, z_{2n_j}^{(j)}), \ldots, (z_{4n_j+1}^{(j)}, \ldots, z_{5n_j}^{(j)}) \right)$$

where the element $z_t^{(j)}$ reflects the individuals' codified response at trial t. The subscript n_j indicates the length of the task phase for individual j, and is calculated in order to obtain equally-sized trial windows. It is important to notice that subjects can vary in the number of observations within windows. The fact that subjects are not homogeneous in the number of trials which constitute each phase is not a matter of concern for our modelling aim, since the task phases are considered to reflect the percentage of progress in the task. At this point, the resulting data structure was organized according to a longitudinal design where a specific block, Y_t, consisted of all the observed responses aggregated across all participants for a specific task phase. As an example, consider the data vector for the time occasion $t = 1$, that is, for the first block of the longitudinal design. It consists of the aggregated responses of all participant's first task phases, and can be formally represented as:

$$Y_1 = \left((z_1^{(1)}, \ldots, z_{n_1}^{(1)}), (z_1^{(2)}, \ldots, z_{n_2}^{(2)}), \ldots, (z_1^{(J)}, \ldots, z_{n_J}^{(J)}) \right).$$

However, one might wonder whether the more natural way to capture changes in set-shifting performances could consist in organizing the longitudinal data structure by partitioning the vector $Z^{(j)}$ in order to take a specific number of trials after a change of the sorting rule occurs. However, individuals in our study differed in the number of trials achieved to complete a category, that is, before a change of the sorting rule occurs. A trial windows clustering based on selecting a specific number of trials after a change of the sorting rule does not ensure the regularity of the longitudinal structure, due to the individual variability in completing a category.

In our model, the windows were equally sized, and the choice of the number of windows, T, directly affected the number of trials within them. For this reason, some trials had to be excluded when the total number of trials

achieved by a given participant was not a multiple of T. There are two main reasons why we fixed T=5:

1. First, we selected the value of T which ensured the least data points loss for the aggregated dataset of healthy and substance dependent individuals (data points removed: 1.9% for T=5, 2.9% for T=6, 2.5% for T=7).

2. The computational machinery of LMM needs longitudinal structures with a great number of observations within a specific longitudinal block (Bartolucci et al., 2012). The model with T=5 maximizes the number of data points within the longitudinal blocks, by ensuring a more reliable parameters estimates.

About the latent process characterization, we adopted a model selection criterion to choose the number of latent states S. Since our dependent variable is a categorical response variable with three levels, the possible choices reduce to a 2-state model and a 3-state model. In order to select the best model we relied on both BIC (Bayesian Information Criterion; (Schwarz et al., 1978)) and AIC (Akaike Information Criterion; (Akaike, 1974)) criteria. Note that, for both criteria smaller values indicate a better model performance. Both 2-state and 3-state models are preferable to a baseline 1-state model which does not account for latent process dynamics (Table 3.1).

TABLE 3.1: Latent States selection

Model	BIC	AIC
1-state	8792	8781
2-states	8561	8490
3-states	8608	8493

However, since results are very similar for the two candidate models, we adopted a further qualitative model selection criteria. In particular, we compared the estimates for the two models to determine which one provided the most useful and realistic substantive description of the data. We concluded that the 3-state model accounted for a more sensible and complete description of set-shifting performances (see (de Haan-Rietdijk et al., 2017) for a similar approach). The reader is referred to Appendix A for a comparison of models' estimates. Thus, we required the model to be based on three distinct latent components, which were expected to have a direct psychological interpretation (see the results section). Moreover, in order to account for group differences in the latent process we also used a binary time-fixed covariate X, codifying the membership of each participant to either Control group ($X = 0$) or Substance Dependent group ($X = 1$). In such a way, we could control for eventual differences between the two sub-populations. Therefore, eventual

differences in set-shifting performance trends between the two groups were completely captured by differences in the latent states dynamics.

In what follows, we describe the model parameters and the main probabilistic relations in the system:

(i) the *conditional response probabilities*

$$\phi_{y|s} = P(Y_t = y|S_t = s),$$

where $y \in \{\mathrm{C}, \mathrm{E}, \mathrm{PE}\}$ and $s = 1, 2, 3$. This parameters set characterizes the measurement model which concerns the conditional distribution of the possible responses given the latent process. It is assumed that the measurement model is conditionally independent of the covariate. Here we are not interested in explaining heterogeneity in the response model between the two groups, since in our view only dynamics in the latent process are responsible for differences in performance trend between groups;

(ii) the *initial probabilities*

$$\begin{aligned} \pi_{s|0} &= P(S_1 = s|X = 0), \\ \pi_{s|1} &= P(S_1 = s|X = 1), \end{aligned}$$

where $s = 1, 2, 3$. This parameter characterizes a distribution for the initial state across the (latent) states. In particular, $\pi_{s|0}$ and $\pi_{s|1}$ refer to the initial probabilities vectors of the states for the control group and for the substance dependent group, respectively;

(iii) the *transition probabilities*

$$\begin{aligned} \pi^{(0)}_{s_t|s_{t-1}} &= P(S_t = s_t|S_{t-1} = s_{t-1}, X = 0), \\ \pi^{(1)}_{s_t|s_{t-1}} &= P(S_t = s_t|S_{t-1} = s_{t-1}, X = 1), \end{aligned}$$

where $t = 2, \ldots, 5$ and $s_t, s_{t-1} = 1, 2, 3$. This parameter characterizes the conditional probabilities of transitions between latent states across the task phases. In particular, $\pi^{(0)}_{s_t|s_{t-1}}$ and $\pi^{(1)}_{s_t|s_{t-1}}$ refer to the transition probabilities for the control group and the substance dependent group, respectively. Here we assume that a specific covariate entails the characterization of a sub-population with its own initial and transition probabilities of the latent process. In this way, accounting for differences in performance trend relies on explaining heterogeneity in the latent states process between the two groups. In order to allow the covariate to condition the characterization of the latent process we adopt a logit parameterization as follows:

$$\log \frac{P(S_1 = s|X_t = x)}{P(S_1 = 1|X_t = x)} = \log \frac{\pi_{s|x}}{\pi_{1|x}} = \mathbf{x}^\top \boldsymbol{\Theta}$$

where $s = 2, \ldots, k$, for the initial probabilities, and

$$\log \frac{P(S_t = s|S_{t-1} = s^*, X_t = x)}{P(S_t = s^*|S_{t-1} = s^*, X_t = x)} = \log \frac{\pi_{s|s^*,x}}{\pi_{s^*|s^*,x}} = \mathbf{x}^\top \boldsymbol{\Gamma}$$

where $t = 2, \ldots, T = 5$, $s, s^* = 1, 2, 3$ and $s \neq s^*$, for the transition probabilities. In both parameterization, $\mathbf{x}^\top$ is a proper design matrix, whilst $\boldsymbol{\Theta}$ and $\boldsymbol{\Gamma}$ are regression coefficients vectors.

In this way the covariate entails the characterization of a sub-population with its own initial and transition probabilities of the latent process, whereas the conditional distribution of the response variable given the latent process does not depend on the specific sub-population.

According to this framework, the identification of the probabilistic relationships between latent states and observed responses, as well as those between latent states themselves, conveys all the information needed to characterize the observed response patterns dynamics.

3.3 Results

The proposed model was fitted using the *LMest* package (Bartolucci et al., 2012) developed within the *R* framework (Team et al., 2013). *LMest* relies on an efficient log-likelihood maximization procedure (e.g., Expectation-Maximization Algorithm) for parameters estimation. Moreover, a model selection criterion was used to evaluate if the model with the group covariate X was preferable to the simpler model without the grouping variable. In particular, we adopted both the BIC (Bayesian Information Criterion; (Schwarz et al., 1978)) and AIC (Akaike Information Criterion; (Akaike, 1974)) to measure the overall model performance. As expected, the model with the group covariate turned out to be the most appropriate model (see Table 3.2) thus confirming that the performance patterns were clearly different between the two groups.

TABLE 3.2: Model Selection criteria

Model	BIC	AIC
Basic	8858	8691
Covariate	**8608**	**8493**

3.3.1 Conditional Response Probabilities

The estimated conditional response probabilities $\hat{\phi}_{y|s}$ are presented in Table 3.3. These probabilities allowed us to characterize the latent states. The first state ($s = 1$) showed the highest probability to respond correctly, indicating that

participants minimized errors within a task phase. By contrast, the second state ($s = 2$) showed an increased probability of the error component, in particular the probabilities that a non-perseverative error or a perseverative error occur were approximately the same. This indicated the adoption of a non-efficient strategy, although the probability to respond correctly was still relatively high. Finally, the third state ($s = 3$) showed a different pattern in which the probability to produce a correct response resulted lower than the probability to produce an error. The errors pattern also entailed a higher perseverative component compared to the second state.

TABLE 3.3: Estimated conditional probabilities of responses given the latent state.

	$\hat{\phi}_{y\|s}$		
y	$s = 1$	$s = 2$	$s = 3$
C	0.93	0.80	0.44
E	0.02	0.10	0.38
PE	0.05	0.10	0.18

These probability distributions represent cognitive response strategies as a function of the latent component or state. In particular, in our context, State 1 may be easily understood as an *Optimal Strategy* whereas State 2 seems to characterize a type of *Sub-Optimal Strategy*. Finally, State 3 indicates a *Perseverative Non-Optimal Strategy*. Therefore, this latent states characterization may be adopted to describe the average ability to operate shifting cognitive strategies.

3.3.2 Initial Probabilities

Table 3.4 reports the model initial probability configurations. These initial probabilities indicated that the two groups performed the early phase of the test very differently. In particular, the control group showed a higher overall probability of starting the initial test phase in State 1. By contrast, the SDI group showed a higher probability to adopt a strategy admitting an error component at the initial phase of the task. This interesting result could reflect the finding that substance dependent individuals usually show an inefficient initial conceptualization of the task (Heaton, 1981).

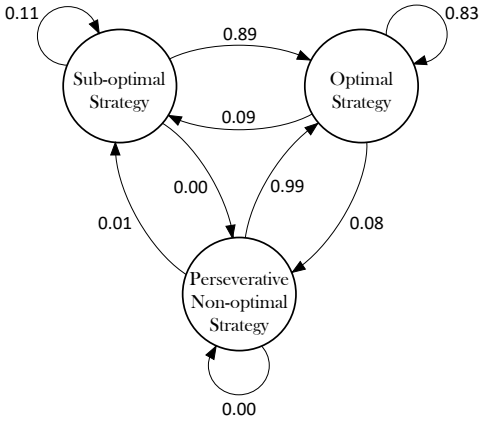
TABLE 3.4: Estimated initial probabilities for each group

	$s=1$	$s=2$	$s=3$
$\hat{\pi}_{s\|0}$	0.57	0.14	0.29
$\hat{\pi}_{s\|1}$	0.38	0.21	0.41

3.3.3 Transitions Probabilities

All the available information on the dynamics of the latent process can be conveyed by the transition probabilities matrices (see Figure 3.3). These matrices represent, at a given task phase t, the probability to transit from a current state s to a different state s^* or to remain in the same state s.

	$\hat{\pi}^{(0)}_{s\|s^*}$		
s^*	$s=1$	$s=2$	$s=3$
1	0.83	0.09	0.08
2	0.89	0.11	0.00
3	0.99	0.01	0.00



	$\hat{\pi}^{(1)}_{s\|s^*}$		
s^*	$s=1$	$s=2$	$s=3$
1	0.01	0.99	0.00
2	0.01	0.99	0.00
3	0.02	0.33	0.65

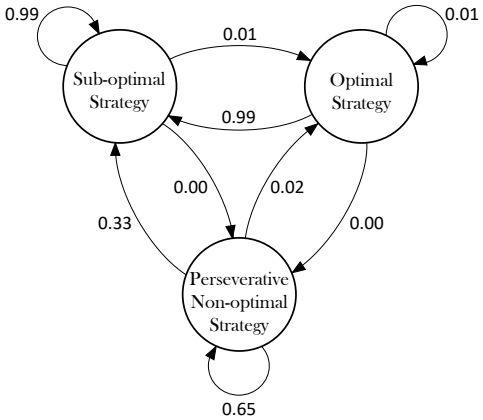


FIGURE 3.3: Transition probability matrices (left) and relative graphical model representations (right) for the control group (top) and the SDI group (bottom).

The transition matrix for the control group showed a clear pattern. First, the diagonal values revealed that the probability to reiterate a certain strategy

decreased as it became less optimal, up to a zero probability to reiterate transitions to State 3, which clearly represents the non-optimal response strategy. Moreover, the very low probability values in the second and third columns indicate that it was nearly impossible to adopt a response strategy affected by the error component, since the overall probabilities to transit to State 2 or State 3 approached zero. It is also worth noting that, the probabilities approaching 1 in the first column indicate a general tendency of the system to transit to State 1, suggesting that these participants tended to switch to the optimal strategy in case they were not in that status, and to maintain that strategy for the rest of the task. Clearly, this pattern reflected the tendency to quickly minimize both the perseverative and the non-perseverative components of the error, as the task unfolded across time.

The transition matrix for the SDI group showed rather different dynamics. Importantly, the system exhibited a general tendency to transit to State 2, the sub-optimal strategy. In particular, it is worth emphasizing that a probability approaching 1 on the main diagonal could be understood as the presence of an *absorbing state*. This means that once in State 2, the system tended to reiterate the same latent state and that SDI participants systematically reiterated the sub-optimal strategy and never transited to the optimal strategy during the task. Further, once in State 3, there was a relatively high probability that an individual remained stacked in that state, indicating the tendency to reiterate the non-optimal strategy and to show a perseverative component of the error. On one hand, this pattern could also reflect the presence of mental rigidity as for substance dependent individuals was nearly impossible to switch to the optimal strategy. On the other hand, the tendency to reiterate a sub-optimal strategy by keeping fixed the error component across the task could also be seen as a probabilistic account for the failure to maintain set phenomenon (Figueroa and Youmans, 2013). This is in accordance with some findings reporting this peculiar behaviour in substance dependent individuals (Tarter, 1973).

3.3.4 Marginal latent states distributions

In order to better understand our model results, we analysed the marginal distribution of the latent states. For each task phase, we derived a probability distribution over the three states for each group. To do so, we relied on basic rules for markov chains. Let π_t be the distribution of the latent states at a certain time step t, or task phase, and let the transition matrix $\pi_{s_t|s_{t-1}}$ be codified as P, for notational convenience. For each time step $t+1, t+2, \ldots, T=5$ we want to compute the quantities $\pi_{t+1}, \pi_{t+2}, \ldots, \pi_T$. The purpose is to move the distribution π_t forward one unit of time, by starting from π_1, which is the initial probability vector. It can be shown that $\pi_{t+n} = \pi_{t+(n-1)}P$ (Dobrow, 2016). Regarding the control group, Figure 3.4 shows that the optimal strategy was maintained for the entire duration of the test, and the probability to adopt a strategy admitting errors component decreased quickly. This indicates that participants in the control group tended to learn immediately how

to minimize the error component. Figure 3.5 shows the marginal distributions of the states for the SDI group. The plot shows that the probability to adopt an optimal strategy decreased faster than the probability to adopt a non-optimal perseverative strategy. The sub-optimal strategy with both error components showed the higher probability to be maintained for the rest of the task, suggesting that substance dependent individuals never minimized the error component during the test.

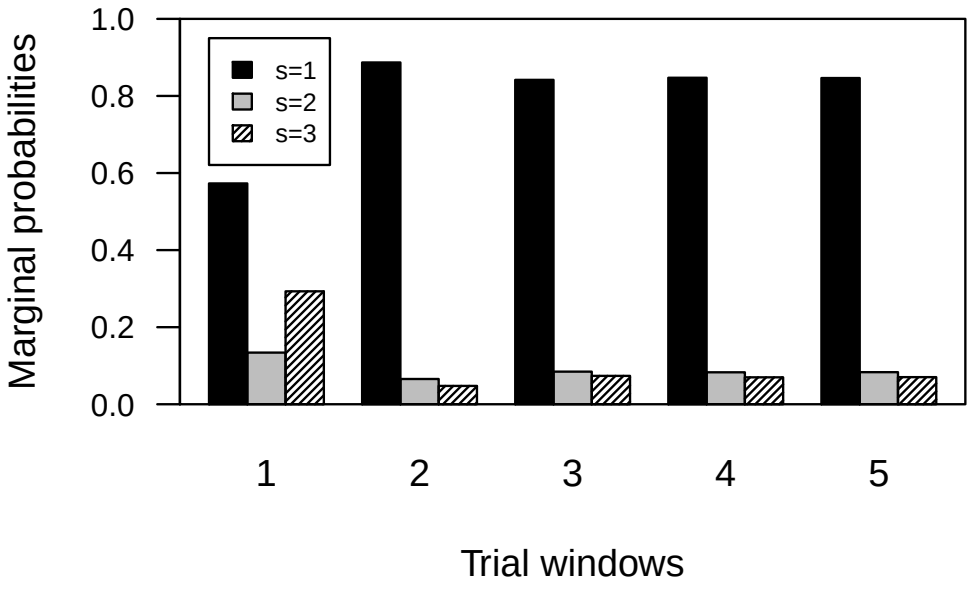


FIGURE 3.4: Marginal distribution of the latent states for each task phase, for the Control group.

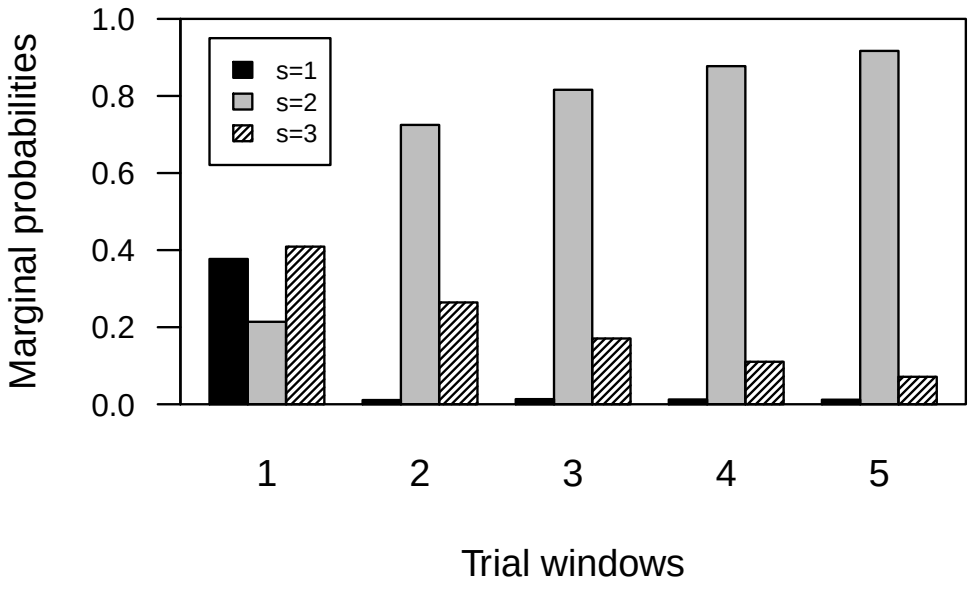


FIGURE 3.5: Marginal distribution of the latent states for each task phase, for the SDI group.

3.4 Discussion of results

Results clearly show that our model is able to capture differences in performance trend between Control and SDI groups in terms of differences in their latent process transition dynamics. The characterization of the conditional response probabilities allows to rephrase the latent states as cognitive strategies adopted in a given phase of the task. Results also show that a 3-state model is a reasonable choice if we want to differentiate dynamics in strategy shifting between groups. In fact, it is unrealistic to think that individuals can rely only on two (latent) cognitive strategies to accomplish the task, as it would be in case of a 2-state model. The three states could be clearly interpreted as error-related strategies with gradually increasing error components. However, one might argue that our state process characterization does not account for three distinct latent components, due to similarities in probability patterns of responses for some of the states (such as State 1 and 2). According to our view, an inspection of the marginal distributions of the latent states in Figure 3.4 and 3.5 can clarify that our model actually accounts for three non-overlapping latent components, as reflected by the differential marginal states probabilities pathways for the two groups.

It is also worth emphasizing that these results can increase the amount of information a researcher can obtain from the assessment of set-shifting performances. Generally, the analysis of data from the WCST reduces to

the computation of summary statistics of the scoring measures, which in turn may provide the input for standard statistical analysis, as well as for classification procedures based on cut-off thresholds (Demakis, 2003). Mean scoring measures across individuals provide a simple way to account for group-level differences in performances (Table 3.5).

TABLE 3.5: Mean scoring measures (*SE* in parenthesis)

	C	*E*	*PE*
Control	66.34 (0.64)	5.25 (0.29)	4.65 (0.42)
SDI	72.34 (2.17)	10.57 (0.96)	14.28 (1.52)

In our case study the groups differ in the number of perseverative ($t(80) = 5.62$, $p < 0.001$) and non-perseverative ($t(80) = 6.48$, $p < 0.001$) errors. However, mean differences cannot account for hypothesis about the underlying causes. From our modelling perspective, differences in mean scoring measures can be explained by the heterogeneity in the latent process affecting the way in which individuals within each group respond at a given phase of the task. Thus, a fundamental additional information provided by our model consists in the data generating process.

3.5 General discussion

The modelling approach proposed in this work was able to map the evolution of response patterns in a set-shifting task with the evolution of a latent states process underlying the observed behaviour. The model provided a parsimonious description of the dynamic processes underlying the data, since we were able to represent the performance trend by using a latent variable with just three categories, representing different cognitive strategies evolving in time. Moreover, the estimated parameters capturing these dynamic aspects could be readily put in relation with some psychological constructs of potential clinical relevance.

However, a crucial issue is that related to the interpretation of these parameters. Although accounting for a data generating process could convey interesting and additional information for the analysis of behaviour outcomes, parameters interpretation is not trivial. Marginal latent states distributions offered a straightforward way to examine dynamic aspects of error-related behaviour. For instance, marginal distributions showed that control group (resp. SDI group) settles to State 1 (resp. State 2) across task phases, which is approximately the same information conveyed by the summary measures of the number of errors for each group. From this perspective, marginal distributions provided no additional information for the analysis of participants' performances. Conversely, transition probabilities matrices provided a more exhaustive source of information at the cost of an increasing difficulty in

results interpretation (e.g. differences in performance trends between groups must rely on row-wise, column-wise, main diagonal values comparison). Therefore, transition probabilities offer the advantage to rely on parameters estimates for simulation and forecasting purposes.

In particular, the transition matrices can be seen as cognitive system profiles and one might be interested in generating data in order to test sensible hypothesis. For example, given the two estimated profiles, one for each group, a sensible question could be: Which system does reach the optimal strategy first, on average, given the assumption that both systems start the task at State 3, the perseverative non-optimal strategy? This kind of investigation could be hard, or even impossible, for standard analytical frameworks based on simple summary statistics of the scoring measures.

In conclusion, our LMM model provides, at least in this first preliminary work, an interesting tool to analyse data presenting a dynamic component. It also illustrates an efficient way to manage differences between groups by accounting for the heterogeneity in the latent process characteristics between them. However, further works are needed in order to solidly establish connections between parameters estimates and more subtle cognitive constructs.

Chapter 4

A Dynamic Bayesian Network account of complex behavioural patterns in set-shifting tasks

4.1 Introduction

Executive functions (EFs) refer to a particular class of higher-order cognitive processes which enable humans to attain a certain goal (Dempster, 1992) by executing appropriate actions, while inhibiting inappropriate ones. They are thought to play an important role in everyday life achievements by affecting not only cognitive, but also social, emotional, and organizational aspects of human behaviour. For example, cognitive measures of executive functions can predict attributes such as occupational status, communicative behaviour, moral behaviour and social cognition across normal and clinical populations (Kibby et al., 1998; Moriguchi et al., 2008; Carlson and Moses, 2001). Several psychological models have been proposed to provide a clear definition of EFs, at both structural and functional level (Miyake et al., 2000; Miller and Cohen, 2001). In general there is a certain agreement in considering these functions as a set of higher-order skills, which serve as a way to monitor, control, and organize complex thoughts and behaviour (Anderson, 2008). More specifically, several evidences suggest that they can be factorized into hierarchically organized separable sub-components (Alvarez and Emory, 2006; Moriguchi et al., 2008). From a functional and neurobiological perspective, experimental cognitive settings requiring the involvement of EFs are known to elicit the activation of a network of cortical and subcortical brain structures which could be conceived as one aspect of a diffuse executive system (Duffy and Campbell III, 2001). This kind of distributed network of neural circuits is activated when task demands involve integrated functioning (Alvarez and Emory, 2006), and are known to be particularly relevant when participants must accomplish cognitive shifting, inhibitory control, working memory and self-regulation tasks (Miyake et al., 2000; Malloy and Richardson, 1994; Stuss et al., 2000), to name a few.

In this work we focus on set-shifting, a fundamental executive function which consists in the ability to alter a behavioural response mode in the face of changing contingencies (Berg, 1948). The choice of relying on set-shifting as a way to formally account for executive functioning is particularly suitable

for our modelling purpose since its psychological assessment involves well-established experimental settings which can be easily taken into account in a computational fashion. The Wisconsin Card Sorting Test (WCST; (Heaton, 1981; Berg, 1948)) is perhaps considered the most renowned neuropsychological setting to measure set-shifting as well as executive functions (Alvarez and Emory, 2006), and we retain it as an ideal framework to develop our cognitive model. In a WCST, participants learn to pay attention and respond to relevant stimuli features, while ignoring irrelevant ones, as a function of experimental feedback. Individuals are asked to sort a target card with one of four stimulus cards, consisting in geometric figures that vary in terms of several features, namely, color (red, green, blue, yellow), shape (triangle, star, cross, circle) and number of objects (1, 2, 3 and 4), according to the proper sorting rule on any given trial (Figure 4.1).

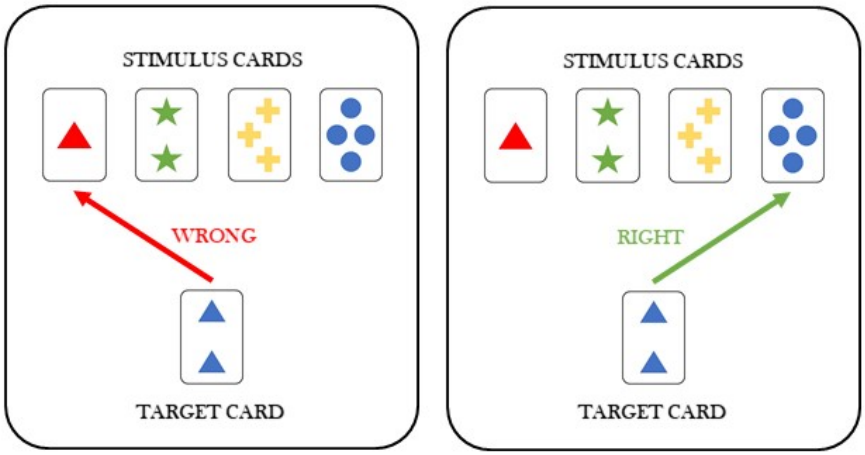


FIGURE 4.1: Suppose that in a certain trial, the target card consists in two blue triangles, whilst the correct sorting rule is the feature color. A correct response requires that the target card is sorted with a stimulus card presenting the color blue, regardless of the other features the stimulus card presents. Sorting by shape results in a wrong response (left). Sorting by color results in a right response (right).

Each response is followed by a feedback telling the individual if the sorting is right or wrong. After a fixed number of consecutive correct responses, the sorting rule is changed by the experimenter without warning, and participants are required to infer the new sorting rule. Sometimes individuals still persist with the old rule and may produce what is called a perseverative response. It is worthwhile noticing that in such cognitively highly demanding task several aspects of executive functions are involved, such as: inhibition and switching (Burgess et al., 1998), working memory (Barceló and Rubia, 1998; Zelazo et al., 1997), attentional processes (Barceló, 2001) and error detection (Lie et al., 2006), among others.

From a neurocognitive perspective, the matching of a card in a given trial allows individuals to receive a feedback which is supposed to directly condition future choices. Here, the hypothesis is that the received feedback is

integrated with information in the working memory about earlier task phases in order to decide whether to maintain or shift the current attentional set (Monchi et al., 2001). Thus, the ability to process feedback related information seems to be crucial in order to correctly perform the task. The relation between feedback and behaviour updating is mediated by lateralized prefrontal brain activities (Lie et al., 2006). Interestingly, there seems to be a functional dissociation in processing negative and positive feedback (Jimura et al., 2004). First of all, when a negative feedback occurs, increasing activity is observed in the anterior cingulate cortex, which is part of an error detection network which in turn alerts the attentional capacities to prevent further errors (Lie et al., 2006). Therefore, when receiving negative feedback, individuals tend to show an increased neural activity in dorsolateral and ventrolateral prefrontal areas, bilaterally, as well as in other subcortical structures such as caudate nucleus and dorsal thalamus. Differently, when a positive feedback occurs, increasing activities are observed in right dorsolateral and posterior prefrontal regions (Monchi et al., 2001). Further dissociations can be found even when individuals have to modulate behaviour after receiving a given experimental feedback. Thus, sorting a target card after receiving a negative or a positive feedback elicit posterior prefrontal and premotor cortex neural activities, respectively, and a shared parietal activation (*ibidem*). Lateralized prefrontal brain regions have also been related to the process of updating behaviour. It was suggested that this activation may be related to updating temporarily maintained internal states, such as cognitive sets (Lie et al., 2006). Importantly, a functional and anatomical dissociation in prefrontal regions in accounting for set-shifting and set-maintenance has been observed (Monchi et al., 2001).

These evidences seem to endorse the speculation that component processes involved in our cognitive setting are functionally separable. In addition, they allow us to outline some functional principles characterizing the interaction between a cognitive agent's system and the set-shifting experimental framework. In general, the sub-processes involved in the task appear to rely on accumulating information over time and operating on them in an efficient way. Both differential feedback information processing and internal attentional processes take place in order to allow the system to update behaviour during the unfolding of the task. A computational model that aims to reproduce a suitable agent behaviour should allow these functional principles to be embedded in a proper cognitive representation in which different executive system's component processes are integrated to produce behaviour. This work introduces a cognitive model based on a probabilistic framework which is particularly suitable to formally account for such sub-processes functional integration, consistently with neuroscientific evidences. It is shown how this computational model simulates behaviour of a cognitive agent facing a WCST and reproduces human performances as measured by clinical neuropsychological assessment criteria. The rest of the chapter is organized as follows. First, a brief description of the mathematical framework adopted is provided. Then, our computational model is shown and motivated. Two simulation experiments are proposed in order to assess the ability of the model to reproduce complex behavioural patterns in different scenarios. Finally, further

remarks on the usefulness of the model for clinical and psychological sciences are discussed.

4.2 The Cognitive Agent Model

In this section, the cognitive model is described in detail. We outline the main features of the model by first providing an overview of the general probabilistic structure. The different components of the model are then described separately.

4.2.1 Probabilistic Structure of the Executive System

In this work, we propose a DBN architecture to account for the relationships between component processes in the cognitive system. The conditional (in)dependence structure characterizing the network reflects our assumptions about how component processes phenomenologically influence each other, based on anatomical and functional evidences. Each node can be seen as a component of the executive system which allows a cognitive agent to evolve and adapt its behaviour based on internal or external information. The DBN is designed to reflect a hierarchical organization of the cognitive system in which the environmental information is processed by high-level functional nodes which coordinate, in a top-down manner, the activities of low-level nodes. The main advantage of our network approach to model cognition is that all the components contribute to achieve the task in an integrative manner. If one of them exhibits a non-efficient behaviour, this may lead to specific deficits in executive functioning, and thus to a particular deficient behavioural pattern. To better reflect this idea, the design of the model is specified by means of specific parameters settings. A given parameters setting could reflect functional deficiencies at different levels of the cognitive hierarchy.

In general, the computational model is thought to integrate components such as: feedback processing, feedback information integration to update behaviour, set-maintaining and set-shifting, stimuli information processing, and internal representation of the response. To clarify these concepts, we introduce the stochastic and deterministic nodes involved in the model. Consider a WCST setting with T trials. An agent delivers a response $Y_t \in \{1, 2, 3, 4\}$ at a given trial t of the task, where $t = 1, 2, \ldots, T$. Y_t indicates the stimulus card which is thought to correctly match the current target card. On each trial, the agent receives an experimental feedback $X_t \in \{0, 1\}$, codifying the response as right ($X_t = 1$) or wrong ($X_t = 0$). Feedback affect the behaviour of the agent by allowing the system to land to a given cognitive state $S_t \in \{1, 2\}$ (1: set-maintenance, 2: set-shifting). However, before transitioning to a certain state, the cognitive system processes the experimental feedback by means of two parallel mechanisms, ω_t and δ_t, which compute an internal value of the positive and negative feedback, respectively. Clearly, the cognitive state plays a relevant role in the internal representation of the response. In particular, we assume that the system internally operates upon a finite set of abstract rules

$\Omega \in \{1, 2, 3\}$ (1: color, 2: shape, 3: number). The card sorting procedure relies on sampling a rule, $R_t \in \Omega$, which allows the system to select the sorting principle to adopt for each trial t. Moreover, the cognitive state affects the way the system samples the rules. Conditioned on the rule the system adopted, a response is delivered. At this point, a new feedback is received and the process repeats. The structure of the probabilistic relationships between the variables is represented in the graphical model in Figure 4.2.

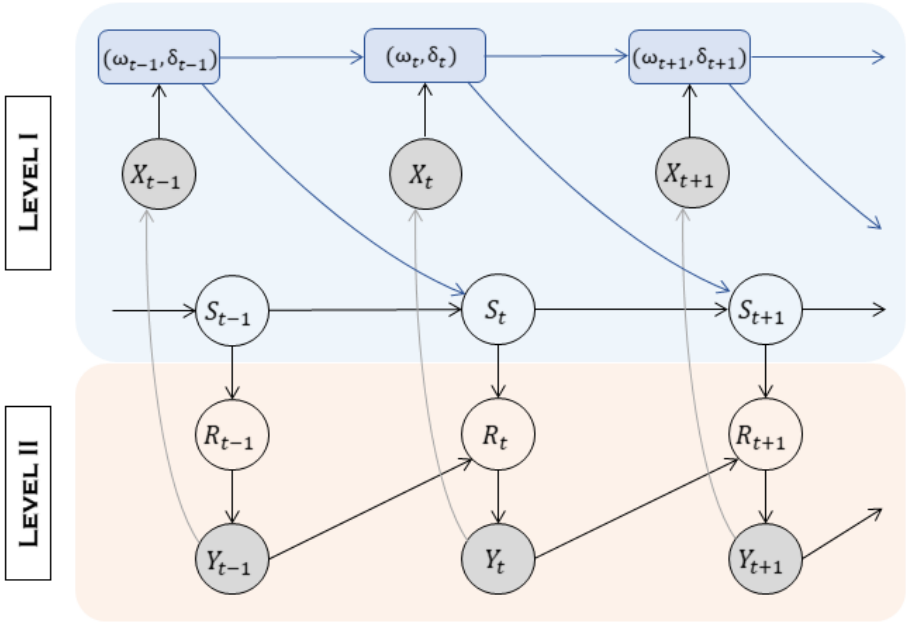


FIGURE 4.2: Graphical model representation.

Here, Level I and Level II codify the hierarchical organization of the computational cognitive system. One might think of Level I as the level in which the most general purpose-oriented processes take place. This means that such level structure holds regardless of the environmental (or experimental) setting. Differently, Level II can be seen as the task-oriented level in which we model the abstract and the experimental factors upon which the system operates (sorting rules and cards to sort within the WCST setting). More specifically, our assumptions about the conditional (in)dependencies between variables result in the following factorization of the join probability distribution:

$$\begin{aligned} P(S_{1:T}, R_{1:T}, Y_{1:T}, X_{1:T}|\Theta) &= P(S_1|\omega_0, \delta_0)P(R_1|S_1)P(Y_1|R_1)P(X_1|Y_1) \\ &\quad \prod_{t=2}^{T} P(S_t|S_{t-1}, \omega_{t-1}, \delta_{t-1})P(R_t|S_t, Y_{t-1})P(Y_t|R_t)P(X_t|Y_t) \end{aligned}$$

where Θ is the set of parameters characterizing the computational formalisation for the nodes. It is worthwhile noticing that our modelling approach does not take into account anatomical and neurobiological elements in representing causal relations between executive components. Here, we rely on a more abstract representation of cognitive, or mental, states. In this way,

neurological knowledge underlying cognitive functioning serves as a basis for the mental (cognitive) modelling level.

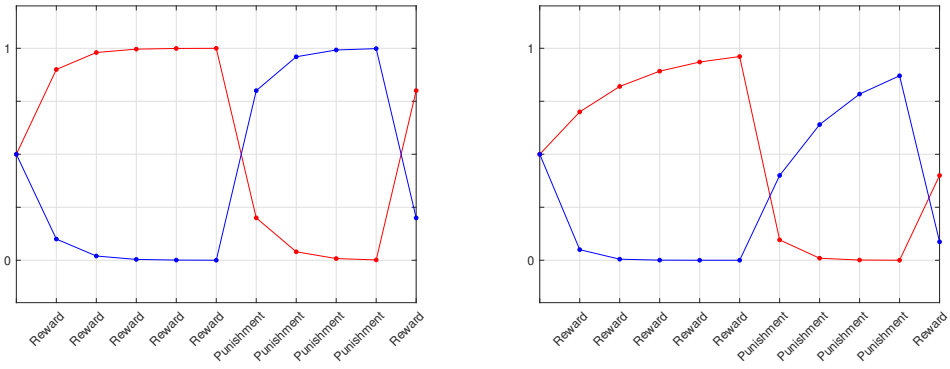
4.2.2 Attention to feedback

As outlined earlier, positive (reward) and negative (punishment) feedback seem to be processed separately at a neural and functional level. Therefore, the characterization of two distinct mechanisms which process feedback may nicely reflect this assumption. In particular, we assume the existence of two parallel processes which, given a certain feedback as input, determine an internal value of that feedback. Let us introduce the following equations, called feedback equations:

$$\omega_t = \lambda X_t + [X_t(1-\lambda) + (1-X_t)(1-k)]\omega_{t-1} \quad (4.1)$$

$$\delta_t = \xi(1-X_t) + [(1-X_t)(1-\xi) + X_t(1-k)]\delta_{t-1} \quad (4.2)$$

These equations represent the ability of the system to integrate feedback related information by allocating an internal attentional value to it. Eq. (4.1) modulates the attention to reward (ω_t) at trial t. Parameter λ codifies the reward updating weight, and as it approaches 1, ω_t increases faster. Eq. (4.2) modulates the attention to punishment (δ_t) at trial t. Parameter ξ codifies the punishment updating weight, and as it approaches 1, δ_t increases faster. Parameter k modulates the promptness of the system to disengage attention from the current attended feedback when environmental contingencies suggest it. We call k flexibility. In this case, the flexibility is considered as a general characteristic of the system, and modulates both attentional mechanisms. Several combinations of parameters lead to particular functional patterns in the processing of the feedback related information (Figure 4.3).



(A) $\lambda = 0.8, \xi = 0.8, k = 0.8$ (B) $\lambda = 0.4, \xi = 0.4, k = 0.9$

FIGURE 4.3: Behaviour of the attention to reward (red) and the attention to punishment (blue) functions under different parameterizations. On the left, the attention to reward (resp. punishment) rapidly increases as a positive (resp. negative) feedback occurs. Similarly, the attention to reward (resp. punishment) rapidly drops down when a negative (resp. positive) feedback is received. On the right figure, the functional dynamics show a different pattern. The attention to reward (resp. punishment) slowly increases when a positive (resp. negative) feedback occurs.

The aforementioned feedback equations are extensions of the basic exponential smoothing function (Zucchini et al., 2008; Hyndman et al., 2002), and show some interesting properties. For a given feedback processing mechanism, we have two sub-components characterizing the behaviour of the functions. To clarify this concept, we focus on the analysis of eq. (4.1), codifying the attention to reward ω_t. However, the same applies to the attention to punishment δ_t. Based on the feedback the system receives, the behaviour of the function can be decompose as follows:

$$\omega_t = \begin{cases} \lambda + (1-\lambda)\omega_{t-1}, & \text{for} X_t = 1 \\ (1-k)\omega_{t-1}, & \text{for} X_t = 0 \end{cases}$$

The increasing step of the value of reward attention depends on parameter λ, which is assumed to be bounded in $[0, 1]$. Differently, the decreasing step of the value of reward attention is only affects by parameter k, which is also bounded in $[0, 1]$. This means that we are assuming a cognitive system which shows a dissociation in sub-components of a given feedback processing. Some limiting behaviour can be observed. Suppose the case in which $k > 0$, then:

$$\lim_{\lambda \to 0} \omega_t = \begin{cases} \omega_{t-1}, & \text{for} X_t = 1 \\ (1-k)\omega_{t-1}, & \text{for} X_t = 0 \end{cases}$$

and

$$\lim_{\lambda \to 1} \omega_t = \begin{cases} 1, & \text{for} X_t = 1 \\ (1-k)\omega_{t-1}, & \text{for} X_t = 0 \end{cases}$$

The scenario where $\lambda = 0$ reflects the case in which our system is totally unable to increase the internal attentional value of the reward when receiving a positive feedback. The main limitation of such a deficient cognitive system is that it never re-allocates attention to reward once a negative feedback occurs. On the other hand, when $\lambda = 1$, we have a perfect system which maximizes the attention to reward immediately as a positive feedback occurs.

Consider now the case in which $\lambda > 0$, then:

$$\lim_{k \to 0} \omega_t = \begin{cases} \lambda + (1-\lambda)\omega_{t-1}, & \text{for} X_t = 1 \\ \omega_{t-1}, & \text{for} X_t = 0 \end{cases}$$

and

$$\lim_{k \to 1} \omega_t = \begin{cases} \lambda + (1-\lambda)\omega_{t-1}, & \text{for} X_t = 1 \\ 0, & \text{for} X_t = 0 \end{cases}$$

In case flexibility k equals zero, the system is unable to decrease the internal attentional value of the reward when receiving a negative feedback. This means that once the system allocates a certain attentional value to reward, it never disengages attention from the positive feedback when receiving a negative one. A direct consequence of this deficient functioning is that the system continues to maximize the probability to reiterate the set-maintenance cognitive state (see section below), even when experimental cues suggest that it is no longer advantageous. On the other hand, when $k = 1$ we have a perfect system which minimizes the attention to reward immediately as a negative feedback occurs.

4.2.3 Shifting between Cognitive States

The internal value of the feedback is further processed in order to characterize the transition dynamics between the two cognitive states responsible for shifting between cognitive sets. More precisely, it is assumed that the system can switch between two states, namely, the set-maintenance state ($S_t = 1$), and the set-shifting state ($S_t = 2$), at each trial. In particular, processing the feedback received at trial t affects the transition to a given cognitive state at trial $t + 1$. The following equations, called transition equations, compute the transition probabilities to move from a state to another. Eq. (4.3) and (4.4) show the probabilities to reiterate a transition to the same state, that is, to settle on a given cognitive state.

$$P(S_{t+1}=1|S_t=1) = \gamma_{11}(t) = (1+\exp(-\alpha_0-\alpha_1\omega_t))^{-1} \quad (4.3)$$
$$P(S_{t+1}=2|S_t=1) = \gamma_{12}(t) = 1-\gamma_{11}(t)$$
$$P(S_{t+1}=2|S_t=2) = \gamma_{22}(t) = (1+\exp(-\beta_0-\beta_1\delta_t))^{-1} \quad (4.4)$$
$$P(S_{t+1}=1|S_t=2) = \gamma_{21}(t) = 1-\gamma_{22}(t)$$

These equations characterize the bias of the system to rely on the internal attentional feedback value in order to update behaviour. As can be noticed, the attention to reward only affects the probability to reiterate a set-maintenance cognitive state across trials. Differently, the attention to punishment only affects the probability to reiterate a set-shifting cognitive state. Parameters α_0 and β_0 are measures of the conservativeness of the system. Lower values of α_0 (resp. β_0) mean that, in order to increase the probability to maintain (resp. shift) the set, the system needs to allocate more attentional resources to reward (resp. punishment). Parameters α_1 and β_1 represent the slopes of the logistic functions. The system can be defined unbiased, or balanced, in case $\alpha_0/\alpha_1 = -1/2$ and $\beta_0/\beta_1 = -1/2$. This condition ensures that the system has the same probability to maintain or shift the set, when the attentional value to feedback is totally uninformative ($\omega_t = 0.5$, or $\delta_t = 0.5$). Differently, the bias of the system is expressed by the extent the intercept-to-slope ratio moves away from the balanced condition. For example, when $\alpha_0/\alpha_1 < -1/2$, the system shows a conservative bias in processing the positive feedback.

4.2.4 Stimuli Information Processing

Once the general purpose-oriented cognitive processes have taken place, the system needs to internally represent the external stimuli in order to operate on them. In our context, these stimuli consist in the target card and the four stimulus card presented in each given trial. The computational model internally represents the target card as a features matching vector $\mathbf{m}_t = (m_1, m_2, m_3)$, where $m_r \in \{1,2,3,4\}$ and the subscript r codifies a given feature (1: color, 2: shape, 3: number). By relying on the vector $\mathbf{m}_t$, the system acquires all the information available in the stimuli. In particular, m_r indicates which stimulus card matches with the current target card for the feature r. For example, in a given trial t the system could codify the target card as $\mathbf{m}_t = (2,1,1)$, meaning that the card matches with the stimulus card $Y_t = 2$, for the feature 1 (color), and with the stimulus card $Y_t = 1$, for both features 2 (shape) and 3 (number). In this way we are assuming that the system represents the relations between the cards in terms of what features they share.

4.2.5 Rule Sampling Process

In order to produce a response, that is, to sort the target card with a proper stimulus card, the system has to rely on the feature matching representation

of the stimuli, $\mathbf{m}_t$. The system has to select which feature (and thus, which sorting rule) is the most appropriate for the current trial. In particular, the model is thought to allow the cognitive system to change its belief about the correct rules based on the integration of information from both trial t and trial $t-1$. Conditioned on the previous response the system delivered and on the current cognitive state, a likelihood function over the rules (features) is computed according to:

$$P(R_t = r|S_t, Y_{t-1}) = \begin{cases} \frac{P(Y_{t-1}|R_t=r)}{\sum_j P(Y_{t-1}|R_t=j)}, & \text{if } S_t = 1 \\ \\ \frac{1-P(Y_{t-1}|R_t=r)}{\sum_j [1-P(Y_{t-1}|R_t=j)]}, & \text{if } S_t = 2 \end{cases} \tag{4.5}$$

Eq. (4.5) codifies an *exact likelihood*, defined as $f_R(t)$, computed on each trial t. It reflects the way the system internally represents responses by allocating a probability distribution over the features. When the feature r is sampled, then the element m_r in the current matching vector $\mathbf{m}_t$ carries information about which stimulus card has to be selected. However, it is assumed that the system can fail to achieve a correct representation of the exact likelihood, which can be degraded to some degree. We refer to this contaminated likelihood as *system's likelihood*, $h_R(t)$, which is computed according to the following equation:

$$P_h(R_t = r|S_t, Y_{t-1}) = \begin{cases} \frac{\pi}{|\Omega|} + (1-\pi)\frac{P(Y_{t-1}|R_t=r)}{\sum_j P(Y_{t-1}|R_t=j)}, & \text{if } S_t = 1 \\ \\ \frac{\pi}{|\Omega|} + (1-\pi)\frac{1-P(Y_{t-1}|R_t=r)}{\sum_j [1-P(Y_{t-1}|R_t=j)]}, & \text{if } S_t = 2 \end{cases} \tag{4.6}$$

where $|\Omega|$ is the total number of internally represented features, that is, the number of rules upon which the lower level computational mechanisms operate. The systems' likelihood reduces to a mixture distribution of the form:

$$h_R(t) = \pi g_R + (1-\pi) f_R(t)$$

where g_R is a discrete uniform distribution allocating equal probability mass to each feature (or rule). Actually, the system relies on this distribution to sample the feature to adopt as the sorting rule. As the mixture weight $\pi \in [0, 1]$ increases, the most unlikely rules, as computed by the exact likelihood, become more and more intrusive. In other words, the entropy of the system's likelihood increases as π increases thus degrading the amount of information available from the exact likelihood. Intuitively, parameter π codifies the level of distractibility of the agent. Figure 4.4 clarifies this concept.

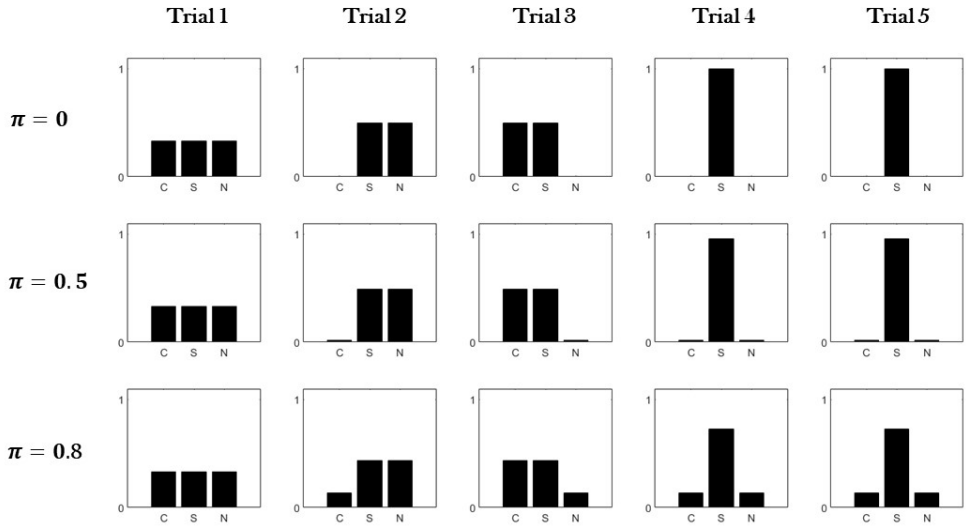


FIGURE 4.4: Probability distribution over the rules: color (C), shape (S), number (N).

Figure 4.4 represents an example scenario of probabilities the system assigns to each rule across five consecutive trials (columns). The first row represents the case in which $\pi = 0$, and the systems' likelihood equals the exact likelihood. The system reaches the solution after few trials by maximizing the probability to select the most appropriate feature for the current trial (for this example, the feature shape). The second row represents the case in which $\pi = 0.5$. The system shows a fair level of distractibility due to the fact that invalid rules have a certain probability to be sampled even after several trials. This immediately reflects an increased probability to wrongly sort the target card. The third row represents the case in which $\pi = 0.8$, and the probability to sample inappropriate features increases.

4.2.6 Generate Agent Behaviour

In our computational model, deterministic and probabilistic laws regulating the relation between the nodes characterize the behaviour of the agent on each trial of the cognitive task. The relational structure of the executive components is the same for the entire duration of the task, which in turn depends on the performance of the agent (as it is shown in the next section). In our context, we are interested in observing how the system evolves during the unfolding of the WCST based on the feedback received, by considering the fact that feedback directly depend on the response the system delivers on each trial. Rather than focusing on raw responses Y_t, indicating the position of the chosen stimulus card, we are more interested in observing the pattern of correct and error responses across the entire trials span. The cognitive agent responses are generated at each time step (trial), by letting the system to adapt automatically based on the experimental feedback, and the parameters setting describing its cognitive characteristics. The generative algorithm is shown in Table 4.1.

TABLE 4.1: Generative Algorithm

Algorithm	Computational Agent Model	
	$\Theta = \{\lambda, \xi, k, \alpha_0, \alpha_1, \beta_0, \beta_1, \pi\}$	SET PARAMETERS
$\mathbf{t = 0}$:	Set ω_0 and δ_0	INITIALIZATION
$\mathbf{t = 1}$:	Set default S_1 and Y_1	FIRST TRIAL
	Generate feedback X_1	
	Compute ω_1 and δ_1 (eq. (1) and (2))	
$\mathbf{t \geq 2}$:	Compute transition probabilities (eq. (3) and (4))	UPDATING
	Transit to a cognitive state S_t	
	Compute the exact likelihood (eq. (5))	
	Compute the system's likelihood (eq. (6))	
	Sample a rule (feature) $R_t = r$	
	Produce a response Y_t	
	Generate feedback X_t	
	Compute ω_t and δ_t (eq. (1) and (2))	
	Loop	

4.3 Simulation Study

In this section we test our computational model in order to assess its ability to reproduce the expected behaviour as a function of parameters modulation. In particular, we are interested in selectively damaging the system in order to produce a given cognitive functioning impairment. To do so, we study how the generative model behaves when facing the WCST in two distinct experiments. In a first experiment we impair two main executive components, namely, flexibility and distractibility, and see how the model is able to reproduce response patterns which are consistent with findings in the clinical literature. In the second experiment we investigate the expressiveness of the computational model in reproducing behaviour when several combinations of parameters are taken into account. Simulations were conducted within the Matlab programming environment.

4.3.1 Computational Model Assessment

For both experiments, the computational agent performances are analysed by relying on two sources of information: (1) the observed sequence of feedback,

$\mathbf{x}_{1:T}$, indicating the configuration of correct and incorrect responses across the trials span; (2) the observed sequence of the stimulus cards selected for the sorting procedure, $\mathbf{y}_{1:T}$. At this point, the array $\mathbf{w}_{1:T} = \big((x_1, y_1), \ldots, (x_T, y_T)\big)$ provides all the information we need to cognitively assess the agent behaviour.

Experimental Setting In both the experiments, the Heaton version of the WCST (Heaton, 1981) is administered to the cognitive agent. In this particular version, the sorting principle changes after a fixed number of consecutive correct responses. In particular, when the system correctly sorts the target card for a series of 10 consecutive trials, the sorting rule is automatically changed by the simulative apparatus. The duration of the task depends on the performance of the agent. When the agent completes six stages (categories) of 10 consecutive responses, the task ends. In case this condition is not met, the task ends after completing a maximum of 128 trials.

Scoring Measures Here, we want to make computational agent and humans performances comparable, at least qualitatively. To this aim we adopt a metric which is usually employed for the clinical assessment of test outcomes in neurological and psychiatric patients (Bechara and Damasio, 2002; Braff et al., 1991; Zakzanis, 1998; Tarter, 1973; Landry and Al-Taie, 2016). Thus, the agent performances as represented by $\mathbf{w}_{1:T}$ are codified according to a neuropsychological criterion (Heaton, 1981; Flashman et al., 1991) which allows to classify responses into several components. These components provide the scoring measures for the test. In particular, we are interested in: *total number of trials* achieved (NT), *number of completed categories* (NCC), *perseverative errors* (PE), *non-perseverative errors* (NPE), *correct responses* (C), *trials to complete the first category* (TFC), and *Failures to Maintain Set* (FMS). The scoring measure NT tells us how many trials the agent needs in order to complete the task. Perseverative errors (PE) occur when the agent applies a sorting rule which was valid for the previous category. Usually, detecting a perseveration is far from trivial, since several responses configurations could be observed when individuals are required to shift sorting rule after completing a category (see (Flashman et al., 1991) for details). Non-perseverative errors are all the errors which are not perseverative, such as casual errors. The scoring measure NCC tells us how many times the agent collects a series of 10 consecutive correct responses. The fact that NCC $= 6$ means that the agent has accomplished the task. However, agents can differ in the number of trials needed to complete six categories. The scoring TFC tells us how many trials the agent needs in order to achieve the first sorting principle, and can be seen as an index of conceptual ability (Anderson, 2008; Singh et al., 2017). A greater attention must be paid to the scoring measure FMS. Failing to maintain a set can be considered a complex behaviour, and which executive component it would actually measure is still debated. For instance, some argue that FMS reflects distractibility characteristics (Barceló and Knight, 2002; Crone et al., 2004) whilst others suggest that it is associated to cognitive flexibility (Zabelina and Robinson, 2010; Greve et al., 2005). In general, a FMS consists in the number of times an individual fails to sort cards by the sorting rule after it

can be determined that he/she has acquired the rule. A given sorting rule is assumed to be acquired when the individual correctly sorts at least five cards in a row (Heaton, 1981; Figueroa and Youmans, 2013). Thus, a failure to maintain a set arises whenever a sorting strategy is changed before this change is appropriate. In our application the FMS scoring measure is computed by collecting the occurrences of first errors after the acquisition of a rule.

4.3.2 Experiment 1

In this experiment, we investigate how the computational agent's response patterns are affected by modulating the levels of flexibility and distractibility of the cognitive system. From a psychological perspective, we are referring to two cognitive constructs: (1) the cognitive flexibility construct, which relates to the ability of an individual to assign new attributes to stimuli (Scott, 1962), and (2) the distractibility construct, which reflects the person's (in)ability to maintain focus on a task (Barceló and Knight, 2002; Crone et al., 2004). In the context of the WCST, this means that the flexibility of the system fosters the agent to change the sorting rule after completing a category (Coulacoglou and Saklofske, 2017). When impaired, it is expected to promote a perseverative behaviour in terms of increase in the number of perseverative responses on the test. This is in line with the main theoretical proposal that cognitive flexibility is related to the degree of perseverative processing of previously relevant representations (Maes et al., 2004; Chevalier and Blaye, 2008). On the other hand, distractibility is supposed to affect the general behaviour of the agent by affecting its ability to provide consistent responses across the entire trials span. We expect distractibility to increase the probability of the system to lose focus on maintaining current task relevant goals (Figueroa and Youmans, 2013), such as a given sorting principle.

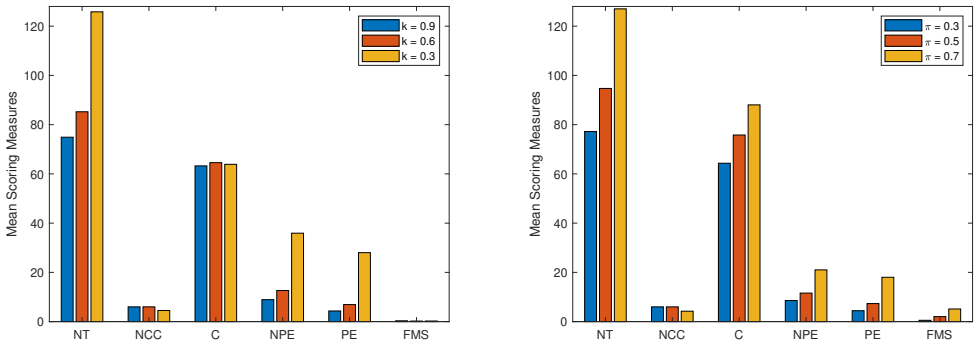
To test these predictions, we rely on two simplified factorial designs in which flexibility and distractibility parameters, k and π, are modulated independently in order to reproduce several degrees of cognitive function impairments. More precisely, in the first simulation design, k can take values on three levels representing no flexibility impairment ($k = 0.9$), mild flexibility impairment ($k = 0.6$), and severe flexibility impairment ($k = 0.3$). The second simulation design consists in letting π to take values on three levels representing low distractibility ($\pi = 0.3$), mild distractibility ($\pi = 0.5$), and high distractibility ($k = 0.7$). Our aim is to explore the main effect of each executive component modulation on agent's performances to assess the model's ability to account for the distractibility/flexibility dissociation. Hence, for each simulation design, the non-varying parameters are fixed in order to produce an optimal behaviour. In this way, only changes in the flexibility, or distractibility, are responsible for eventual differences in response patterns. For each level of the factorial designs, performances of $N = 100$ cognitive agents are simulated. For each factorial level, mean scoring measures across the N agents' responses are considered. Results are shown in Table 4.2 and 4.3 (Figure 4.5).

TABLE 4.2: Simulations results for the flexibility simulation design.

	Mean Scoring Measures					
Flexibility (k)	NT	NCC	C	NPE	PE	FMS
0.9	74.9	6	63.25	8.92	4.33	0.33
0.6	85.23	6	64.56	12.64	6.92	0.23
0.3	125.82	4.52	63.89	35.92	28.01	0.25

TABLE 4.3: Simulations results for the distractibility simulation design.

	Mean Scoring Measures					
Distractibility (π)	NT	NCC	C	NPE	PE	FMS
0.3	77.2	6	64.33	8.55	4.43	0.54
0.5	94.71	5.99	75.79	11.59	7.33	2.03
0.7	127.03	4.23	88.03	21.01	18.01	5.13



(A) Flexibility simulation design (B) Distractibility simulation design

FIGURE 4.5: Simulations results.

As can be noticed, the number of perseverations increases especially as system's flexibility decreases. This means that the cognitive agent shows a deficient behaviour when switching the sorting rule is needed. Perseverative errors usually occur after the agent completes a given category. In the flexibility simulation study, the fact that the averaged number of completed

categories approaches the task completion criterion, suggests that the computational agent is able to settle to a set-maintenance cognitive state, but fails when transitioning to the other state is required. For these reasons, the impaired agent needs more trials in order to complete the task, and this seems to be only due to a selective impairment of cognitive flexibility. This is also consistent with the fact that the number of failures to maintain the set approaches zero across all the level of the factorial design. Thus, at least in our computational cognitive representation, the scoring measure FMS does not seem to be related to cognitive flexibility. In the second simulation study, although system's flexibility is not impaired, a positive relation between distractibility level and number of perseverative errors is observed. Given the model architecture, this can be explained by the increase in the number of failures to maintain sets which leads the agent to randomly switch attention to a sorting rule which is no longer valid for the current trial. Accordingly, the higher the occurrences of failures to maintain a set during the trials span, the higher the chance to randomly sort a card according to a previous rule. In this case, the averaged FMS scoring measure seems to be selectively affected across all the levels of the factorial design. This suggests that this scoring measure is related to system's distractibility.

4.3.3 Experiment 2

In this experiment, we investigate the expressiveness of the computational model in reproducing behaviour when several combinations of parameters are taken into account. However, due to the high number of parameters involved, we adopt a theoretical psychological criterion in order to select which are the best parameters to include in the factorial design. More precisely, we focus on a particular aspect of cognitive functioning which is of interest for computational research, namely, the reward system (Dehaene and Changeux, 1991; Cella et al., 2014). In our model representation, such system involves both the dynamics of the reward processing, ω_t, and the bias to evaluate the internal feedback value to update the cognitive state $\gamma_{11}(t)$. Here, we consider observed response patterns as a function of the interaction between distractibility, flexibility and reward system modulations. In particular, we consider five cognitive profiles characterized by modulating parameters $\lambda, \alpha_0, \alpha_1, \pi$ and k. We let the parameters to take values on two levels, representing low and high reward updating weight ($\lambda = 0.3$ and $\lambda = 0.9$, respectively), severe and no flexibility impairments ($k = 0.3$ and $k = 0.9$, respectively), low and high distractibility ($\pi = 0.3$ and $\pi = 0.9$). Regarding the parameters of the transition equations, α_0 and α_1, we let the system to be either balanced or conservative ($\alpha_0/\alpha_1 = -1/2$ and $\alpha_0/\alpha_1 = -2/3$). Therefore, we consider five generative models: (1) Optimal Behaviour; (2) low reward updating weight and no flexibility impairment; (3) high reward updating weight and severe flexibility impairment; (4) low reward updating weight and high distractibility; (5) high reward updating weight and high distractibility. For each model, we consider both balanced and conservative cognitive profiles. As for the previous experiment, the non-varying parameters across all the models are

fixed to an optimal status. For each model, performances of $N = 100$ cognitive agents are simulated, and mean scoring measure are computed. Results are shown in Table 4.4.

TABLE 4.4: Simulations results for the Experiment 1.

Cognitive Profile	Bias	Mean Scoring Measures					
		NT	NCC	TFC	NPE	PE	FMS
Model 1	Balanced	72.39	6	12.15	7.74	3.65	0.07
	Conservative	80.06	6	14.29	9.43	4.48	0.78
Model 2	Balanced	92.98	5.99	14.71	12.94	8.72	0.09
	Conservative	128	2.22	67.33	29.56	27.51	0.19
Model 3	Balanced	126.7	3.07	38.68	33.93	33.93	0.02
	Conservative	128	2.53	57.36	37.97	35.31	0.44
Model 4	Balanced	127.88	3.6	42.64	24.25	23.43	3.99
	Conservative	128	1.31	90.5	34.42	32.77	1.3
Model 5	Balanced	125.93	4.37	33.98	20.09	17.09	4.98
	Conservative	127.84	3.99	38.01	20.24	18.48	5.08

On the one hand, the selectivity of the effect of distractibility and flexibility impairments on agent's performances is consistent with the first experiment's results. Modulating the functionality of the reward system does not affect this functional dissociation, as can be noticed by the fact that the mean number of FMS only increases when distractibility is impaired (Model 4 and Model 5). On the other hand, the reward system seems to play a role in modulating the overall quality of agent's performance. In particular, a system's bias toward conservativeness entails an increase in the number of perseverations and, more importantly, an increase in the number of trials needed to complete the first category. Such dysfunctional behaviour is emphasized when the system is particularly slow to update the internal attention to reward due to a low reward updating weight (Model 2 and Model 4).

4.3.4 Discussion of results

The simulation studies show that our computational model is able to account for differential behavioural patterns when parameters are modulated. In particular, we focused on three cognitive aspects of the agent's executive system, namely, cognitive flexibility, distractibility and reward system. In general, agents' performances worsen as one of the executive component is selectively impaired. However, a common performance trend can be observed

in any case: number of perseverations and trials needed to accomplish the task increase as a general indicator of cognitive impairment. In the first experiment, although the general performance indicators (i.e. NT, PE, NPE) suggest a shared behavioural pattern when either distractibility or flexibility are degraded, results tell us that the underlying reasons could be different. This is due to the fact that the variable FMS seems to be selectively affected by distractibility modulations, whilst it does not show association with the level of flexibility of the system. In general, we can assume that flexibility impairments allow the agent to incur in errors at a particular phase of the task (i.e. the trials window after a category completion). In this case, an increase in the number of trials needed to complete the task can be due to inefficient cognitive dynamics within that specific task phase. To clarify this concept, consider Figure 4.6. It shows some of the executive component processes dynamics during the unfolding of the first 50 trials of a WCST. Here, we focus on the occurrences of the total error component (NPE + PE) as a function of feedback processing and cognitive states transitions. The depicted system presents a severe cognitive flexibility impairment ($k = 0.3$). Other parameters are fixed in order to attain an optimal behaviour, as in experiment 1. The error pattern (white/black dots) is clear. The system is able to settle to a set-maintenance cognitive state (green squares), and to complete a category (sequence of black dots). The dysfunctional behaviour occurs within the trials window after completing a category.

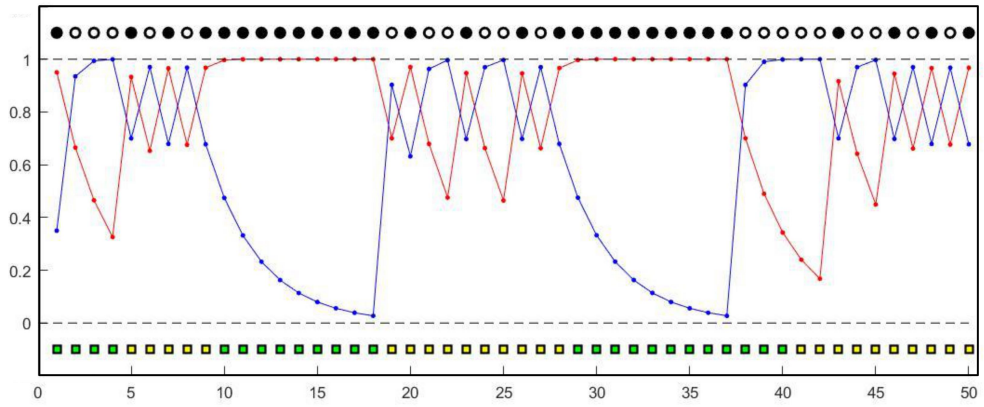


FIGURE 4.6: The figure shows the evolution of feedback processing in time, and the configuration of both cognitive states and responses sequences. Attention to reward, ω_t, is represented by the red function, whilst the attention to punishment, δ_t, is represented by the blue function. Dots represent correct responses (black dots) and general errors (white dots). The sequence of states is represented by green (set-maintenance state) and yellow (set-shifting) squares. The impaired system's flexibility affects the dynamics of the feedback processing as the task unfolds. For instance, after completing a category (sequence of 10 black dots), the agent receives a punishment. At this point, the internal attentional value of the reward slowly drops down, increasing the chance to incur in a series of perseverative and non-perseverative errors.

Differently, we can assume that distractibility impairments allow the agent to incur in errors at any time during the unfolding of the task. In this case, an increase in the number of trials needed to complete the task can be due to dysfunctional behavioural dynamics which are randomly distributed along the trials span. A graphical inspection of the computational behaviour of the system can help clarifying these aspects (Figure 4.7). Here, the system is characterized by an high distractibility ($\pi = 0.7$). Even in this case, other parameters are fixed in order to attain an optimal behaviour. Differently from the previously depicted cognitive agent behaviour, the feedback processing is more efficient, allowing the system to suddenly switch between cognitive states (colored squares) in a consistent way. However, the system seems to be unable to complete a category. Errors occur randomly along the trials span, despite the efficiency of the feedback processing components.

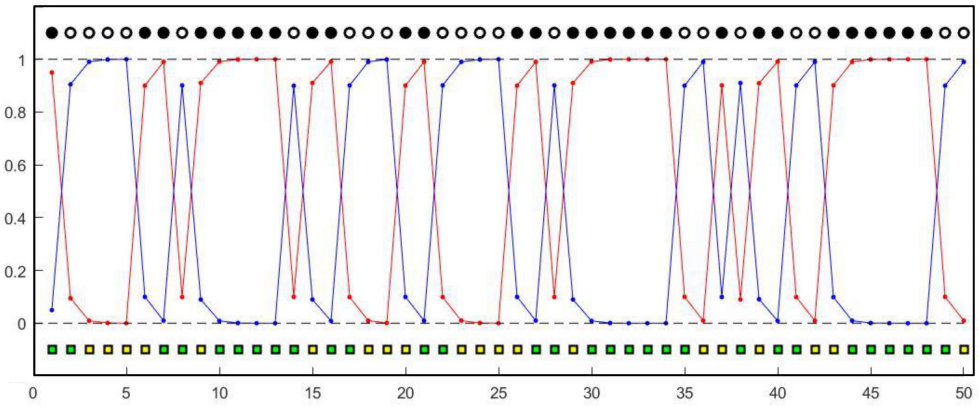


FIGURE 4.7: The figure shows the evolution of feedback processing in time, and the configuration of both cognitive states and responses sequences. Attention to reward, ω_t is represented by the red function, whilst the attention to punishment, δ_t, is represented by the blue function. Dots represent correct responses (black dots) and general errors (white dots). The sequence of states is represented by green (set-maintenance state) and yellow (set-shifting) squares. The impaired system's distractibility affects the responses consistency across the trials. Although the updating of the internal attentional value to feedback shows an efficient behaviour, the system commits errors before completing a category.

Therefore, as outlined in experiment 2, a further source of performance variability consists in the efficiency of the reward system. We can assume that when information processing at such basic functional level is compromised, a global performances deterioration is observed. This seems to be independent of the status of the other main executive components. For instance, consider the aforementioned response styles characterization: the task phase-dependent and the task phase-independent response profiles entailed by flexibility and distractibility components, respectively. A compromised reward system does not directly affect the structural properties of the response patterns characterizing the two profiles (selective effect of distractibility on FMS scoring, Table 4.4). However, an additional error component is thought to further penalize the system in achieving the best possible performance attainable for a given cognitive profile. This is due to the fact that a conservative (and thus, impaired) reward system selectively affects the responsiveness of the system to increase the internal value of a received positive feedback. Thus, errors occur meanwhile the system accumulates reward related information when a positive feedback is processed. We can assume that performances globally deteriorate, as reflected by the overall increase of the error component, due the fact that the system is too slow in updating evidences for the positive feedback (as reflected by increased TFC). At this point, the identification of the real cognitive source of the errors, at a functional level, turns out to be

problematic from a neuropsychological assessment perspective (as we will discuss in the next section).

4.4 General discussion

Set-shifting tasks are highly demanding cognitive settings which require integrated and distributed neural functioning to be accomplished. Cognitive neuropsychology rephrases the problem of investigating the property of such distributed neural architecture into that of accounting for separable cognitive constructs with the aid of tests (such as the WCST) measuring multiple cognitive processes. The presented computational cognitive model relies on evidences from cognitive neuroscience as a starting point to build a model of integrated cognitive functioning which has the potential to account for complex behavioural patterns observed in normal and clinical populations. Several computational models have been proposed to analyse performances in the WCST (Dehaene and Changeux, 1991; Amos, 2000; Monchi et al., 2000; Levine et al., 1993). These models of cognitive functioning differ from our proposed model, since the focus in these models lies in structuring a plausible, biologically inspired, neural network capable of reproducing general qualitative patterns of observed data, such as, simulating different number of errors in different conditions. Our DBN representation is not a neural level model. However, according to our view, it provides a suitable and flexible cognitive description, at a representational and algorithmic level (Marr, 1982), of neural states involved in set-shifting tasks. Therefore, the probabilistic representation of the cognitive system is thought to directly reflect the stochastic nature and the trial-to-trial variability characteristic of neural systems in the brain (Harrison et al., 2005; Buesing et al., 2011; Rolls and Deco, 2010). In this way, such modelling approach retains some of the main features of brain functioning by letting a substrate-independent functional architecture to provide a biologically plausible, but abstract, model of cognitive functioning.

There are two main advantages in adopting our modelling approach to represent executive functions characteristics and, more generally, higher-level cognitive processes. First, a discrete states representation of the cognitive system's functional sub-components provides a flexible way to account for psychological hypothesis about the hierarchical and functional organization of mental processes (Kopp, 2012). On the other hand, our basic probabilistic structure can be easily upgraded by adding more hierarchical cognitive levels and functional nodes in a consistent way. By extension, these two observations make our computational approach also valuable from a theoretical perspective, since, given a set of assumptions, it enables the exploration of a cognitive phenomenon at a given desired level of detail. From this perspective, the model provides a mechanistic, process-based theory of cognitive functioning (Sun, 2009) whose formal instantiation can be useful for empirical research. For instance, in the computational model, we avoided ambiguity in describing what the executive component distractibility is. In our context we assume that the system allocates attention to internally represented concepts (features),

as accounted by the entropy of the probability distributions over the abstract rules. The adopted computational formalisations allow us to explicitly refer to distractibility impairments as the inefficiency to allocate neural resources towards internally directed attention states, a process which involves the recruitment of lateral prefrontal regions and is essential for optimal cognitive performances (Kam et al., 2018; Buckner et al., 2008). This theoretically-driven choice formalizes the kind of process we are aiming to explore. However, different distractibility representations, as well as related theoretical assumptions, can be taken into account. Another way to take advantage of our simple model representation is that of relying on the flexible probabilistic structure to make predictions on cognitive performances in different experimental or environmental settings. For instance, one might be interested in how the system could behave when facing a set-shifting task with an arbitrary number of abstract concepts (features and rules) to functionally operate on. As another example, one might investigate how executive components modulations affect performances when an agent is required to deal with a complex multidimensional features space (e.g. when stimuli have to be sorted by taking into account multiple rules simultaneously). In order to explore these possibilities, the computational formalisations characterizing Level II have to be suitably rewritten in order to adapt to the cognitive setting of interest and to provide a mathematical description of the new environment. At this point, a monte carlo simulation approach could help exploring the role of several executive components on performances in such hypothesised environmental context.

As can be noticed from the simulation results, shared qualitative response patterns can be observed when several executive components are degraded. However, from a generative perspective, it could be easier to provide suitable explanations on the role of examined cognitive components in producing behaviour. Problems might occur when the observed behaviour serves as a basis to infer the characteristics of the underlying cognitive generative process. In this case, neuropsychological criteria based on computation of scoring measures of test outcomes are employed. A great emphasis is placed on the error component since accounting for sub-types of error may help to discriminate cognitive processes that disrupt set-shifting performances in clinical population (Miller et al., 2015). However, our results outlined the fact that general performance indicators, such as errors and number of trials needed to achieve the task, show a similar pattern when either cognitive flexibility, distractibility, and reward system are impaired. In such case, different scoring measures have to be employed in order to investigate more in depth the underlying reasons (e.g. trials to complete the first category (TFC) and failures to maintain set (FMS) scoring measures to account for reward system and distractibility impairments, respectively). Nonetheless, this approach could be too simplistic for the complexity of the phenomenon under investigation. Several profiles of executive component impairments might give rise to response patterns whose properties can only emerge at a micro detailed level of analysis. Thus, given the set of assumptions characterizing the computational cognitive system, the joint analysis of both generated response patterns and simulated executive

components processing dynamics can yield new insights for the development of more sophisticated scoring measures.

As a final consideration, we highlight the potential of our modelling approach in providing an interesting psychometric tool for the formal assessment of the cognitive characteristics of both healthy and non healthy individuals. For instance, parameter estimation procedures on real human data could provide a way to directly measure and quantify cognitive processes at individual or group levels. At this point, parameter estimates could be employed in further statistical analysis of clinical and psychological relevance. However, a detailed analytic study of the model to make it suitable for parameters estimation is left for future work.

Chapter 5

A Bayesian brain model of adaptive functioning under uncertainty

The content of the chapter has been accepted for publication as: D'Alessandro, M., Radev, S., Voss, A., & Lombardi, L. (2020). A Bayesian brain model of adaptive behavior: An application to the Wisconsin Card Sorting Task. *PeerJ*.

5.1 Introduction

Computational models of cognition provide a way to formally describe and empirically account for mechanistic, process-based theories of adaptive cognitive functioning (Sun, 2009; Cooper et al., 1996; Lee and Wagenmakers, 2014). A foundational theoretical framework for describing functional characteristics of neurocognitive systems has recently emerged under the hood of Bayesian brain theories (Knill and Pouget, 2004; Friston, 2010). Bayesian brain theories owe their name to their core assumption that neural computations resemble the principles of Bayesian statistical inference.

In a Bayesian theoretical framework, cognitive agents interact with an uncertain and changeable sensory environment. This requires a cognitive system to infer sensory contingencies based on an internal generative model of the environment. Such a generative model represents subjective hypotheses, or beliefs, about the causal structure of events in the environment (Friston, 2005; Knill and Pouget, 2004) and forms a basis for adaptive behavior. It is assumed that internal beliefs are constantly updated and refined to match the current state of the world as new observations become available. The core idea behind the Bayesian brain hypothesis is that computational mechanisms underlying such an internal belief updating follow the logic of Bayesian probability theory. In this respect, information about the external world provided by sensory inputs is represented as a conditional probability distribution over a set of environmental states. Consequently, the brain relies on this probabilistic representation of the world to infer the most likely environmental causes (states) which generate those inputs, and such a process follows the computational principles of Bayesian inference (Friston and Kiebel, 2009; Friston, 2010; Buckley et al., 2017).

To clarify this concept, consider a simple example of a perceptual task in which a cognitive agent is required to judge whether an item depicted on a flat plane is concave or convex. Its judgment is based solely on the

basis of a set of observed perceptual features, such as, shape, orientation, texture and brightness. Here, the concave-to-convex gradient entails the set of environmental states which must be inferred. The internal generative model of the agent codifies beliefs about how different degrees of convexity might give rise to certain configurations of perceptual inputs. From a Bayesian perspective, the problem is solved by *inverting* the generative model of the environment in order to turn assumptions about how environmental states generate sensory inputs into beliefs about the most likely states (e.g., degree of convexity) given the available sensory information.

Potentially, there are no limitations regarding the complexity of environmental settings (e.g., items and rules in experimental tasks) and cognitive processes to be described in light of the Bayesian brain framework. Indeed, the latter has proven to be a consistent computational modeling paradigm for the investigation of a variety of neurocognitive mechanisms, such as motor control (Friston et al., 2010), oculomotor dynamics (Friston et al., 2012), object recognition (Kersten et al., 2004), attention (Feldman and Friston, 2010), perceptual inference (Petzschner et al., 2015; Knill and Pouget, 2004), multisensory integration (Körding et al., 2007), as well as for providing a foundational theoretical account of general neural systems' functioning (Lee and Mumford, 2003; Friston, 2005; Friston, 2003) and complex clinical scenarios such as Schizophrenia (Stephan et al., 2006), and Autistic Spectrum Disorder (Haker et al., 2016; Lawson et al., 2014). For this reason, such a modeling approach might provide a comprehensive and unified framework under which several cognitive impairments can be measured and understood in the light of a general process-based theory of neural functioning.

In this work, we address the challenging problem of modeling adaptive behavior in a dynamic environment. The empirical assessment of adaptive functioning often relies on dynamic reinforcement learning scenarios which require participants to adapt their behavior during the unfolding of a (possibly) demanding task. Typically, these tasks are designed with the aim to figure out how adaptive behavior unfolds through multiple trials as participants observe certain environmental contingencies, take actions, and receive feedback based on their actions. From a Bayesian theoretical perspective, optimal performance in such adaptive experimental paradigms require that agents infer the probabilistic model underlying the hidden environmental states. Since these models usually change as the task progresses, agents, in turn, need to adapt their inferred model, in order to take optimal actions.

Here, we propose and validate a computational Bayesian model which accounts for the dynamic behavior of cognitive agents in the Wisconsin Card Sorting Test (WCST; (Berg, 1948; Heaton, 1981)), which is perhaps the most widely adopted neuropsychological setting employed to investigate adaptive functioning, due to its specificity in accounting for executive components underlying observed behavior, such as set-shifting, cognitive flexibility and impulsive response modulation (Bishara et al., 2010; Alvarez and Emory, 2006). For this reason, we consider the WCST as a fundamental paradigm for investigating adaptive behavior from a Bayesian perspective.

The environment of the WCST consists of a target and a set of stimulus

cards with geometric figures which vary according to three perceptual features. The WCST requires participants to infer the correct classification principle by trial and error using the examiner's feedback. The feedback is thought to carry a positive or negative information signaling the agent whether the immediate action was appropriate or not. Modeling adaptive behavior in the WCST from a Bayesian perspective is straightforward, since observable actions emerge from the interaction between the internal probabilistic model of the agent and a set of discrete environmental states.

Performance in WCST is usually measured via a rough summary metric such as the number of correct/incorrect responses or pre-defined psychological scoring criteria (see for instance (Heaton, 1981)). These metrics are then used to infer the underlying cognitive processes involved in the task. A major shortcoming of this approach is that it simply assumes the cognitive processes to be inferred without specifying an explicit *process model*. Moreover, summary measures do not utilize the full information present in the data, such as trial-by-trial fluctuations or various interesting agent-environment interactions. For this reason, crude scoring measures are often insufficient to disentangle the dynamics of the relevant cognitive (sub)processes involved. Consequently, an entanglement between processes at the metric level can prevent us from answering interesting research questions about aspects of adaptive behavior.

In our view, a sound computational account for adaptive behavior in the WCST needs to provide at least a quantitative measure of effective belief updating about the environmental states at each trial. This measure should be complemented by a measure of how feedback-related information influences behavior. The first measure should account for the integration of meaningful information. In other words, it should describe how prior beliefs about the current environmental state change after an observation has been made. The second measure should account for signaling the (im)probability of observing a certain environmental configuration (e.g., an (un)expected feedback given a response) (Schwartenbeck et al., 2016).

Indeed, recent studies suggest that the meaningful information content and the pure unexpectedness of an observation are processed differently at the neural level. Moreover, such disentanglement appears to be of crucial importance to the understanding of how new information influences adaptive behavior (Nour et al., 2018; Schwartenbeck et al., 2016; O'Reilly et al., 2013). Inspired by these results and previous computational proposals (Koechlin and Summerfield, 2007), we integrate these different information processing aspects into the current model from an information-theoretic perspective.

Our computational cognitive model draws heavily on the mathematical frameworks of Bayesian probability theory and information theory (Sayood, 2018). First, it provides a parsimonious description of observed data in the WCST via two neurocognitively meaningful parameters, namely, *flexibility* and *information loss* (to be motivated and explained in the model section). Moreover, it captures the main response patterns obtainable in the WCST via different parameter configurations. Second, we formulate a functional

connection between cognitive parameters and underlying information processing mechanisms related to belief updating and prediction formation. We formalize and distinguish between *Bayesian surprise* and *Shannon surprise* as the main mechanisms for adaptive belief updating. Moreover, we introduce a third quantity, which we named predictive *Entropy* and which quantifies an agent's subjective uncertainty about the current internal model. Finally, we propose to measure these quantities on a trial-by-trial basis and use them as a proxy for formally representing the dynamic interplay between agents and environments.

The rest of the work is organized as follows. First, the WCST is described in more detail and a mathematical representation of the new Bayesian computational model is provided. Afterwards, we explore the model's characteristics through simulations and perform parameter recovery on simulated data using a powerful Bayesian deep neural network method (Radev et al., 2020). We then apply the model to real behavioral data from an already published dataset. Finally, we discuss the results as well as the main strengths and limitations of the proposed model.

5.2 The Wisconsin Card Sorting Test

In a typical WCST (Heaton, 1981; Berg, 1948), participants learn to pay attention and respond to relevant stimulus features, while ignoring irrelevant ones, as a function of experimental feedback. In particular, Individuals are asked to match a target card with one of four stimulus cards according to a proper sorting principle, or sorting rule. Each card depicts geometric figures that vary in terms of three features, namely, color (red, green, blue, yellow), shape (triangle, star, cross, circle) and number of objects (1, 2, 3 and 4). For each trial, the participant is required to identify the sorting rule which is valid for that trial, that is, which of the three feature has to be considered as a criterion to matching the target card with the right stimulus card (see Figure 5.1). Notice that both features and sorting rules refer to the same concept. However, the feature still codifies a property of the card, whilst the sorting rule refers to the particular feature which is valid for the current trial.

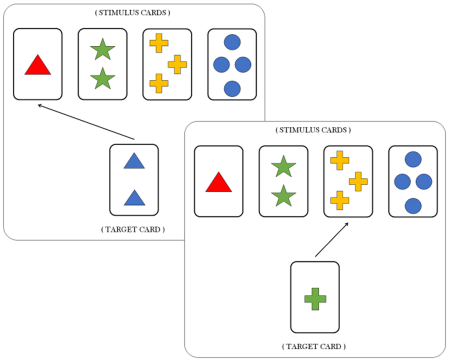


FIGURE 5.1: Suppose that the current sorting rule is the feature shape. The target card in the first trial (left box) contains two blue triangles. A correct response requires that the agent matches the target card with the stimulus card containing the single triangle (arrow represents the correct choice), regardless of the features color and number. The same applies for the second trial (right box) in which matching the target card with the stimulus card containing three yellow crosses is the correct response.

Each response in the WCST is followed by a feedback informing the participant if his/her response is correct or incorrect. After some fixed number of consecutive responses, the sorting rule is changed by the experimenter without warning, and participants are required to infer the new sorting rule. Clearly, the most adaptive response would be to explore the remaining possible rules. However, participants sometimes would persist responding according to the old rule and produce what is called a *perseverative response*.

5.3 Methods

5.3.1 The Model

The core idea behind our computational framework is to encode the concept of *belief* into a generative probabilistic model of the environment. Belief updating then corresponds to recursive Bayesian updating of the internal model based on current and past interactions between the agent and its environment. Optimal or sub-optimal actions are selected according to a well specified or a misspecified internal model and, in turn, cause perceptible changes in the environment.

We assume that the cognitive agent aims to infer the *true hidden state* of the environment by processing and integrating sensory information from the environment. Within the context of the WCST, the hidden environmental states might change as a function of both the structure of the task and the (often sub-optimal) behavioral dynamics, so the agent constantly needs to rely on environmental feedback and own actions to infer the current state. We assume that the agent maintains an internal probability distribution over the states at each individual trial of the WCST. The agent then updates this distribution upon making new observations. In particular, the hidden environmental

states to be inferred are the three features, $s_t \in \{1, 2, 3\}$, which refer the three possible sorting rules in the task environment such that 1: color, 2: shape and 3: number of objects. The posterior probability of the states depends on an observation vector $\boldsymbol{x}_t = (a_t, f_t)$, which consists of the pair of agent's response $a_t \in \{1, 2, 3, 4\}$, codifying the action of choosing deck 1, 2, 3 or 4, and received feedback $f_t \in \{0, 1\}$, referring to the fact that a given response results in a failure (0) or in a success (1), in a given trial $t = 0, ..., T$. The discrete response a_t represents the stimulus card indicator being matched with a target card at trial t. We denote a sequence of observations as $\boldsymbol{x}_{0:t} = (\boldsymbol{x}_0, \boldsymbol{x}_1, ..., \boldsymbol{x}_t) = ((a_0, f_0), (a_1, f_1), (a_2, f_2), ..., (a_t, f_t))$ and set $\boldsymbol{x}_0 = \varnothing$ in order to indicate that there are no observations at the onset of the task. Thus, trial-by-trial belief updating is recursively computed according to Bayes' rule:

$$p(s_t|\boldsymbol{x}_{0:t}) = \frac{p(\boldsymbol{x}_t|s_t, \boldsymbol{x}_{0:t-1})p(s_t|\boldsymbol{x}_{0:t-1})}{p(\boldsymbol{x}_t|\boldsymbol{x}_{0:t-1})} \tag{5.1}$$

Accordingly, the agent's posterior belief about the task-relevant features s_t after observing a sequence of response-feedback pairs $\boldsymbol{x}_{0:t}$ is proportional to the product of the likelihood of observing a particular response-feedback pair and the agent's prior belief about the task-relevant feature in the current trial. The likelihood of an observation is computed as follows:

$$p(\boldsymbol{x}_t|s_t, \boldsymbol{x}_{0:t-1}) = \frac{f_t p(a_t|s_t = i) + (1 - f_t)(1 - p(a_t|s_t = i))}{f_t \sum_j p(a_t|s_t = j) + (1 - f_t) \sum_j (1 - p(a_t|s_t = j))} \tag{5.2}$$

where $j = 1, 2, 3$ and $p(a_t|s_t = i)$ indicates the probability of a matching between the target and the stimulus card assumed that the current feature is i. Here, we assume the likelihood of a current observation to be independent from previous observations without loss of generality, that is:

$$p(\boldsymbol{x}_t|s_t, \boldsymbol{x}_{0:t-1}) = p(\boldsymbol{x}_t|s_t)$$

The prior belief for a given trial t is computed based on the posterior belief generated in the previous trial, $p(s_{t-1}|\boldsymbol{x}_{0:t-1})$, and the agent's belief about the probability of transitions between the hidden states, $p(s_t|s_{t-1})$. The prior belief can also be considered as a predictive probability over the hidden states. The predictive distribution for an upcoming trial t is computed according to the Chapman-Kolmogorov equation:

$$p(s_{t+1} = k|\boldsymbol{x}_{0:t}) = \sum_{i=1}^{3} p(s_{t+1} = k|s_t = i, \boldsymbol{\Gamma}(t))p(s_t = i|\boldsymbol{x}_{0:t}) \tag{5.3}$$

where $\boldsymbol{\Gamma}(t)$ represents a stability matrix describing transitions between the states (to be explained shortly). Thus, the agent combines information from the updated belief (posterior distribution) and the belief about the transition properties of the environmental states to predict the most probable future

state. The predictive distribution represents the internal model of the cognitive agent according to which actions are generated.

The stability matrix $\mathbf{\Gamma}(t)$ encodes the agent's belief about the probability of states being stable or likely to change in the next trial. In other words, the stability matrix reflects the cognitive agent's internal representation of the dynamic probabilistic model of the task environment. It is computed on each trial based on the response-feedback pair, $\boldsymbol{x}_t$, and a matching signal, $\boldsymbol{m}_t$, which are observed.

The matching signal $\boldsymbol{m}_t$ is a vector informing the cognitive agent which features are currently relevant (meaningful), such that $m_t^{(i)} = 1$ when a positive feedback is associated with a response implying feature $s_t = i$, and $m_t^{(i)} = 0$ otherwise. Note, that the matching signal is not a free parameter of the model, but is completely determined by the task contingencies. The matching signal vector allows the agent to compute the *state activation level* $\omega_t^{(i)} \in [0, 1]$ for the hidden state $s_t = i$, which provides an internal measure of the (accumulated) evidence for each hidden state at trial t. Thus, the activation levels of the hidden states are represented by a vector $\boldsymbol{\omega}_t$. The stability matrix is a square and asymmetric matrix related to hidden state activation levels such that:

$$\mathbf{\Gamma}(t) = \begin{bmatrix} \omega_t^{(1)} & \frac{1}{2}(1-\omega_t^{(1)}) & \frac{1}{2}(1-\omega_t^{(1)}) \\ \frac{1}{2}(1-\omega_t^{(2)}) & \omega_t^{(2)} & \frac{1}{2}(1-\omega_t^{(2)}) \\ \frac{1}{2}(1-\omega_t^{(3)}) & \frac{1}{2}(1-\omega_t^{(3)}) & \omega_t^{(3)} \end{bmatrix} \tag{5.4}$$

where the entries $\Gamma_{ii}(t)$ in the main diagonal represent the elements of the activation vector $\boldsymbol{\omega}_t$, and the non-diagonal elements are computed so as to ensure that rows sum to 1. The state activation vector is computed in each trial as follows:

$$\begin{bmatrix} \omega_t^{(1)} \\ \omega_t^{(2)} \\ \omega_t^{(3)} \end{bmatrix} = f_t \boldsymbol{\omega}_{t-1}^{\delta} \begin{bmatrix} m_t^{(1)} \\ m_t^{(2)} \\ m_t^{(3)} \end{bmatrix} + \lambda \begin{bmatrix} (1-f_t)\boldsymbol{\omega}_{t-1}^{\delta} \begin{bmatrix} 1-m_t^{(1)} \\ 1-m_t^{(2)} \\ 1-m_t^{(3)} \end{bmatrix} \end{bmatrix} \begin{bmatrix} \omega_{t-1}^{(1)} \\ \omega_{t-1}^{(2)} \\ \omega_{t-1}^{(3)} \end{bmatrix}. \tag{5.5}$$

This equation reflects the idea that state activations are simultaneously affected by the observed feedback, f_t, and the matching signal vector, $\boldsymbol{m}_t$. However, the matching signal vector conveys different information based on the current feedback. Matching a target card with a stimulus card makes a feature (or a subset of features) informative for a specific state. The vector $\boldsymbol{m}_t$ contributes to increase the activation level of a state if the feature is informative for that state when a positive feedback is received, as well as to decrease the activation level when a negative feedback is received.

The parameter $\lambda \in [0, 1]$ modulates the efficiency to disengage attention to a given state-activation configuration when a negative feedback is processed. We therefore term this parameter *flexibility*. We also assume that information from the matching signal vector can degrade by slowing down the rate of evidence accumulation for the hidden states. This means that the matching

signal vector can be re-scaled based on the current state activation level. The parameter $\delta \in [0, 1]$ is introduced to achieve this re-scaling. When $\delta = 0$, there is no re-scaling and updating of the state activation levels relies on the entire information conveyed by $\boldsymbol{m}_t$. On the other extreme, when $\delta = 1$, several trials have to be accomplished before converging to a given configuration of the state activation levels. Equivalently, higher values of δ affect the entropy of the distribution over hidden states by decreasing the probability of sampling of the correct feature. We therefore refer to δ as *information loss*.

The free parameters λ and δ are central to our computational model, since they regulate the rate at which the internal model converges to the true task environmental model. Eq. (5) can be expressed in compact notation as follows:

$$\boldsymbol{\omega}_t = f_t \boldsymbol{\omega}_{t-1}^{\delta} \boldsymbol{m}_t + \lambda \left[(1 - f_t) \boldsymbol{\omega}_{t-1}^{\delta} (1 - \boldsymbol{m}_t) \right] \boldsymbol{\omega}_{t-1} \tag{5.6}$$

Note that the information loss parameter δ affects the amount of information that a cognitive agent acquires from environmental contingencies, irrespective of the type of feedback received. Global information loss thus affects the rate at which the divergence between the agent's internal model and the true model is minimized. Figure 5.2 illustrates these ideas.

The probabilistic representation of adaptive behaviour provided by our Bayesian agent model allows us to quantify latent cognitive dynamics by means of meaningful information-theoretic measures. Information theory has proven to be an effective and natural mathematical language to account for functional integration of structured cognitive processes and to relate them to brain activity (Koechlin and Summerfield, 2007; Friston et al., 2017a; Collell and Fauquet, 2015; Strange et al., 2005; Friston, 2003). In particular, we are interested in three key measures, namely, *Bayesian surprise*, $\mathcal{B}_t$, *Shannon surprise*, $\mathcal{I}_t$, and *entropy*, $\mathcal{H}_t$. The subscript t indicates that we can compute each quantity on a trial-by-trial basis. Each quantity is amenable to a specific interpretation in terms of separate neurocognitive processes. Bayesian surprise $\mathcal{B}_t$ quantifies the magnitude of the update from prior belief to posterior belief. Shannon surprise $\mathcal{I}_t$ quantifies the improbability of an observation given an agent's prior expectation. Finally, entropy $\mathcal{H}_t$ measures the degree of epistemic uncertainty regarding the true environmental states. Such measures are thought to account for the ability of the agent to manage uncertainty as emerging as a function of competing behavioral affordances (Hirsh et al., 2012). We expect an adaptive system to attenuate uncertainty over environmental states (current features) by reducing the entropy of its internal probabilistic model.

Bayesian surprise can be computed as the Kullback–Leibler ($\mathbb{KL}$) divergence between prior and posterior beliefs about the environmental states. Thus, Bayesian surprise accounts for the divergence between the predictive model for the current trial and the updated predictive model for the upcoming trial. It is computed as follows:

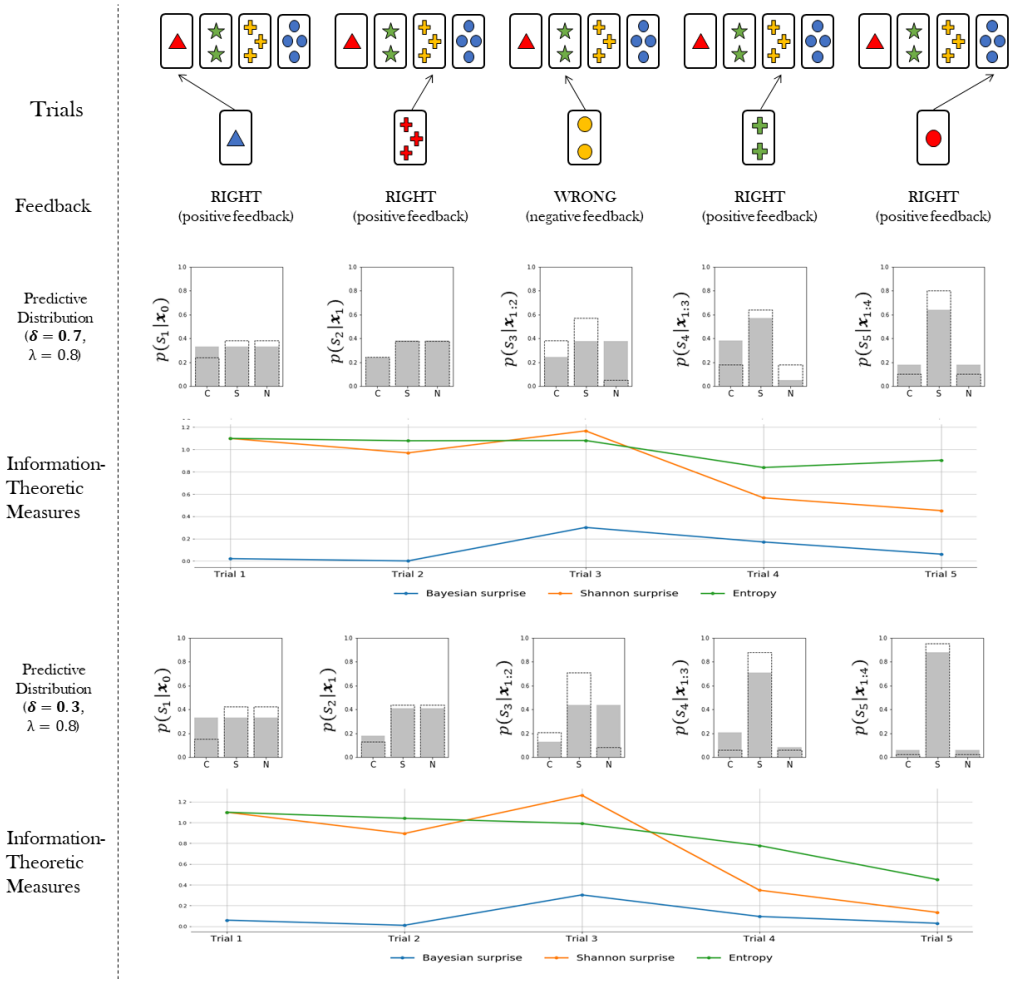


FIGURE 5.2: Suppose the correct sorting rule is the feature shape. The figure shows the rate of convergence of the predictive distributions to the true task environmental model. The predictive distributions at trial $t + 1$ depends on the sorting action a_t (first row) and the received feedback f_t (second row). Two examples of updating a predictive distribution are shown: one in which information loss is high ($\delta = 0.7$, third row), and one in which information loss is low ($\delta = 0.3$, fifth row). High information loss slows down the convergence of the internal model to the true environmental model. The gray bar plots represent the predictive probability distribution over the rules from which an action is sampled at each trial. Dotted bars represent the updated predictive distribution after the feedback observation. For each scenario, trial-by-trial information-theoretic measures are shown.

$$\begin{aligned}\mathcal{B}_t &= \mathbb{KL}[p(s_{t+1}|\boldsymbol{x}_{0:t})||p(s_t|\boldsymbol{x}_{0:t-1})] \\ &= \sum_{i=1}^{3}\left[p(s_{t+1}=i|\boldsymbol{x}_{0:t})\log\left(\frac{p(s_{t+1}=i|\boldsymbol{x}_{0:t})}{p(s_t=i|\boldsymbol{x}_{0:t-1})}\right)\right]\end{aligned} \tag{5.7}$$

The Shannon surprise of a current observation given a previous one is computed as the conditional information content of the observation:

$$\begin{aligned}\mathcal{I}_t &= -\log p(\boldsymbol{x}_t|\boldsymbol{x}_{0:t-1}) \\ &= -\log\sum_{i=1}^{3}[p(\boldsymbol{x}_t|s_t=i)p(s_t=i|\boldsymbol{x}_{0:t-1})]\end{aligned} \tag{5.8}$$

Finally, the entropy is computed over the predictive distribution in order to account for the uncertainty in the internal model of the agent in trial t as follows:

$$\begin{aligned}\mathcal{H}_t &= \mathbb{E}\left[-\log p(s_t|\boldsymbol{x}_{0:t-1})\right] \\ &= -\sum_{i=1}^{3}p(s_t=i|\boldsymbol{x}_{0:t-1})\log p(s_t=i|\boldsymbol{x}_{0:t-1})\end{aligned} \tag{5.9}$$

Once the flexibility (λ) and information loss (δ) parameters are estimated from data, the information-theoretic quantities can be easily computed and visualized for each trial of the WCST (see Figure 5.2). This allows to rephrase standard neurocognitive constructs in terms of measurable information-theoretic quantities. Moreover, the dynamics of these quantities, as well as their interactions, can be used for formulating and testing hypotheses about the neurocognitive underpinnings of adaptive behavior in a principled way, as discussed later in the text. A summary of all quantities relevant for our computational model is provided in Table 5.1.

5.3.2 Simulations

In this section we evaluate the expressiveness of the model by assessing its ability to reproduce meaningful behavioral patterns as a function of its two free parameters. We study how the generative model behaves when performing the WCST in a 2-factorial simulated Monte Carlo design where flexibility (λ) and information loss (δ) are systematically varied.

In this simulation, the Heaton version of the task (Heaton, 1981) is administered to the Bayesian cognitive agent. In this particular version, the sorting rule (true environmental state) changes after a fixed number of consecutive correct responses. In particular, when the agent correctly matches the target card in 10 consecutive trials, the sorting rule is automatically changed. The task ends after completing a maximum of 128 trials.

Expression	Name	Description
$s_t \in \{1,2,3\}$	Sorting rule	Card feature relevant for the sorting criterion in trial t.
$a_t \in \{1,2,3,4\}$	Choice action	Action of choosing one of the four stimulus cards in trial t.
$f_t \in \{0,1\}$	Feedback	Indicates whether the action of matching a stimulus to a target card is correct or not in trial t.
$x_t = (a_t, f_t)$	Observation	Pair of action and feedback which constitutes the agent's observation in trial t.
$\mathbf{\Gamma}(t)$	Stability matrix	Matrix encoding the agent's beliefs about state transitions from trial t to the next trial $t+1$.
$\lambda \in [0,1]$	Flexibility	Parameter encoding the efficiency to disengage attention from a currently attended hidden state when signaled by the environment.
$\delta \in [0,1]$	Information loss	Parameter encoding how efficiently the agent's internal model converges to the true environmental model based on experience.
$m_t^{(i)} \in \{0,1\}$	Matching signal	Signal indicating whether feature i is relevant in trial t based on the feedback received.
$\omega_t^{(i)} \in [0,1]$	State activation level	Agent's internal measure of the accrued evidence for the hidden environmental state i in trial t.
$\mathcal{B}_t \in \mathbb{R}^+$	Bayesian surprise	Kullback-Leibler divergence between prior and posterior beliefs about hidden environmental states in trial t.
$\mathcal{I}_t \in \mathbb{R}^+$	Shannon surprise	Information-theoretic surprise encoding the improbability or unexpectedness of an observation in trial t.
$\mathcal{H}_t \in \mathbb{R}^+$	Entropy	Degree of epistemic uncertainty in the internal model of the environment in trial t.

TABLE 5.1: Descriptive summary of all quantities involved in our model representation.

Generative Model

The cognitive agent's responses are generated at each time step (trial) by processing the experimental feedback. Its performance depends on the parameters governing the computation of the relevant quantities. The generative algorithm is outlined in **Algorithm 1**.

Algorithm 1 Bayesian cognitive agent

Set parameters $\boldsymbol{\theta} = (\lambda, \delta)$.
Set initial activation levels $\boldsymbol{\omega}_0 = (0.5, 0.5, 0.5)$.
Set initial observation $\boldsymbol{x}_0 = \varnothing$ and $p(s_1|\boldsymbol{x}_0) = p(s_1)$.
for $t = 1, ..., T$ **do**
    Sample feature from prior/predictive internal model $s_t \sim p(s_t|\boldsymbol{x}_{0:t-1})$.
    Obtain a new observation $\boldsymbol{x}_t = (a_t, f_t)$.
    Compute state posterior $p(s_t|\boldsymbol{x}_{0:t})$.
    Compute new activation levels $\boldsymbol{\omega}_t$.
    Compute stability matrix $\boldsymbol{\Gamma}(t)$.
    Update prior/predictive internal model to $p(s_{t+1}|\boldsymbol{x}_{0:t})$.
end for

Simulation 1: Clinical Assessment of the Bayesian Agent

Ideally, the qualitative performance of the Bayesian cognitive agent will resemble human performance. To this aim, we adopt a metric which is usually employed in clinical assessment of test results in neurological and psychiatric patients (Braff et al., 1991; Zakzanis, 1998; Bechara and Damasio, 2002; Landry and Al-Taie, 2016). Thus, agent performance is codified according to a neuropsychological criterion (Heaton, 1981; Flashman et al., 1991) which allows to classify responses into several response types. These response types provide the scoring measures for the test.

Here, we are interested in: 1) non-perseverative errors (E); 2) perseverative errors (PE); 3) number of trials to complete the first category (TFC); and 4) number of failures to maintain set (FMS). Perseverative errors occur when the agent applies a sorting rule which was valid before the rule has been changed. Usually, detecting a perseveration error is far from trivial, since several response configurations could be observed when individuals are required to shift a sorting rule after completing a category (see (Flashman et al., 1991) for details). On the other hand, non-perseverative errors refer to all errors which do not fit the above description, or in other words, do not occur as a function of changing the sorting rule, such as casual errors.

The number of trials to complete the first category tells us how many trials the agent needs in order to achieve the first sorting principle, and can be seen as an index of conceptual ability (Anderson, 2010; Singh et al., 2017). Finally, a failure to maintain a set occurs when the agent fails to match cards according to the sorting rule after it can be determined that the agent has acquired the rule. A given sorting rule is assumed to be acquired when the individual correctly sorts at least five cards in a row (Heaton, 1981; Figueroa

and Youmans, 2013). Thus, a failure to maintain a set arises whenever a participant suddenly changes the sorting strategy in the absence of negative feedback. Failures to maintain a set are mostly attributed to distractibility. We compute this measure by counting the occurrences of first errors after the acquisition of a rule.
We run the generative model by varying flexibility across four levels, $\lambda \in \{0.3, 0.5, 0.7, 0.9\}$, and information loss across three levels, $\delta \in \{0.4, 0.7, 0.9\}$. We generate data from 150 synthetic cognitive agents per parameter combination and compute standard scoring measures for each of the agents simulated responses. Results from the simulation runs are depicted in Table 5.2 and a graphical representation is provided in Figure 5.3.

Measure	Info. Loss (δ)	Flexibility (λ)			
		$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.7$	$\lambda = 0.9$
E	$\delta = 0.4$	9.07 (2.68)	7.95 (2.07)	7.50 (2.13)	6.85 (1.75)
	$\delta = 0.7$	10.84 (2.35)	9.60 (2.2)	8.25 (2.23)	7.37 (1,74)
	$\delta = 0.9$	12.75 (2.96)	11.25 (2.43)	9.12 (2.09)	7.79 (1.73)
PE	$\delta = 0.4$	20.81 (2.27)	18.18 (1.88)	14.99 (1.88)	12.37 (1.12)
	$\delta = 0.7$	19.77 (2.55)	17.65 (2.26)	15.42 (1.94)	12.39 (1.47)
	$\delta = 0.9$	18.56 (2.76)	16.58 (2.53)	14.49 (2.03)	12.33 (1.44)
TFC	$\delta = 0.4$	12.20 (1.46)	11.91 (1.35)	11.83 (1.24)	11.67 (1.04)
	$\delta = 0.7$	13.82 (2.76)	13.32 (2.52)	12.97 (2.13)	12.29 (1.53)
	$\delta = 0.9$	17.27 (4.21)	16.63 (4.04)	14.39 (3.58)	12.91 (1.91)
FMS	$\delta = 0.4$	0.11 (0.31)	0.09 (0.31)	0.05 (0.32)	0.02 (0.14)
	$\delta = 0.7$	1.65 (1.4)	1.41 (1.3)	0.84 (0.91)	0.35 (0.69)
	$\delta = 0.9$	4.44 (1.96)	3.88 (1.86)	2.79 (1.56)	1.54 (1.25)

TABLE 5.2: Mean clinical scoring measures as functions of flexibility (λ) and information loss (δ). Cells show the average scores across simulated agents (standard deviation is shown in parenthesis).

The simulated performance of our Bayesian cognitive agents demonstrates that different parameter combinations capture different meaningful behavioral patterns. In other words, flexibility and information loss seem to interact in a theoretically meaningful way.

First, overall errors increase when flexibility (λ) decreases, which is reflected by the inverse relation between the number of casual, as well as perseverative, errors and the values of parameter λ. Moreover, this pattern is consistent across all the levels of parameter δ. More precisely, information loss (δ) seems to contribute to the characterization of the casual and the perseverative components of the error in a different way. Perseverative errors are likely to occur after a sorting rule has changed and reflect the inability of the agent to use feedback to disengage attention from the currently attended

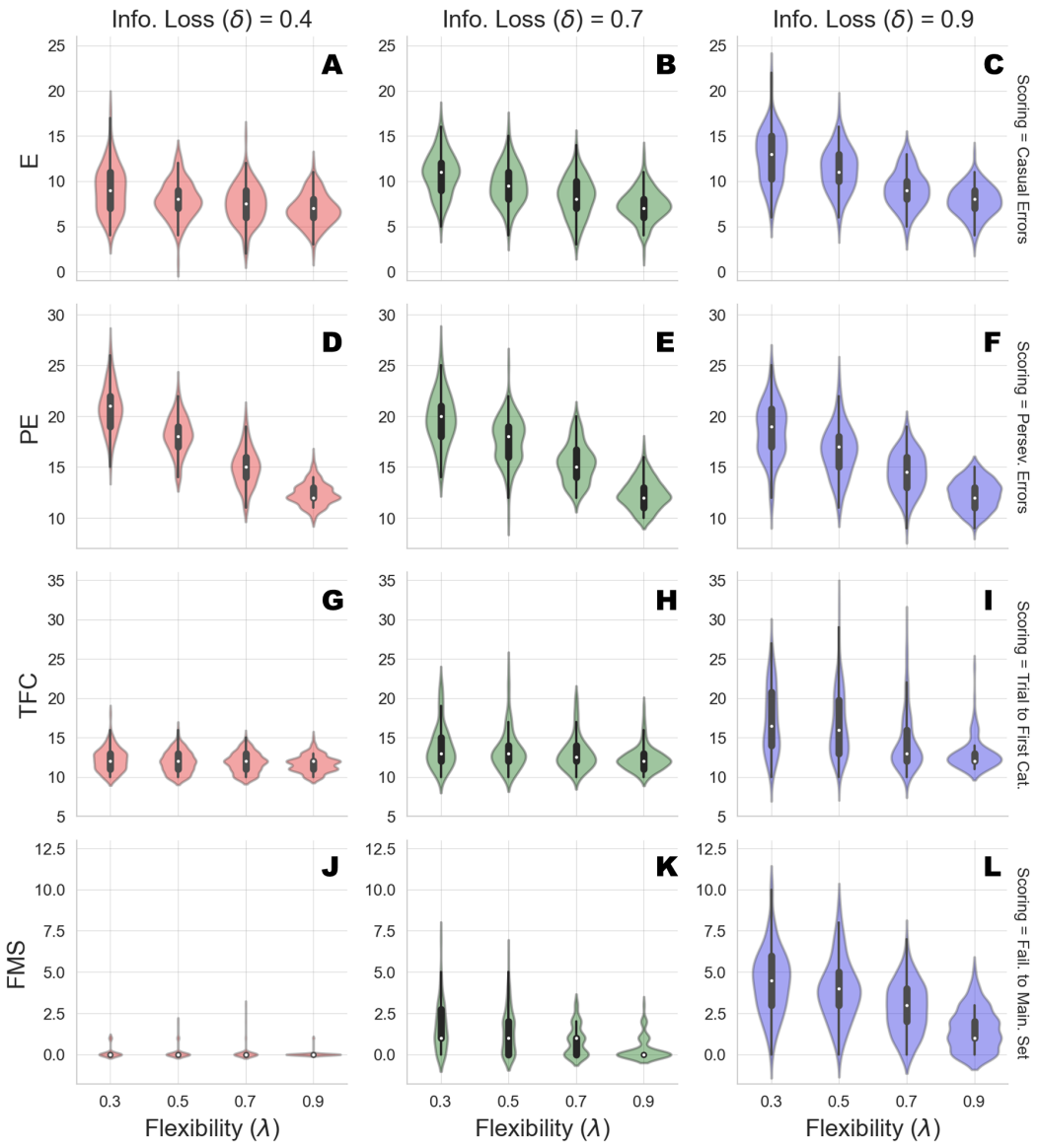


FIGURE 5.3: Clinical scoring measures as functions of flexibility (λ) and information loss (δ) - simulated scenarios. The different cells show the violin plots for the estimated distribution densities of the scoring measures obtained from the group of synthetic individuals, for the levels of λ across different levels of δ. In particular, they show the distribution of non-perseverative errors (E), perseverative errors (PE), number of trials to complete the first category (TFC), number of failures to maintain set (FMS) obtained from 150 synthetic agent's response simulations for each cell of the factorial design.

feature. They therefore result from local cognitive dynamics conditioned on a particular stage of the task (e.g., after completing a series of correct responses).

Second, information loss does not interact with flexibility when perseverative errors are considered. This is due to the fact that high information loss affects general performance by yielding a dysfunctional response strategy which increases the probability of making an error at any stage of the task. The lack of such interaction provides evidence that our computational model can disentangle between error patterns due to perseveration and those due to general distractibility, according to neuropsychological scoring criteria.

However, in our framework, flexibility (λ) is allowed to yield more general and non-local cognitive dynamics as well. Indeed, λ plays a role whenever belief updating is demanded as a function of negative feedback. An error classified as non-perseverative (e.g., casual error) by the scoring criteria might still be processed as a feedback-related evidence for belief updating. Consistently, the interaction between λ and δ in accounting for causal errors shows that performance worsens when both flexibility and information loss become less optimal, and that such pattern becomes more pronounced for lower values of δ.

On the other hand, a specific effect of information loss (δ) can be observed for the scoring measures related to slow information processing and distractibility. The number of trials to achieve the first category reflects the efficiency of the agent in arriving at the first true environmental model. Flexibility does not contribute meaningfully to the accumulation of errors before completing the first category for some levels of information loss. This is reflected by the fact that the mean number of trials increases as a function of δ, and do not change across levels of λ for low and mid values of δ. A similar pattern applies for failures to maintain a set. Both scoring measures index a deceleration of the process of evidence accumulation for a specific environmental configuration, although the latter is a more exhaustive measures of dysfunctional adaptation.

Therefore, an interaction between parameters can be observed when information loss is high. A slow internal model convergence process increases the amount of errors due to improper rule sampling from the internal environmental model. However, internal model convergence also plays a role when a new category has to be accomplished after completing an older one. On the one hand, compromised flexibility increases the amount of errors due to inefficient feedback processing. This leads to longer trial windows needed to achieve the first category. On the other hand, when information loss is high, belief updating upon negative feedback is compromised due to high internal model uncertainty. At this point, the probability to err due to distractibility increases, as accounted by the failures to maintain a set measures.

Finally, the joint effect of δ and λ for high levels of information loss suggests that the roles played by the two cognitive parameters in accounting for adaptive functioning can be entangled when neuropsychological scoring criteria are considered.

Simulation 2: Information-Theoretic Analysis of the Bayesian Agent

In the following, we explore a different simulation scenario in which information-theoretic measures are derived to assess performance of the Bayesian cognitive agent. In particular, we explore the functional relationship between cognitive parameters and the dynamics of the recovered information-theoretic measures by simulating observed responses by varying flexibility across three levels, $\lambda \in \{0.1, 0.5, 0.9\}$, and information loss across three levels, $\delta \in \{0.1, 0.5, 0.9\}$.

For this simulation scenario, we make no prior assumptions about subtypes of error classification. Instead, we investigate the dynamic interplay between Bayesian surprise, $\mathcal{B}_t$, Shannon surprise, $\mathcal{I}_t$, and entropy, $\mathcal{H}_t$ over the entire course of 128 trials in the WCST.

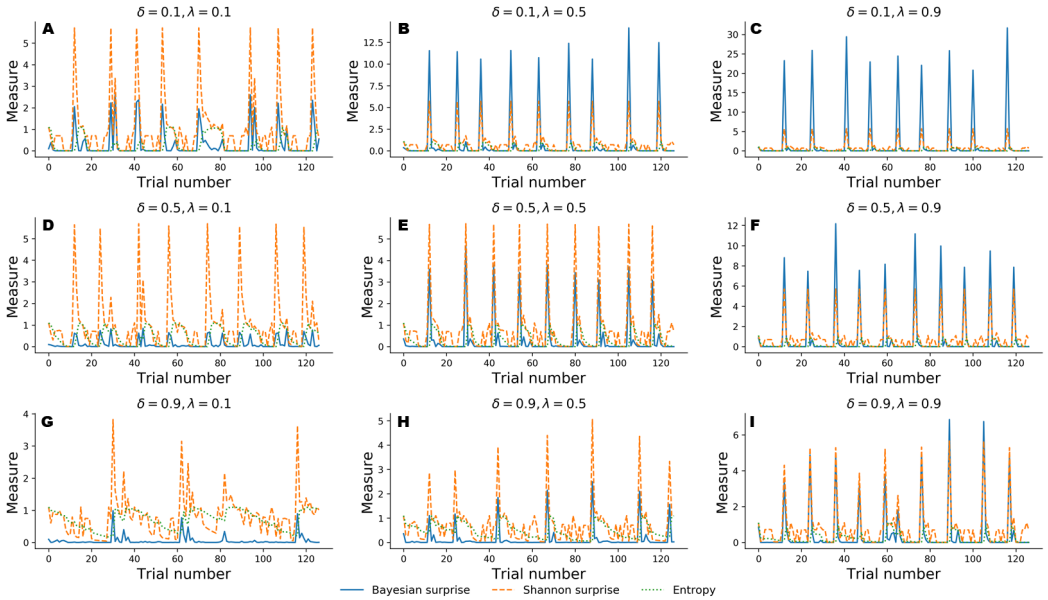


FIGURE 5.4: Information-theoretic measures varying as a function of flexibility λ and information loss δ across 128 trials of the WCST. Optimal belief updating and uncertainty reduction are achieved with low information loss and high flexibility (first row, third column).

Figure 5.4 depicts results from the nine simulation scenarios. Although an exhaustive discussion on cognitive dynamics should couple information-theoretic measures with patterns of correct and error responses, we focus solely on the information-theoretic time series for illustrative purposes. We refer to the application section for a more detailed description of the relation between observed responses and estimated information-theoretic measures in the context of data from a real experiment.

Again, simulated performance of the Bayesian cognitive agent shows that different parameter combinations yield different patterns of cognitive dynamics. Observed spikes and their related magnitudes signal informative task events (e.g., unexpected negative feedback), as accounted by Shannon

surprise, or belief updating, as accounted by Bayesian surprise. Finally, entropy encodes the epistemic uncertainty about the environmental model on a trial-by-trial basis.

In general, low information loss (δ) ensures optimal behavior by speeding up internal model convergence by decreasing the number of trials needed to minimize uncertainty about the environmental states. Low uncertainty reflects two main aspects of adaptive behavior. On the one hand, the probability that a response occurs due to sampling of improper rules decreases, allowing the agent to prevent random responses due to distractibility. On the other hand, model convergence entails a peaked Shannon surprise when a negative feedback occurs, due to the divergence between predicted and actual observations.

Flexibility (λ) plays a crucial role in integrating feedback information in order to enable belief updating. The first row depicted in Figure 5.4 shows cognitive dynamics related to low information loss, across the levels of flexibility. As can be noticed, there is a positive relation between the magnitude of the Bayesian surprise and the level of flexibility, although unexpectedness yields approximately the same amount of signaling, as accounted by peaked Shannon surprise. From this perspective, surprise and belief updating can be considered functionally separable, where the first depends on the particular internal model probability configuration related to δ, whilst the second depends on flexibility λ.

However, more interesting patterns can be observed when information loss increases. In particular, model convergence slows down and several trials are needed to minimize predictive model entropy. Casual errors might occur within trial windows characterized by high uncertainty, and interactions between entropy and Shannon surprise can be observes in such cases. In particular, Shannon surprise magnitude increases when model's entropy decreases, that is, during task phases in which the internal model has already converged. As a consequence, negative feedback could be classified as informative or uninformative, based on the uncertainty in the current internal model. This is reflected by the negative relation between entropy and Shannon surprise, as can be noticed by inspecting the graphs depicted in the third row of Figure 5.4. Therefore, the magnitude of belief updating depends on the interplay between entropy and Shannon surprise, and can differ based on the values of the two measures in a particular task phase.

In sum, both simulation scenarios suggest that the simulated behavior of our generative model is in accord with theoretical expectations. Moreover, the flexibility and information loss parameters can account for a wide range of observed response patterns and inferred dynamics of information processing.

5.3.3 Parameter Estimation

In this section, we discuss the computational framework for estimating the parameters of our model from observed behavioral data. Parameter estimation is essential to inferring the cognitive dynamics underlying observed behavior

in real-world applications of the model. This section is slightly more technical and can be skipped without significantly affecting the flow of the text.

Computational Framework

Rendering our cognitive model suitable for application in real-world contexts also entails accounting for uncertainty about parameter estimates. Indeed, uncertainty quantification turns out to be a fundamental and challenging goal when first-level quantities, that is, cognitive parameter estimates, are used to recover (second-level) information-theoretic measures of cognitive dynamics. The main difficulties arise when model complexity makes estimation and uncertainty quantification intractable at both analytical and numerical levels. For instance, in our case, probability distributions for the hidden model are generated at each trial, and the mapping between hidden states and responses changes depending on the structure of the task environment.

Identifying such a dynamic mapping is relatively easy from a generative perspective, but it becomes challenging, and almost impossible, when inverse modeling is required. Generally, this problem arises when the likelihood function relating model parameters to the data is not available in closed-form or too complex to be practically evaluated (Sisson and Fan, 2011). To overcome these limitations, we apply the first version of the recently developed *BayesFlow* method (see (Radev et al., 2020) for mathematical details). At a high-level, BayesFlow is a simulation-based method that estimates parameters and quantifies estimation uncertainty in a unified Bayesian probabilistic framework when inverting the generative model is intractable. The method is based on recent advances in deep generative modeling and makes no assumptions about the shape of the true parameter posteriors. Thus, our ultimate goal becomes to approximate and analyze the joint posterior distribution over the model parameters. The parameter posterior is given via an application of Bayes' rule:

$$p(\boldsymbol{\theta}|\boldsymbol{x}_{0:T}, \boldsymbol{m}_{0:T}) = \frac{p(\boldsymbol{x}_{0:T}, \boldsymbol{m}_{0:T}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{x}_{0:T}, \boldsymbol{m}_{0:T}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{5.10}$$

where we set $\boldsymbol{\theta} = (\lambda, \delta)$ and stack all observations and matching signals into the vectors $\boldsymbol{x}_{0:T} = (\boldsymbol{x}_0, \boldsymbol{x}_1, ..., \boldsymbol{x}_T)$ and $\boldsymbol{m}_{0:T} = (\boldsymbol{m}_0, \boldsymbol{m}_1, ..., \boldsymbol{m}_T)$, respectively. The BayesFlow method uses simulations from the generative model to optimize a neural density estimator which learns a probabilistic mapping between raw data and parameters. It relies on the fact that data can easily be simulated by repeatedly running the generative model with different parameter configurations $\boldsymbol{\theta}$ sampled from the prior. During training, the neural network estimator iteratively minimizes the divergence between the true posterior and an approximate posterior. Once the network has been trained, we can efficiently obtain samples from the approximate joint posterior distribution of the cognitive parameters of interest, which can be further processed in order to extract meaningful summary statistics (e.g., posterior means, medians, modes, etc.). Importantly, we can apply the same pre-trained inference network to

an arbitrary number of real or simulated data sets (i.e., the training effort *amortizes* over multiple evaluations of the network).

For our purposes of validation and application, we train the network for 50 epochs which amount to 50000 forward simulations. As a prior, we use a bivariate continuous uniform distribution $p(\boldsymbol{\theta}) \sim \mathcal{U}([0,0],[1,1])$. We then validate performance on a separate validation set of 1000 simulated data sets with known *ground-truth* parameter values. Training the networks took less than a day on a single machine with an NVIDIA® GTX1060 graphics card (CUDA version 10.0) using TensorFlow (version 1.13.1) ((Abadi et al., 2016)). In contrast, obtaining full parameter posteriors from the entire validation set took approximately 1.78 seconds. In what follows, we describe and report all performance validation metrics.

Performance Metrics and Validation Results

To assess the accuracy of point estimates, we compute the root mean squared error (RMSE) and the coefficient of determination (R^2) between posterior means and true parameter values. To assess the quality of the approximate posteriors, we compute a calibration error (Radev et al., 2020) of the empirical coverage of each marginal posterior Finally, we implement simulation-based calibration (SBC, (Talts et al., 2018)) for visually detecting systematic biases in the approximate posteriors.

Point Estimates. Point estimates obtained by posterior means as well as corresponding RMSE and R^2 metrics are depicted in Figure 5.5A-B. Note, that point estimates do not have any special status in Bayesian inference, as they could be misleading depending on the shape of the posteriors. However, they are simple to interpret and useful for ease-of-comparison. We observe that pointwise recovery of λ is better than that of δ. This is mainly due to suboptimal pointwise recovery in the lower $(0, 0.1)$ range of δ. This pattern is evident in Figure 5.5A-B and is due to the fact that δ values in this range produce almost indistinguishable data patterns. Bootstrap estimates yielded an average RMSE of 0.155 ($SD = 0.004$) and an average R^2 of 0.708 ($SD = 0.015$) for the δ parameter. An average RMSE of 0.094 ($SD = 0.002$) and an average R^2 of 0.895 ($SD = 0.007$) were obtained for the λ parameter. These results suggest good global pointwise recovery but also warrant the inspection of full posteriors, especially in the low ranges of δ.

Full Posteriors. Average bootstrap calibration error was 0.011 ($SD = 0.005$) for the marginal posterior of δ and 0.014 ($SD = 0.007$) for the marginal posterior of λ. Calibration error is perhaps the most important metric here, as it measures potential under- or overconfidence across all confidence intervals of the approximate posterior (i.e., an α-confidence interval should contain the true posterior with a probability of α, for all $\alpha \in (0, 1)$). Thus, low calibration error indicates a faithful uncertainty representation of the approximate posteriors. Additionally, SBC-histograms are depicted in Figure 5.5C-D. As shown by (Talts et al., 2018), deviations from the uniformity of the rank statistic (also know as a PIT histogram) indicate systematic biases in the posterior estimates. A visual inspection of the histograms reveals that the posterior means

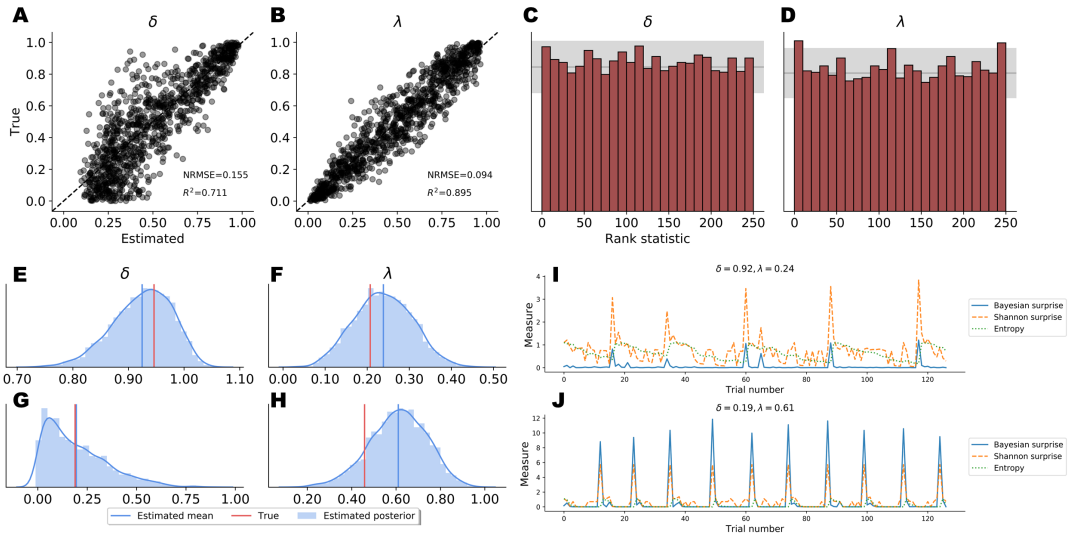


FIGURE 5.5: Parameter recovery results on validation data; (**A** and **B**) Posterior means vs. true parameter values; (**C** and **D**) Histograms of the rank statistic used for simulation-based calibration; (**E**-**H**) Example full posteriors for two validation data sets; (**I** and **J**) Example information-theoretic dynamics recovered from the parameter posteriors.

slightly overestimate the true values of δ. This corroborates the pattern seen in Figure 5.5A-B for the lower range of δ.

Finally, Figure 5.5E-H depicts the full marginal posteriors on two example validation sets. Even on these two data sets, we observe strikingly different posterior shapes. The marginal posterior of δ obtained from the first data set is slightly left-skewed and has its density concentrated over the $(0.8, 1.0)$ range. On the other hand, the marginal posterior of δ from the second data set is noticeably right-skewed and peaked across the lower range of the parameter. The marginal posteriors of λ appear more symmetric and warrant the use of the posterior mean as a useful summary of the distribution. These two examples underline the importance of investigating full posterior distributions as a means to encode epistemic uncertainty about parameter values. Moreover, they demonstrate the advantage of imposing no distributional assumptions on the resulting posteriors, as their form and sharpness can vary widely depending on the concrete data set.

5.4 Application

In this section we fit the Bayesian cognitive model to real clinical data. The aim of this application is to evaluate the ability of our computational framework to account for dysfunctional cognitive dynamics of information processing in substance dependent individuals (SDI) as compared to healthy controls.

5.4.1 Rationale

The advantage of modeling cognitive dynamics in individuals from a clinical population is that model predictions can be examined in light of available evidence about individual performance. For instance, SDIs are known to demonstrate inefficient conceptualization of the task and dysfunctional, error-prone response strategies. This has been attributed to defective error monitoring and behavior modulation systems, which depend on cingulate and frontal brain regions functionality (Kübler et al., 2005; Willuhn et al., 2003). On the other hand, the WCST should be a rather easy and straightforward task for healthy participants to obtain excellent performance. Therefore, we expect our model to consistently capture such characteristics. To test these expectations, we estimate the two relevant parameters λ and δ from both clinical patients and healthy controls from an already published dataset (Bechara and Damasio, 2002).

5.4.2 The Data

The dataset used in this application consists of responses collected by administering the standard Heaton version of the WCST (Heaton, 1981) to healthy participants and SDIs. In this version of the task, the sorting rule changes when a participant collects a series of 10 consecutive correct responses, and the task ends when this happens for 6 times. Participants in the study consisted of 39 SDIs and 49 healthy individuals. All participants were adults ($>$ 18 years old) and gave their informed consent for inclusion which was approved by the appropriate human subject committee at the University of Iowa. SDIs were diagnosed as substance dependent based on the Structured Clinical Interview for DSM-IV criteria (First, 1997).

5.4.3 Model Fitting

We fit the Bayesian cognitive agent separately to data from each participant in order to obtain individual-level posterior distributions. We apply the same BayesFlow network trained for the previous simulation studies, so obtaining posterior samples for each participant is almost instant (due to amortized inference).

5.4.4 Results

The means of the joint posterior distributions are depicted for each individual in Figure 5.6, and provide a complete overview of the heterogeneity in cognitive sub-components at both individual and group levels (individual-level full joint posterior distributions can be found in the Appendix B).

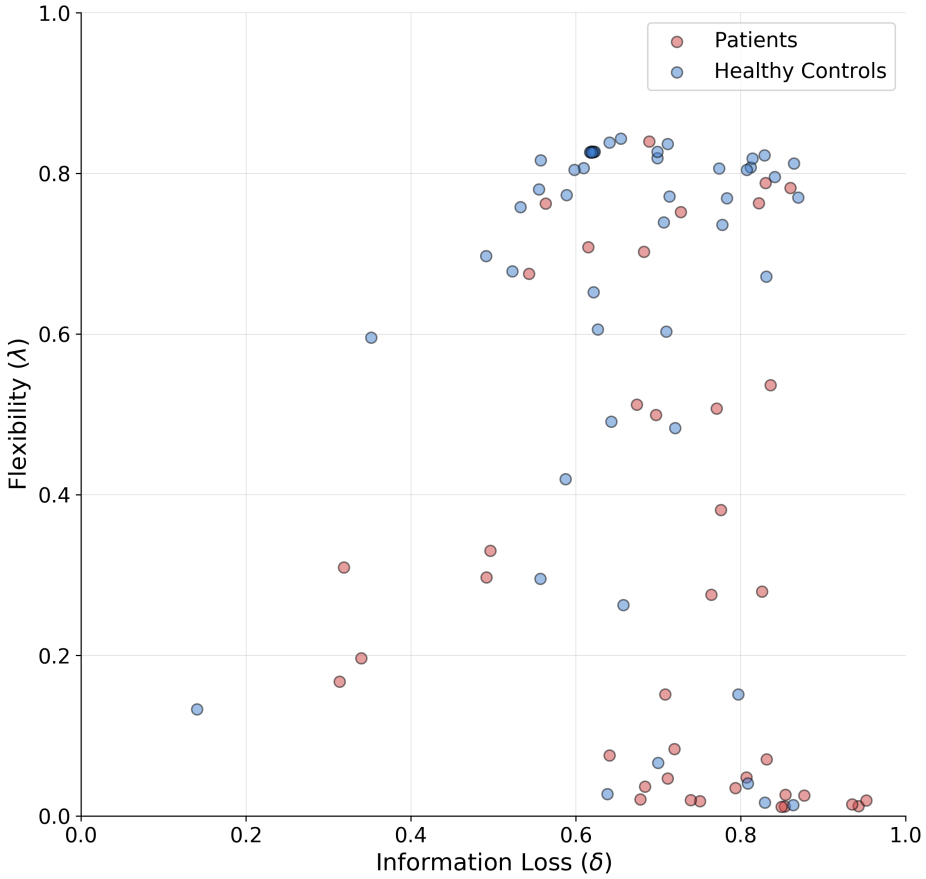


FIGURE 5.6: Joint posterior mean coordinates of the cognitive parameters, flexibility (λ) and information loss (δ), estimated for each individual. We observe a great heterogeneity in the distribution of posterior means, most pronouncedly for the flexibility parameter. However, a moderate between-subject variability in information loss can still be observed in both groups.

The estimates reveal a rather interesting pattern across both healthy and SDI participants. In particular, in both clinical and control groups, individuals with a poor flexibility (e.g., low values of λ) can be detected. However, the group parameter space appears to be partitioned into two main clusters consisting of individuals with high and low flexibility, respectively. As can be noticed, the majority of SDIs belongs to the latter cluster, which suggests that the model is able to capture error-related defective behavior in the clinical population and attribute it specifically to the flexibility parameter. On the other hand, individual performance seems hardly separable along the information loss parameter dimension.

As a further validation, we compare the classification performance of two logistic regression models. The first uses the estimated parameter means as

inputs and the participants' binary group assignment (patient vs. control) as an outcome. The second uses the four standard clinical measures (non-perseverative errors (E), perseverative errors (PE), number of trials to complete the first category (TFC), number of failures to maintain set (FMS) computed from the sample as inputs and the same outcome. Since we are interested solely in classification performance and want to mitigate potential overfitting due to small sample size, we compute leave-one-out cross-validated (LOO-CV) performance for both models. Interestingly, both logistic regression models achieve the same accuracy of 0.70, with a sensitivity of 0.71 and specificity of 0.70. Thus, it appears that our model is able to differentiate between SDIs and healthy individuals as good as the standard clinical measures.

However, as pointed out in the previous sections, estimated parameters serve merely as a basis to reconstruct cognitive dynamics by means of the trial-by-trial unfolding of information-theoretic measures. Moreover, cognitive dynamics can only be analysed and interpreted by relying on the joint contribution of both estimated parameters and individual-specific observed response patterns.

To further clarify this concept, we investigate the reconstructed time series of information-theoretic quantities based on the response patterns of two exemplary individuals (Figure 5.7). In particular, Figure 5.7A depicts the behavioral outcomes of a SDI with sub-optimal performance where the information-theoretic trajectories are reconstructed by taking the corresponding posterior means ($[\bar{\lambda} = 0.07, \bar{\delta} = 0.82]$), thus representing compromised flexibility and high information loss. Differently, Figure 5.7B shows the information-theoretic path related to response dynamics of an optimal control participant, according to the parameter set $[\bar{\lambda} = 0.60, \bar{\delta} = 0.35]$, representing relatively high flexibility, and low information loss. Note, that in both cases, the reconstructed information-theoretic measures are based on the estimated posterior means for ease of comparison (see Appendix B for the full joint posterior densities of the two exemplary individuals and the rest of the sample).

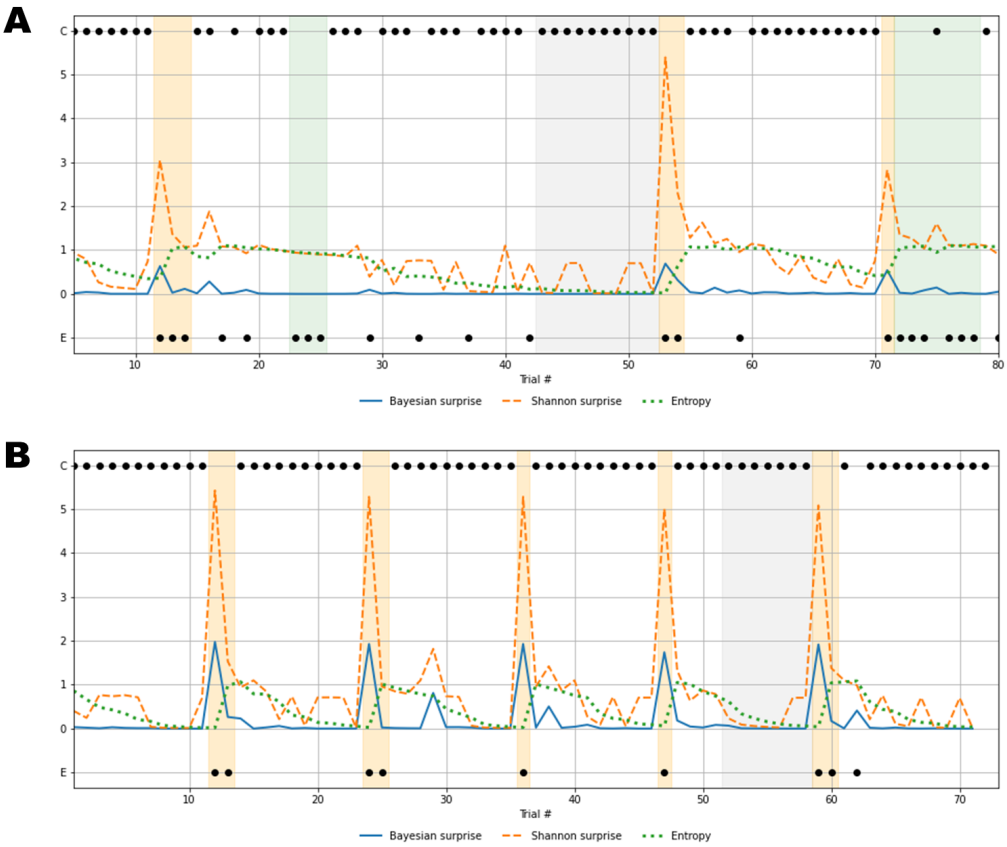


FIGURE 5.7: Recovered cognitive dynamics of two exemplary individuals. (**A**) Trial-by-trial information-theoretic measures of a SDI characterized by very low flexibility and very high information loss; (**B**) Trial-by-trial information-theoretic measures of a healthy individual characterized by relatively high flexibility and low information loss. Labels C and E indicate correct and error responses.

Results in Figure 5.7A account for a typical sub-optimal behavior observed in the SDI group, where several errors are produced in different phases of the task. The error patterns produced by such an individual might be induced by a non-trivial interaction between cognitive sub-components. Lower values of flexibility imply that errors are likely to be produced by generating responses from an internal environmental model which is no longer valid. In other words, the agent is unable to rely on local feedback-related information in order to update beliefs about hidden states. On the other hand, higher values of information loss reflect a general inefficiency of belief updating processes due to slow convergence to the optimal probabilistic environmental model. From this perspective, Bayesian surprise $\mathcal{B}_t$ and Shannon surprise $\mathcal{I}_t$ might play different roles in regulating behavior based on different internal model probability configurations. In addition, errors might be processed differently based on the status of the internal environmental representation, as reflected by the entropy of the predictive model, $\mathcal{H}_t$. Thus, information-theoretic measures allow to describe cognitive dynamics on a trial-by-trial basis and,

further, to disentangle the effect that different feedback-related information processing dynamics exert on adaptive behavior.

Processing unexpected observations is accounted by the quantification of surprise upon observing a response-feedback pair which is inconsistent with the current internal model of the task environment. Negative feedback is maximally informative when errors occur after the internal model has converged to the true task model (grey area, Figure 5.7A), or the entropy approaches zero (grey line, Figure 5.7A). The Shannon surprise (orange line) is maximal when errors occur within trial windows in which the agent's uncertainty about environmental states is minimal (orange areas, Figure 5.7A). However, internal model updates following an informative feedback are not optimally performed, which is reflected by very small Bayesian surprise (blue line, Figure 5.7A). This can be attributed to impaired flexibility and reflects the fact that after internal model convergence, informative feedback is not processed adequately and the internal model becomes impervious to change.

Conversely, errors occurring when the agent is uncertain about the true environmental state carry no useful information for belief updating, since the system fails to conceive such errors as unexpected and informative. The information loss parameter plays a crucial role in characterizing this cognitive behavior. The slow convergence to the true environmental model, accompanied by the slow reduction of entropy in the predictive model, leads to a large number of trials required to achieve a good representation of the current task environment (white areas, Figure 5.7A). Errors occurring within trial windows with large predictive model entropy (green area, Figure 5.7A) do not affect subsequent behavior, and feedback is maximally uninformative.

Rather different cognitive dynamics can be observed in Figure 5.7B, accounting for a typical optimal behavior where the errors produced fall within the trial windows which follow a rule completion (e.g. when the individual completes a sequence of 10 consecutive correct responses), and, thus, the environmental model becomes obsolete. However, the high flexibility, λ, allows to rely on local feedback-related information to suddenly update beliefs about the hidden states, that is, the most appropriate sorting rule. In this case, negative feedback become maximally informative after model convergence (grey area, Figure 5.7B) and the process of entropy reduction (green line, Figure 5.7B) is faster (e.g. less trials are needed) compared to the sub-optimal behavior scenario. Since uncertainty about the environmental states decreases faster, the Shannon surprise is always highly peaked when errors occur (orange line, Figure 5.7B), thus ensuring an efficient employment of the local feedback-related information. Accordingly, higher values of Bayesian surprise are observed (blue line, Figure 5.7B), revealing optimal internal model updating.

In general, the role that predictive (internal) model uncertainty plays in characterizing the way the agent processes feedback allows to disentangle sub-types of errors based on the information they convey for subsequent belief updating. From this perspective, error classification is entirely dependent on the status of the internal environmental model across task phases. Identifying

such a dynamic latent process is therefore fundamental, since the error codification criterion evolves with respect to the internal information processing dynamics. Otherwise, the problem of inferring which errors are due to perseverance in maintaining an older (converged) internal model and which due to uncertainty about the true environmental state becomes intractable, or even impossible.

5.5 Discussion

Investigating information processing related to changing environmental contingencies is fundamental to understanding adaptive behavior. For this purpose, cognitive scientists mostly rely on controlled settings in which individuals are asked to accomplish (possibly) highly demanding tasks whose demands are assumed to resemble those of natural environments. Even in the most trivial cases, such as the WCST, optimal performance requires integrated and distributed neurocognitive processes. Moreover, these processes are unlikely to be isolated by simple scoring or aggregate performance measures.

In the current work, we developed and validated a new computational Bayesian model which maps distinct cognitive processes into separable information-theoretic constructs underlying observed adaptive behavior. We argue that these constructs could help describe and investigate the neurocognitive processes underlying adaptive behavior in a principled way.

Furthermore, we couple our computational model with a novel neural density estimation method for simulation-based Bayesian inference (Radev et al., 2020). Accordingly, we can quantify the entire information contained in the data about the assumed cognitive parameters via a full joint posterior over plausible parameter values. Based on the joint posterior, a representative summary statistic can be computed to simulate the most plausible unfolding of information-theoretic quantities on a trial-by-trial basis.

Several computational models have been proposed to describe and explain performance in the WCST, ranging from behavioral (Bishara et al., 2010; Gläscher et al., 2019; Steinke et al., 2020) to neural network models (Dehaene and Changeux, 1991; Amos, 2000; Levine et al., 1993; Monchi et al., 2000). These models aim to provide psychologically interpretable parameters or biologically inspired network structures, respectively, accounting for specific qualitative patterns of observed data. Behavioral models, in particular, abstract the main cognitive features underlying individual performance in the WCST according to different theoretical frameworks (e.g., attentional updating (Bishara et al., 2010), or reinforcement learning (Steinke et al., 2020)) and disentangle psychological sub-processes explaining observed task performance. However, the main advantage of our Bayesian model is that it provides both a cognitive and a measurement model which coexist within the overarching theoretical framework of Bayesian brain theories. More precisely, the presented model is specifically designed to capture trial-by-trial fluctuations in information processing as described by second-order information-theoretic quantities. The latter can be seen as a multivariate quantitative account of the interaction between the agent and its environment. Moreover, it is worth

noting that such a model representation might not be applicable outside a Bayesian theoretical framework.

Even though our computational model is not a neural model, it might provide a suitable description of cognitive dynamics at a representational and/or a computational level (Marr, 1982). This description can then be related to neural functioning underlying adaptive behavioral. Indeed, there is some evidence to suggest that neural processes related to belief maintenance/updating and unexpectedness are crucial for performance in the WCST. In particular, brain circuits associated with cognitive control and belief formation, such as the parietal cortex and prefrontal regions, seem to share a functional basis with neural substrates involved in adaptive tasks (Nour et al., 2018). Prefrontal regions appear to mediate the relation between feedback and belief updating (Lie et al., 2006) and efficient functioning in such brain structures seems to be heavily dependent on dopaminergic neuromodulation (Ott and Nieder, 2019). Moreover, the dopaminergic system plays a role in the processing of salient and unexpected environmental stimuli, in learning based on error-related information, and in evaluating candidate actions (Nour et al., 2018; Daw et al., 2011; Gershman, 2018). Accordingly, dopaminergic system functioning has been put in relation with performance in the WCST (Hsieh et al., 2010; Rybakowski et al., 2005) and shown to be critical for the main executive components involved in the task, that is, cognitive flexibility and set-shifting (Bestmann et al., 2014; Stelzel et al., 2010). Further, neural activity in the anterior cingulate cortex (ACC) is increased when a negative feedback occurs in the context of the WCST (Lie et al., 2006). This finding corroborates the view that the ACC is part of an error-detection network which allocates attentional resources to prevent future errors. The ACC might play a crucial role in adaptive functioning by encoding error-related or, more generally, feedback-related information. Thus, it could facilitate the updating of internal environmental models (Rushworth and Behrens, 2008).

The neurobiological evidence suggests that brain networks involved in the WCST might endow adaptive behavior by accounting for maintaining/updating of an internal model of the environment and efficient processing of unexpected information. Is it noteworthy, that these processing aspects are incorporated into our computational framework. At this point, we briefly outline the empirical and theoretical potentials of the proposed computational framework for investigating adaptive functioning and discuss future research vistas.

Model-Based Neuroscience. Recent studies have pointed out the advantage of simultaneously modeling and analyzing neural and behavioral data within a joint modeling framework. In this way, the latter can be used to provide information for the former, as well as the other way around (Turner et al., 2017; Turner et al., 2013; Forstmann et al., 2011). This involves the development of joint models which encode assumptions about the probabilistic relationships between neural and cognitive parameters.

Within our framework, the reconstruction of information-theoretic discrete time series yields a quantitative account of the agent's internal processing of environmental information. Event-related cognitive measures of belief

updating, epistemic uncertainty and surprise can be put in relation with neural measurements by explicitly providing a formal account of the statistical dependencies between neural and cognitive (information-theoretic) quantities. In this way, latent cognitive dynamics can be directly related to neural event-related measures (e.g., fMRI, EEG). Applications in which information-theoretic measures are treated as dependent variables in standard statistical analysis are also possible.

Neurological Assessment. Although neuroscientists have considered performance in the WCST as a proxy for measuring high-level cognitive processes, the usual approach to the analysis of human adaptive behavior consists in summarizing response patterns by simple heuristic scoring measures (e.g., occurrences of correct responses and sub-types of errors produced) and classification rules (Flashman et al., 1991). However, the theoretical utility of such a summary approach remains questionable. Indeed, adaptive behavior appears to depend on a complex and intricate interplay between multiple network structures (Barcelo et al., 2006; Monchi et al., 2001; Lie et al., 2006; Barceló and Rubia, 1998; Buchsbaum et al., 2005). This posits a great challenge for disentangling high-level cognitive constructs at a model level and further investigating their relationship with neurobiological substrates. It appears that standard scoring measures might not be able to fulfil these tasks. Moreover, there is a pronounced lack of anatomical specificity in previous research concerning the neural and functional substrates of the WCST (Nyhus and Barceló, 2009).

Thus, there is a need for more sophisticated modeling approaches. For instance, disentangling errors due to perseverative processing of previously relevant environmental models from those due to uncertainty about task environmental states, is important and nontrivial. Sparse and distributed error patterns might depend on several internal model probability configurations. Such internal models are latent, and can only be uncovered through cognitive modeling. Therefore, information-based criteria to response (error) classification can enrich clinical evaluation beyond heuristically motivated criteria.

Generalizability. Another important advantage of the proposed computational framework is that it is not solely confined to the WCST. In fact, one can argue that the seventy-year old WCST does not provide the only or even the most suitable setting for extracting information about cognitive dynamics from general populations or maladaptive behavior in clinical populations. One can envision tasks which embody probabilistic (uncertain) or even chaotic environments (for instance with partially observable or unreliable feedback or partially observable states) and demand integrating information from different modalities (O'Reilly et al., 2013; Nour et al., 2018). These settings might prove more suitable for investigating changes in uncertainty-related processing or cross-modal integration than deterministic and fully observable WCST-like settings.

Despite these advantages, our proposed computational framework has certain limitations. A first limitation might concern the fact that the new Bayesian cognitive model accounts for the main dynamics in adaptive tasks

by relying on only two parameters. Although such a parsimonious proposal suffices to disentangle latent data-generating processes, a more exhaustive formal description of cognitive sub-components might be envisioned. However, parameter estimation can become challenging in such a scenario, especially when one-dimensional response data is used as a basis for parameter recovery. Second, the information loss parameter appears to be more challenging to estimate than the flexibility parameter in some datasets. There are at least two possible remedies for this problem. On the one hand, global estimation of information loss might be hampered due to the model's current functional (algorithmic) formulation and can therefore be optimized via an alternative formulation/parameterization. On the other hand, it might be the case that the data obtainable in the simple WCST environment is not particularly informative about this parameter and, in general, not suitable for modeling more complex and non-linear cognitive dynamics in general. Future works should therefore focus on designing and exploring more data-rich controlled environments which can provide a better starting point for investigating complex latent cognitive dynamics in a principled way. Additionally, the information loss parameter seems to be less effective in differentiating between substance abusers and healthy controls in the particular sample used in this work. Thus, further model-based analyses on individuals from different clinical populations are needed to fully understand the potential of our 2-parameter model as a clinical neuropsychological tool. Finally, in this work, we did not perform formal model comparison, as this would require an extensive consideration of various nested and non-nested model within the same theoretical framework and between different theoretical frameworks. We therefore leave this important endeavor for future research.

5.6 Conclusions

In conclusion, the proposed model can be considered as the basis for a (bio)psychometric tool for measuring the dynamics of cognitive processes under changing environmental demands. Furthermore, it can be seen as a step towards a theory-based framework for investigating the relation between such cognitive measures and their neural underpinnings. Further investigations are needed to refine the proposed computational model and systematically explore the advantages of the Bayesian brain theoretical framework for empirical research on high-level cognition.

Chapter 6

A Probabilistic Graphical Model to jointly analyse structural neural and behavioural data in a risky task

The content of the chapter has been in part published as: D'Alessandro, M., Gallitto, G., Greco, A., & Lombardi, L. (2020). A joint modelling approach to analyze risky decisions by means of diffusion tensor imaging and behavioural data. *Brain Sciences*, 10, 138.

6.1 Introduction

In cognitive neuroscience, relations between neural and behavioural characteristics of individuals are usually analyzed using a two-step approach which first summarizes performances on a given experimental task, and then applies standard statistical analysis on the neural and behavioural measures. However, several studies have highlighted the limitations of this approach in investigating and selecting theories to explain the relation between neural functioning and cognition (Turner et al., 2013; Hawkins et al., 2017; Bridwell et al., 2018).

Advances in the understanding of this relation are due to the development of different computational tools, allowing for a finer analysis of several sources of information. Some examples are: (1) cognitive modelling (Lee and Wagenmakers, 2014; Lewandowsky and Farrell, 2010) which formally accounts for the generative cognitive processes which are assumed to produce the observed data; (2) Bayesian graphical models (Lee, 2011a; Barber, 2012) which provide a powerful and flexible way to perform hierarchical Bayesian analysis, allowing to account for group and individual differences; (3) joint neurocognitive modelling (Forstmann et al., 2011; Nunez et al., 2017; Turner et al., 2013; **?**; Palestro et al., 2018) which provides a framework to simultaneously model and analyze neural and behavioural data by allowing the latter to be informative for the former, and vice versa.

The latter modelling framework has demonstrated to be an effective way to increase knowledge about the underlying neural substrates of cognitive functioning by bridging the gap between neuroscience and mathematical

psychology. Here, the main advantage consists of using formal cognitive models as tools to isolate and quantify cognitive processes in order to effectively associate them with some brain measurements (Forstmann et al., 2011)).

In this work we aimed to put the emphasis on the mutual dependency between measures of structural integrity of brain regions of interest and cognitive functioning as assessed by the analysis of the outcomes of a given experimental task.

Several works ranging from perception (Brouwer and Heeger, 2011), attention (Lu et al., 2011), memory (Kragel et al., 2015), categorization (Mack et al., 2013), and decision in two alternatives forced choices (Turner et al., 2013; van Ravenzwaaij et al., 2017) have demonstrated the need to formally account for reciprocal relations between mathematical behavioural models and brain functional or structural data.

In this contribution we proposed an architecture for jointly modelling such reciprocal relation in the context of risky decision-making. Although risk–decision tasks can be considered highly popular and effective experimental tools to investigate cognitive control and decision-making characteristics under risk conditions, a model-based approach to the joint analysis of brain and behavioural data in such contexts is still lacking.

Here, we proposed a novel way to relate structural information from Diffusion Tensor Imaging (DTI) to psychological parameters of a computational cognitive model accounting for the behavioural outcomes in the Balloon Analogue Risk Task (BART; (Lejuez et al., 2002)), from a confirmatory perspective.

The BART represents an ideal scenario to model decision-making since it has been correlated to "real-world" risk taking (Aklin et al., 2005; Lejuez et al., 2002). The task has proven to reliably account for risk-taking propensity, response strategy and risk-related behaviour modulation in a broad range of normal and clinical populations (Goldenberg et al., 2017; Cazzell et al., 2012; Bornovalova et al., 2005; Lejuez et al., 2003). In a typical BART setting, participants are required to decide whether to risk by inflating a balloon to earn a cumulative small monetary reward, being informed that the balloon might explode with a certain probability, thus causing the loss of accrued earnings. If participants decide to stop inflating they can cash out the current winnings. Optimizing total earnings in such a scenario is not trivial. In general, it requires a balanced risky-oriented strategy, learning from experience and modulating choices consistently (van Ravenzwaaij et al., 2011).

In the present work we adopted a hierarchical Bayesian framework to relate neural and cognitive parameters inferred from performances of healthy participants on the BART. Analysis of posterior distributions was then used to assess relationships between the neural and cognitive variables. The model was applied to data from an already published dataset. Finally, the potentials in applying the method to the analysis of neural substrates underlying risk-taking behaviour and decision making were outlined and discussed.

6.2 Materials and Methods

6.2.1 The BART Data

The dataset used in this work was selected from the OpenfMRI database repository (http://www.openfmri.org; (Poldrack et al., 2013)) and refers to the experimental data reported in (Cohen and Poldrack, 2014). The dataset contains both behavioural performances and MRI scans from 24 healthy participants on a slightly modified version of the BART. Participants were adults recruited from UCLA's campus with ages in the range 18–33, with no history of neurological illness and no use of psychoactive medication or illegal substances.

In the adopted version of the task, individuals saw a balloon on the monitor and were asked to select one of two possible options at each choice occasion for a given trial. The first option consisted in inflating the balloon, and is referred to as pump . The second option ended up the current trial by deciding to stop inflating the balloon, and is referred to as cash. Pumping the balloon increased the amount of possible monetary reward by 25 cents for each pump. If the participant decided to stop inflating the balloon, the accrued money was moved to a permanent store of winnings and a new balloon was presented. After a variable number of pumps the balloon exploded, in which case the participant lost all the money in the temporary pool. Participants did not receive any cue about the bursting probability. However, probabilities of explosions were not fixed and the actual number of pumps before an explosion followed a uniform distribution across trials, with an average of 6 pumps ($SD = 2$ pumps). Each balloon was presented on each trial for a total of 36 trials.

6.2.2 The Cognitive Model

As previously outlined, performances of the BART are employed as a measure of risk-related behavioural tendencies and are usually analyzed by means of standard summary measures on test outcomes (e.g., total number of pumps, frequency of pumps across trials, number of cashes, number of explosions). However, such measures do not provide a suitable account for the data-generating process, that is, for the cognitive sub-processes involved in the task.

In this work we proposed a parsimonious computational account of the cognitive mechanisms underlying the observed response pattern of pumps and cashes (Wallsten et al., 2005; van Ravenzwaaij et al., 2011). In particular, we relied on a modified version of a robust model representation which has shown to be particularly stable to parameter recovery and estimation (van Ravenzwaaij et al., 2011).

The model assumes a subjective probability estimate that a pump will make the balloon bursts in a given trial k. It also assumes that individuals determine the number of pumps for that trial prior to the first actions, and do not make adjustments during pumping. The number of pumps that individuals consider optimal on trial k is defined as ω_k, and depends on the propensity

of risk taking, γ, and on the current subjective bursting probability p_k^*, as follows:

$$\omega_k = -\frac{\gamma}{\log(1-p_k^*)} \tag{6.1}$$

where $\gamma \geq 0$. Equation (6.1) provides a parsimonious and effective representation of an individual decision strategy. Intuitively, ω_k places an upper bound on the pump attainable at a given trial, which is proportional to risk propensity, γ. The term p_k^* in the denominator has the role of shrinking the number of pumps an individual considers as optimal. Moreover, the probability of pumping in trial k, at a given occasion j, is defined as θ_{kj} and depends on ω_k and on behavioural consistency, β, which can be though to account for response variability:

$$\theta_{kj} = [1+\exp(\beta(j-\omega_k)]^{-1} \tag{6.2}$$

where $\beta \geq 0$. High values (resp. low values) of β mean less variable responding (resp. more variable responding). Equation (6.2) represents the fact that behaviour is generally determined by the divergence between the current choice occasion j (e.g., pump opportunity) and the optimal number of pumps, ω_k. When the optimal number of pumps is exceeded ($j > \omega_k$ for the trial k), the probability of pumping, θ_{kj}, approaches zero. However, parameter β reflects the degree to which a response is determined by such a divergence. When $\beta = 0$, the individual decision to pump or cash is random. Differently, decisions become more consistently determined by the divergence criterion as β increases.

However, the original formulation of the model (van Ravenzwaaij et al., 2011) assumed parameter p_k^* to be fixed (which implies removing subscript k) and known from participants at the beginning of the task. This supports the assumption that the subjective probability of burst is constant across the task trials. However, fitting such model to our data could be problematic at least for two main reasons: (1) participants were not informed about the true bursting probability which in the task is uniformly distributed across trials; (2) in general, it is not possible to ensure that subjective bursting probabilities are consistent among participants and constant across trials, whatever the information they receive prior to the task.

In our model representation, subjective bursting probability and its dynamics were taken into account and inferred by relying on the history of participant's choices. Let the variable C_k indicate the cumulative success rate up to trial k according to:

$$C_k = \frac{\sum_{t=1}^{k-1} s_t}{\sum_{t=1}^{k-1} n_t}$$

where s_t and n_t are the number of successful (non-bursting) balloon pumps

and total pumping attempts at trial t, respectively. Modelling C_k as a Beta distributed random variable yields the statistical solution to the task of inferring the subjective bursting probability as follows:

$$\begin{aligned} C_k &\sim \text{Beta}(\mu_{\alpha_k}\sigma_\alpha, (1-\mu_{\alpha_k})\sigma_\alpha) \\ \mu_{\alpha_k} &= \text{logit}^{-1}(\alpha_0 + \alpha_1 k) \quad (6.3) \\ p_k^* &= 1 - \mu_{\alpha_k} \quad (6.4) \end{aligned}$$

where the cumulative success rate is regressed on trial numbers, and parameters α_0 and α_1, denoting the intercept and slope respectively, are the regression coefficients. For computational convenience we adopted the parameterization proposed by Ferrari and Cribari-Neto (Ferrari and Cribari-Neto, 2004). Such a parameterization has proven to be convenient in our computational setting since it allowed to model the observed cumulative success rate at trial k as sampled from a Beta distribution with expected value μ_{α_k} and concentration σ_α. Thus, the (conditional) expected value of the Beta distribution, given the specific trial, has been modelled as a function of cognitive parameters α_0 and α_1, and the specific trial k. The inverse logit function allowed to map such parameters to the natural domain of the expected value of the Beta distribution according to the specified parameterization. Here, α_0 indicates the baseline subjective bursting probability and α_1 represents the rate of change of bursting belief. In particular, if $\alpha_1 < 0$ (rep. $\alpha_1 > 0$), then this reflects an indicator that perceived bursting probability increases (resp. decreases) as the participant's responses start accumulating balloon bursts as the trials unfold (resp., start decreasing balloon bursts).

At this point, the resulting cognitive model can be thought to account for response configurations of pumps and cashes by means of two hierarchically organized sub-models. In the first sub-model, a trial-specific bursting probability, p_k^*, is computed based on the baseline subjective bursting probability, α_0, and the bursting belief dynamic yielded by α_1. In the second sub-model, the decision process is instantiated by allowing the system to estimate an optimal number of pumps, ω_k, conditioned on the computed trial-specific bursting probability, p_k^*, and a response is delivered based on θ_{kj}.

Therefore, model representation allows to test the hypothesis that participants do not modify their initial bursting belief during the task. When $\alpha_1 = 0$, behaviour depends only on the baseline bursting probability and cognitive parameters γ, and β.

From a generative perspective, an observed pumping action y_{kj} (1 if pump, 0 if cash) can be modelled according to a Bernoulli distribution:

$$y_{kj} \sim \text{Bernoulli}(\theta_{kj})$$

and θ_{kj} depends on both cognitive parameters and the specific choice occasion within a specific trial. The likelihood function is then defined as follows:

$$p(\boldsymbol{Y}|\boldsymbol{\Omega}) = \prod_{k=1}^{K} \prod_{j=1}^{J(k)} \theta_{kj}^{y_{kj}} (1 - \theta_{kj})^{(1-y_{kj})} \tag{6.5}$$

where $\boldsymbol{\Omega} = (\gamma, \beta, \alpha_0, \alpha_1)$ is the array of parameters of the behavioural model, and $J(k)$ is the total number of observed actions for trial k.

6.2.3 The Neural Model

The cognitive model decomposition allowed to isolate individual cognitive characteristics and to rephrase them in terms of model parameters. A further step to model neural and behavioural data simultaneously consisted in bringing individual brain characteristics into the joint model. To this purpose, we focused on neural structural information at individual level. More precisely, we wanted the neural model to account for properties of structural connectivity in the brain. Consistently, we adopted Fractional Anisotropy (FA) as the founding measure to parameterize individual brain structural connectivity.

FA is the most commonly used index for estimation of anisotropy using DTI, and reflects fiber tracts characteristics such as the extent of alignment of cellular structures within the fibers and their structural integrity (Pierpaoli and Basser, 1996; Beppu et al., 2003). Therefore, such a measure also proved to be a promising index to study the relation between brain structural integrity and both response variability and risky behaviour in both clinical and normal population (Kwon et al., 2014; Lane et al., 2010; Goldenberg et al., 2017; Kohno et al., 2017).

In this work, we were interested in relating cognitive functioning with connectivity measures of networks of regions of interest (ROIs). This choice was motivated by substantial evidences on the potential of functional-structural properties of distributed neural networks to account for complex decision processes (Fukunaga et al., 2012; Kohno et al., 2017; Krain et al., 2006). Thus, the main purpose of our neurocognitive modelling approach consisted of linking connectivity-related information of network structures with the latent mechanisms captured by the computational cognitive model.

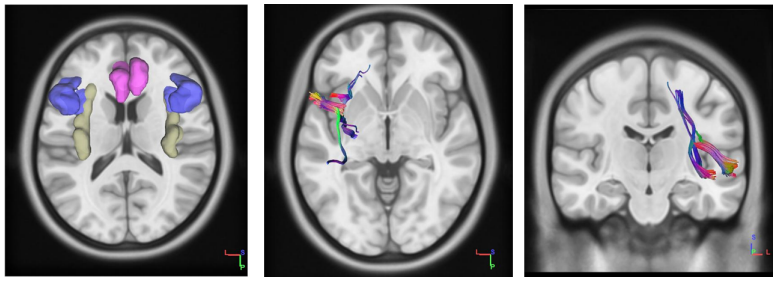
As a measure of connectivity we quantified the FA of white matter tracts relating specific regions of interest in a specific neural network. The confirmatory aspect of our approach was reflected by the choice of relying on a subset of the whole-brain structural connectivity matrix. To do this, a custom connectivity matrix was obtained by focusing on the following ROIs array and related white matter connectivity paths: left and right thalamus, striatum, dorsolateral prefrontal cortex, anterior cingulate cortex, inferior frontal gyrus, insular cortex. A network was then defined as the vector of elements of any (biologically) consistent subset of the ROIs array.

More formally, consider a square and symmetric connectivity matrix $\boldsymbol{F}$, such that:

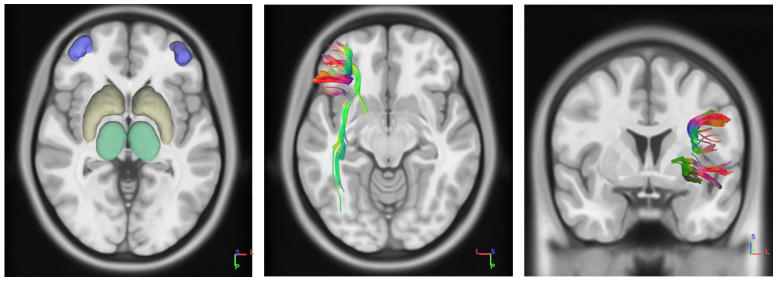
$$\boldsymbol{F} = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1J} \\ f_{21} & f_{22} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ f_{I1} & f_{I2} & \cdots & f_{IJ} \end{bmatrix}$$

where $f_{ij} = 0$ for $i = j$, and $I = J$. The entries f_{ij} specify the fractional anisotropy of the custom tract connecting ROIs i and j. We refer to Network FA (Kohno et al., 2017) to represent structural connectivity in a given network. Thus, network FA consists of the collection of $f_{ij}^{(x)}$ such that $i, j \in \boldsymbol{x}$, and $\boldsymbol{x}$ is the indicator variable reflecting the vector of ROI labels which constitute a defined network. Potentially, Network FA can be obtained for several combinations of ROI labels, and thus for several subsets of $\boldsymbol{F}$.

However, in this application we focused on two brain networks. The first network involves white matters connections between anterior cingulate cortex (ACC), insula, and inferior frontal gyrus (IFG), regions which are thought to be involved in loss-aversion modulations (Fukunaga et al., 2012). In particular, ACC is critically involved in cognitive control and decision-making processes in signaling anticipated risk and potential loss (Krawitz et al., 2010), whilst insula and IFG are thought to be implied in risk aversion signaling and risk avoidance during risky decisions (Christopoulos et al., 2009; **?**). The network implies fiber connections between ACC and IFG, and those implied by insular–cingulate and insular–frontal projections, bilaterally (Figure 6.1a). The second network involves projections between dorsolateral prefrontal cortex (dlPFC) and both striatum and thalamic nuclei, involved in top-down modulation of goal-directed behaviour (Furman et al., 2011) and, in part, in response variability in risky task (Goldenberg et al., 2017). Such a network is part of the Cortico–Striatal–Thalamic path, and consists of fiber connections between striatum and thalamus, and those implied by the striatal–dlPFC and thalamic–dlPFC projections, bilaterally (Figure 6.1b).



(A)



(B)

FIGURE 6.1: Pictures on the left show the regions of interest (ROIs) which constitute the networks. The network containing anterior cingulate (red), insula (yellow) and inferior frontal gyrus (blue) consists of the anterior cingulate cortex (ACC)–Insula–inferior frontal gyrus (IFG) Network (**a**). The network containing thalamus (green), striatum (yellow) and dorsolateral prefrontal cortex (dlPFC) (blue) consists of the dlPFC–Thalamus–Striatum Network (**b**). The central and rightmost pictures represent tracts of white matters connections for the first and the second network, respectively. For simplicity, figures show networks tracts for the left brain hemisphere, but the same applies to the opposite hemisphere. Network Fractional Anisotropy (FA) is intended to account for bilateral network tracts fractional anisotropy.

We refer to $\delta_{x=1}$ and $\delta_{x=2}$ as the neural parameters accounting for network FA measure for ACC–Insula–IFG and dlPFC–Thalamus–Striatum networks, respectively. Here, we followed a strategy proposed in (Turner et al., 2013) to easily provide a probabilistic account of the neural measures. In particular, tracts fractional anisotropy was assumed to be drawn from a Gaussian Distribution, which provided a computationally convenient and tractable probability model of Network FA:

$$\text{logit}(f_{ij}^{(x)}) \sim \text{Normal}(\delta_x, \sigma_x). \tag{6.6}$$

Here, δ_x is thought to be the latent neural parameter accounting for structural property of network x, and $f_{ij}^{(x)}$ is the fractional anisotropy for tract connecting ROIs i and j in network x. Parameter σ_x was thought to represent the inter-tracts variability of FA in the network x, and it was not conceived as

accounting for Network FA since we were not interested in relating the variance in the inter-tracts FA measurements to cognitive parameters in the joint framework. However, it is worth noticing that such assumption might be infeasible when high inter-tracts variation is empirically detected, as in the case of measurements on the clinical population. In this case, more consistent parametric models might be considered to better account for Network FA.

The likelihood function is then defined as follows:

$$p(\boldsymbol{f}^{(x)}|\delta_x, \sigma_x) = \prod_{n=1}^{N^{(x)}} \mathrm{N}(\mathrm{logit}(f_n^{(x)})|\delta_x, \sigma_x) \tag{6.7}$$

where $\mathrm{N}(\cdot)$ denotes the normal density with mean δ_x and standard deviation σ_x, and $N^{(x)}$ the number of connected tracts within the network x. We referred to $f_n^{(x)}$ as a simplified notation which reflects the FA value of the n-th tract connection between ROIs i and j in the network x.

Thus, neural parameters were inferred based on processed neural data and served to feed the system to account for the neural counterpart of the joint neurocognitive model, as will become clearer later in the next sections.

6.2.4 DTI Data Processing

The FA value computation was based on the eigendecomposition of the diffusion tensor (Basser and Pierpaoli, 1996). In order to extract tracts' FA, DTI diffusion images with a total of 64 volumes (diffusion sampling directions) with a b-value of $1000 s/\mathrm{mm}^2$, in-plane resolution of 1.97917 mm and slice thickness of 2 mm, were used for analysis. All images have been corrected for eddy currents with FSL's eddy toolbox using one $b0$ image as structural reference to account for geometrical distortions. The diffusion data were normalized in the MNI (Montreal Neurological Institute) space using affine registration and the ICBM-152 template, and a deterministic fiber tracking algorithm (Yeh et al., 2013) was used. The tractography and connectivity matrix were calculated using DSI Studio (`http://dsi-studio.labsolver.org`). A seeding region was placed at whole brain and the fiber tracking procedure was performed with the thresholds of minimum FA value at 0.15, and maximum angle at 27° according to previously utilized protocols (Christidi et al., 2016). The step size was randomly selected from 0.5 voxel to 1.5 voxels and tracks with length shorter than 30 or longer than 300 mm were discarded. A custom template with 12 ROIs (consisting of the brain regions whose tracts constitute $\boldsymbol{F}$), six left and six right, was created using AAL2 (Rolls et al., 2015), Desikan-Killiany-Tourville (Desikan et al., 2006) and HCP842 (Yeh et al., 2017) atlases and used as the brain parcellation. The connectivity matrix was calculated by using the FA of the connecting tracks.

6.2.5 Joint Modelling

A fundamental characteristic of joint models relies on their particular flexibility in allowing several assumptions about probabilistic (or deterministic) relations

between neural and behavioural variables to be taken into account through model's architecture.

In neurocognitive modelling are proposed two relevant architectures to account to the modelling of relationships between different sources of data: the *Directed Approach* and the *Covariance Approach* (Palestro et al., 2018; Turner et al., 2017).

In the directed approach, a statistical model of neural data is defined and it is assumed that behavioural model parameters are directly affected by neural model parameters, codifying a non-reciprocal relation between the two sources of information. By contrast, the covariance approach does not assume such restrictions on parameters dependencies, but relies on specifying a joint model in which cognitive and neural parameters share a multivariate structure with covariance.

In this work, we adopted the latter as an adapted version of the joint model proposed by (Turner et al., 2013). The primary reason for relying on the covariance approach was that we wanted to be agnostic in specifying the causal role of each source of information, that is, the directional statistical influence between neural and cognitive measures. To say it differently, the proposed covariance model combined both behavioural and neural models' parameters in a unified framework, which characterizes the way behavioural and neural parameters coexist to explain the underlying cognitive process (Turner et al., 2017).

In our context, the covariance model has been thought to account for individual differences in task performances and brain structural characteristics by letting individual-level parameters to be modelled by a multivariate distribution connecting the two sources of information. Such connection allowed the information yielded by the neural data, as represented by $\boldsymbol{F}$, to affect the information we learned about key cognitive parameters (e.g., γ, β).

We proposed a Multivariate Student's t-distribution (Welsh and Richardson, 1997; Pinheiro et al., 2001) as the multivariate probability model in order to account for robust relations between neural and cognitive parameters. Such relations were learned through hierarchical modelling and accounted by the (hyper-)covariance matrix of the multivariate distribution. Both cognitive and neural parameters were treated as latent variables.

For the behavioural model, we assumed structured individual differences in parameters γ, β, α_0, and α_1. The assumed model was also the one showing the best general fitting performances when compared to other possible models.

More precisely, we consider 4 possible behavioural models: (1) a model assuming structured individual differences in three parameters, namely, γ, β and p, where p^* is a fixed bursting belief. Here, parameter p^* assumes a fixed and known bursting probability by removing the dependency on the specific trial k; (2) a model assuming structured individual differences in three parameters, namely, γ, β and α_0. The latter indicates the baseline bursting probability. Parameter α_0 is inferred based on the beta regression parameterization outlined in the main text; (3) a model assuming structured individual differences in four parameters, namely, γ, β and α_0, and a group-level parameter α_1 accounting for the dynamics of the subjective bursting

belief; (4) the last model is the four parameter model used in the main text, which assumes structured individual differences for all the four parameters.

We assess models' performances by relying on both the *Deviance Information Criterion* (DIC) and the (mean) $\widehat{\mathrm{R}}$ statistic (Gelman et al., 1992) for each parameter across all individual estimates. Results are shown in the 6.1.

TABLE 6.1: Behavioural model comparison

Model	DIC	Parameters	$\hat{R}$ (mean)
1	2071.3	γ	2.368
		β	1.001
		p^*	2.551
2	-1497.6	γ	1.001
		β	1.001
		α_0	1.001
3	-1531.6	γ	1.001
		β	1.001
		α_0	1.001
		α_1	1.001
4	-1595.6	γ	1.001
		β	1.002
		α_0	1.023
		α_1	1.005

As can be noticed, Model 4 has the lowest DIC, and $\widehat{\mathrm{R}}$ approaching 1, and can be selected as the best model.

We further put a constraint on the relation between neural and cognitive parameters accounted by the covariance matrix of the multivariate probability model. In particular, we assumed that individual-level baseline bursting probability, α_0 and its updating, α_1, condition the behavioural model outside the covariance structure, and that risk-taking, γ and response variability, β, were then recovered within the multivariate probability model. To say it differently, we let α_0 and α_1 play the role of providing conditions for (possible) unbiased estimates of individual-level parameters γ and β, given that subjective probabilities have been taken into account.

The Multivariate Student's t-distribution was then specified by the hyper-parameters vector:

$$\boldsymbol{\mu} = (\mu_\gamma, \mu_\beta, \mu_{\delta_1}, \mu_{\delta_2})$$

containing the hyper-mean parameters for each of the individual-level parameters sharing a covariance matrix. The hyper-covariance matrix is thought to reflect research question and model assumptions. In our case we aimed

to investigate the relation between pairs of network FA and cognitive sub-processes in a confirmatory perspective, and it was defined as follows:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{\gamma}^2 & 0 & \sigma_{\gamma}\sigma_{\delta_1}\rho_1 & \sigma_{\gamma}\sigma_{\delta_2}\rho_2 \\ 0 & \sigma_{\beta}^2 & \sigma_{\beta}\sigma_{\delta_1}\rho_3 & \sigma_{\beta}\sigma_{\delta_2}\rho_4 \\ \sigma_{\gamma}\sigma_{\delta_1}\rho_1 & \sigma_{\beta}\sigma_{\delta_1}\rho_3 & \sigma_{\delta_1}^2 & 0 \\ \sigma_{\gamma}\sigma_{\delta_2}\rho_2 & \sigma_{\beta}\sigma_{\delta_2}\rho_4 & 0 & \sigma_{\delta_2}^2 \end{bmatrix}$$

where ρ_1 and ρ_2 account for the relation between risk-taking, γ, and both ACC–Insula–IFG and dlPFC–Thalamus–Striatum networks FA, δ_1 and δ_2, respectively. Correlation parameters ρ_3 and ρ_4 account for the relation between behavioural consistency, β, and both δ_1 and δ_2, respectively. Eventual relations between brain structural and cognitive characteristics were, thus, estimated on a model-based perspective. A graphical representation of the relation between the variables in the system is shown in Figure 6.2, which depicts the joint model's architecture.

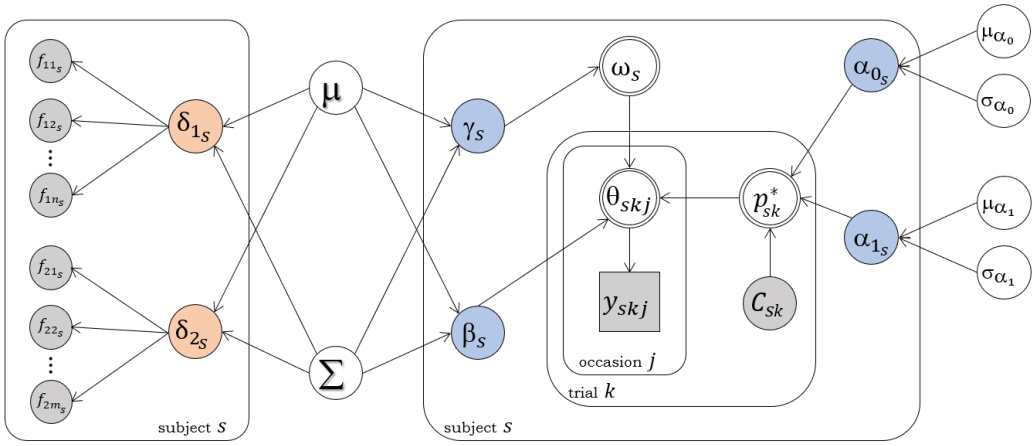


FIGURE 6.2: Covariance model's architecture. Square and circular nodes indicate discrete and continous variables, respectively. Grey nodes indicate observed variables. Blue and red nodes represent behavioural and neural node parameters, respectively. Double-circled nodes represent deterministic nodes.

The graphical model represents all the (in)dependencies assumptions between the variables in the system. Given the model's assumptions we can compute the joint posterior distribution of the model parameters conditional on observed data as follows:

$$
\begin{aligned}
p(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \boldsymbol{\Omega}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, &\sigma_1, \sigma_2, \mu_{\alpha_0}, \sigma_{\alpha_0}, \mu_{\alpha_1}, \sigma_{\alpha_1} | \boldsymbol{Y}, \boldsymbol{F}) \propto \\
&\prod_s \left[p(\boldsymbol{Y}_s | \gamma_s, \beta_s, \alpha_{0s}, \alpha_{1s}) p(\boldsymbol{f}_s^{(1)} | \delta_{1s}, \sigma_1) p(\boldsymbol{f}_s^{(2)} | \delta_{2s}, \sigma_2) \right] \\
&\prod_s p(\delta_{1s}, \delta_{2s}, \gamma_s, \beta_s | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \, p(\alpha_{0s} | \mu_{\alpha_0}, \sigma_{\alpha_0}) \, p(\alpha_{1s} | \mu_{\alpha_1}, \sigma_{\alpha_1}) \\
&p(\boldsymbol{\mu}) p(\boldsymbol{\Sigma}) p(\sigma_1) p(\sigma_2) p(\mu_{\alpha_0}) p(\sigma_{\alpha_0}) p(\mu_{\alpha_1}) p(\sigma_{\alpha_1})
\end{aligned}
\tag{6.8}
$$

where s represents individuals, and $\boldsymbol{f}_s^{(1)}$ (resp. $\boldsymbol{f}_s^{(2)}$) reflects the FA values for the brain tract connections in ACC–Insula–IFG network (resp. dlPFC–Thalamus–Striatum network) for individual s.

The first row on the right side of Equation (6.8) represents the likelihood of the joint structure which simultaneously includes behavioural and neural model likelihoods, the second row represents the related priors according to model factorization, that is, the multivariate probability model for the random vector of individuals neural and cognitive parameters and the probability models for the two regression coefficients. The third row specifies the hyper priors. Such factorization allows computation of marginal posterior distributions via Markov Chain Monte Carlo algorithms (MCMC; (Gilks et al., 1995)). The following probability distributions were used for the hyper priors:

$$
\begin{aligned}
\mu_\gamma, \mu_\beta &\sim \text{Normal}(0, 10^3)_{I(0,\infty)} \\
\mu_{\delta_1}, \mu_{\delta_2} &\sim \text{Normal}(0, 10^3) \\
\sigma_\gamma, \sigma_\beta &\sim \text{Gamma}(0.01, 0.01) \\
\sigma_{\delta_1}, \sigma_{\delta_2} &\sim \text{Gamma}(0.01, 0.01) \\
\sigma_1, \sigma_2 &\sim \text{Gamma}(0.01, 0.01) \\
\rho_1, \rho_2, \rho_3, \rho_4 &\sim \text{Uniform}(-1, 1) \\
\mu_{\alpha_0} &\sim \text{Normal}(0, 10^3)_{I(0,3)} \\
\sigma_{\alpha_0} &\sim \text{Gamma}(0.01, 0.01) \\
\mu_{\alpha_1} &\sim \text{Normal}(0, 1)_{I(-0.2,0.2)} \\
\sigma_{\alpha_1} &\sim \text{Uniform}(0, 1)
\end{aligned}
$$

Parameter reflecting the degrees of freedom of the Multivariate Student's t-distribution has been treated as a tuning parameter in order to ensure algorithm convergence and chains mixing (see Appendix C for details).

6.3 Results

For the model fitting, one participant was excluded from the analysis due to corrupted and unreliable MRI scan. The joint model was then fitted to data from the remaining 23 participants. The data array consisted of the collection of pumps and cashes across trials and the custom structural connectivity matrix for each subject $(\boldsymbol{Y}_s, \boldsymbol{F}_s)$.

All calculations were performed with the aim of the efficient interaction between R (Team et al., 2013) and JAGS (Plummer et al., 2003) using the package "R2jags" (Su et al., 2015). A probabilistic programming implementation of the bayesian graphical model architectures was then provided and posterior distributions were computed using Gibbs Sampling algorithm (Casella and George, 1992). We ran 12 chains of 15,000 iterations each, with a burn-in period of 5000 iterations and a thinning size of 1, parallelized on an Intel i7 6 cores CPU. Thus, we obtained 120,000 samples from the joint posterior. The total time required to perform the computation was about 35 minutes. Table 6.2 summarizes some of the posterior densities of interest.

TABLE 6.2: Marginal posterior distributions statistics: Posterior mean (μ_{post}), 95% credible intervals $[q_{0.05}, q_{0.975}]$, chains convergence ($\widehat{\mathrm{R}}$).

Parameter	μ_{post}	$q_{0.05}$	$q_{0.975}$	$\widehat{\mathrm{R}}$
μ_γ	0.442	0.374	0.474	1.012
μ_β	1.471	1.211	1.571	1.013
μ_{α_0}	2.653	2.460	2.722	1.001
μ_{α_1}	-0.004	-0.007	-0.001	1.001
ρ_1	-0.341	-0.85	0.365	1.019
ρ_2	-0.483	-0.86	0.072	1.010
ρ_3	0.021	-0.645	0.750	1.013
ρ_4	-0.250	-0.761	0.371	1.008

Posterior marginals were sampled efficiently and the 12 chains showed an optimal convergence as measured by the $\widehat{\mathrm{R}}$ statistic (Gelman et al., 1992), and the trace plot of the log joint posterior density (Figure 6.3). Values of $\widehat{\mathrm{R}}$ approaching 1 indicate better convergence.

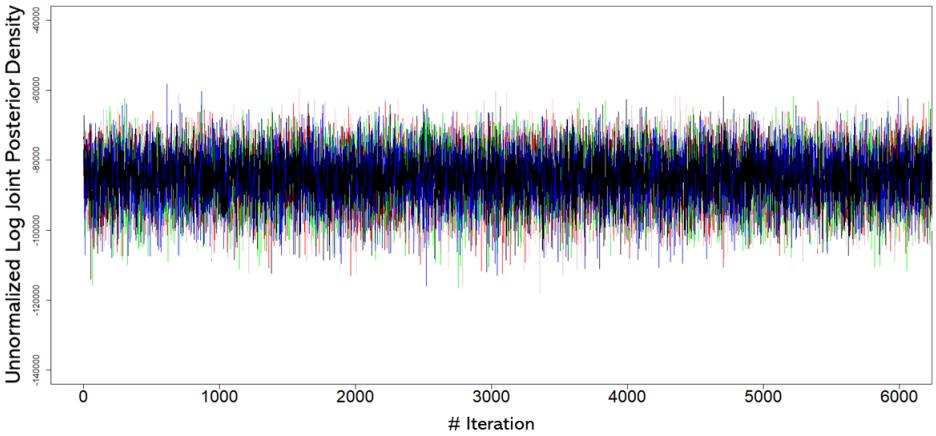


FIGURE 6.3: Trace plot of the (unnormalized) log posterior density computed for all the chains, for the first 6000 iterations. The burn-in period was removed to show the whole convergence dynamic. As can be noticed, the log posterior seems to show no trends.

Therefore, the joint model seemed to fit the data adequately by allowing a reliable recovery of cognitive parameters describing observed behaviour. Figure 6.4 shows results from posterior predictive check, which, in bayesian modelling, is the benchmark method to assess effective model fit (Gelman and Shalizi, 2013). We compared observed data to synthetic model-generated data produced by parameters drawn from the posterior distribution. Model fit adequacy was evaluated based on how much synthetic data resemble empirical data. We generated posterior predictives of 1000 datasets of pumps and cashes patterns on 36 trials, for 1000 cognitive parameter sets $(\gamma_s, \beta_s, \alpha_{0s}, \alpha_{1s})$ sampled from the joint posterior distribution corresponding to each individual s. Empirical distributions of number of pumps were then compared with recovered distributions.

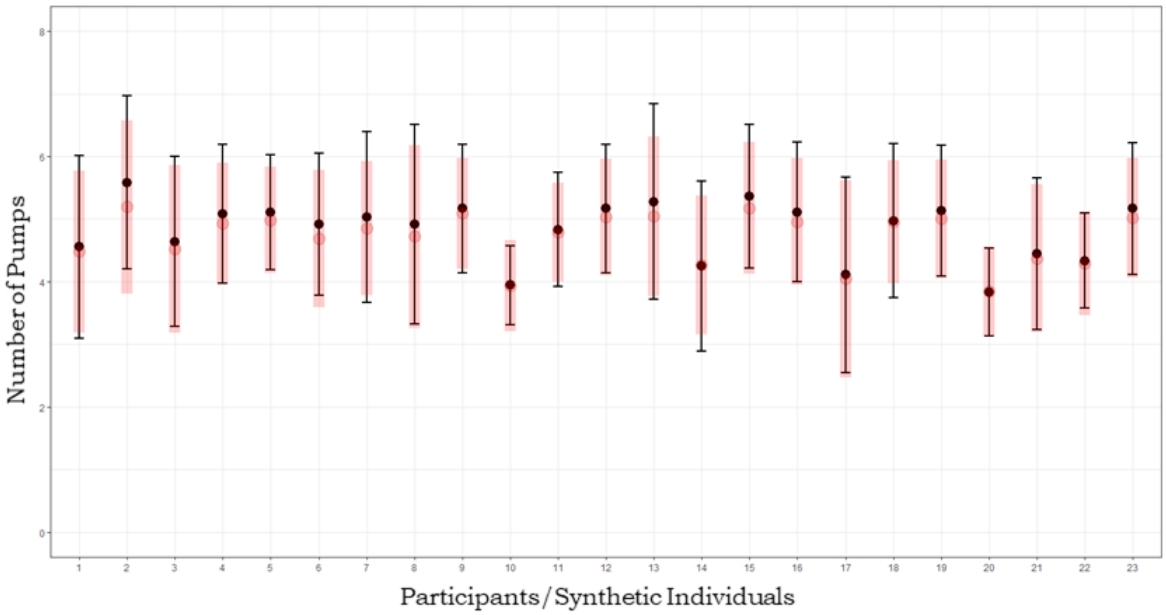


FIGURE 6.4: Posterior predictive check. Black dots and boundaries represent mean pumps and standard deviations for each individual from the empirical dataset. Red dots and lines represent mean pumps and standard deviations of predicted synthetic individual datasets.

Population (hyper-)means allow to interpret individual differences in performance in terms of few parameters reflecting the assumptions about the process generating individual-level parameters (Lee, 2011b).

At the population level, individuals seemed to modify their bursting belief only very slightly during the unfolding of the task (posterior mean $\mu_{\alpha_1} = -0.004$), and in general subjective bursting probabilities can be considered constant across the trials span. Therefore, individuals showed a relatively low level of risk-taking (posterior mean $\mu_\gamma = 0.442$) and a relatively high level of behaviour consistency (posterior mean $\mu_\beta = 1.471$) leading to low response variability.

In our confirmatory framework we aimed to verify whether such cognitive parameters configuration was related to Network FA. The multivariate distribution of the joint model allows to characterize such relation by treating groups of individual-level parameters as covariates. Thus, posterior densities of correlation parameters of the covariance matrix convey information on how individual differences in brain networks structural integrity and cognitive characteristics account for differences in performance. Figure 6.5 shows the estimated posterior distribution for the correlations between risk-taking, γ, and both ACC–Insula–IFG and dlPFC–Thalamus–Striatum networks FA, ρ_1 and ρ_2, respectively, and that between behavioural consistency, β, and dlPFC–Thalamus–Striatum network FA, ρ_4. The correlation between behavioural consistency and ACC–Insula–IFG is not shown since it has not substantial evidence.

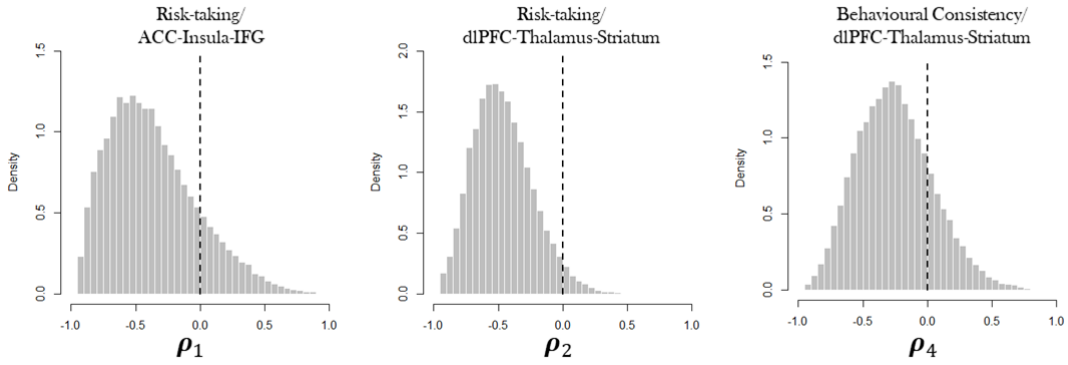


FIGURE 6.5: Marginal posterior distributions of the correlation parameters of interest in the covariance matrix.

In general, the relationships between parameters are weak, except for ρ_2, but the figure indicates a moderate inverse relation in all the three cases. Increase in risk-taking propensity was related to decreased white matter microstructure integrity in two networks which codify for loss and risk aversion and for goal-directed behaviour. Such results might posit some constraints on the quantification of the actual role of risk propensity in enacting an optimal decision strategy. Participants were indeed required to adopt a balanced risky-oriented strategy in order to maximise earnings, and taking more risk at different stages of the task could be seen as an adaptive strategy which increases the chance to produce more positive outcome (Dean et al., 2011; Kohno et al., 2017). From this perspective, however, risk-taking propensity might not be the main component to fulfil the task of optimizing earnings. The finding of an inverse relation between behavioural consistency and dlPFC–Thalamus–Striatum network seems to clarify the role of white matter structural properties in predicting the adoption of a functional cognitive strategy when performing the BART. Individuals presenting an increased fractional anisotropy in such network showed an increased response variability (decreased behavioural consistency). This relation might reflect, in healthy individuals, the tendency to approach and explore the environment by choosing actions whose outcome is uncertain but potentially advantageous (Humphries et al., 2012; Goldenberg et al., 2017), as reflected by the functionality of the network.

6.4 General discussion

In the present work, we proposed an approach to the modelling of the neural structural substrates underlying risky behaviour within a joint modelling framework, inspired by previous works on joint analysis by means of hierarchical bayesian models (Turner et al., 2013; Turner et al., 2017). A behavioural model allowing for estimation of meaningful cognitive parameters was developed and coupled with neural parameters in a multivariate probability model.

This made the analysis of the relation between decision-making and brain structural connectivity interpretable on a model-based perspective.

The presented methodology application is thought to provide an example for cognitive scientists who are interested in investigating dependencies between behavioural and neural data via computational models. When applied to the experimental context of the BART, our approach shows several useful advantages.

First, the proposed computational framework is extremely flexible and has the potential to combine neural and cognitive models with several assumptions and complexities. This comes in handy when different BART configurations are considered (e.g., a priori knowledge of the bursting probability) and consistent behavioural model assumptions have to be made accordingly (e.g., removing the node related to the bursting belief dynamic, α_1, from the graphical model).

Second, our method provides a way to infer the relationships between the biological properties of the brain and higher-level cognitive processes involved in risky decision-making by overcoming some limitations of the standard approach. For instance, inferences about the role of given cognitive mechanisms and their neural correlates in producing behavioural outcome, such as total monetary earning and relative frequency of pumps and cashes, implicitly assume a mapping between cognitive processes and summary measures of individuals task output. As a consequence, neurocognitive theories are built upon resulting relations between such statistics and brain measurements. Decomposing the data-generating process in several psychological sub-processes allows, instead, to relate brain measurements directly to cognitive variables of interest. In this respect, our computational model is valuable from a theoretical perspective since it can be used to test hypothesis about how neural variables predict cognitive functioning and behaviour in risk conditions, in a substantial formal way.

Therefore, the proposed architecture can be modularly extended to account for the presence of several explanatory variables. Thus, different covariates in the joint structure can be employed based on conditional dependence assumptions. As an example, one might think of explicitly modelling subjective bursting probabilities as predicted by discrete or continuous covariates such as sex, age, or self-report measures.

Moreover, several brain networks can be put in relation to cognitive variables of interest by extending the connectivity matrix $\boldsymbol{F}$ and the possible network subsets, and by extending the covariance matrix, $\boldsymbol{\Sigma}$, to include more correlation parameters on a model-based perspective.

Despite these advantages, the proposed computational framework has some limitations. A first technical limitation might concern the fact that posterior probability computations become demanding and potentially unstable when the number of free parameters in the covariance matrix of the multivariate neurocognitive probability model increases. However, the Multivariate Student's t-distribution adopted in the proposed application has shown to overcome some computational problems by ensuring posterior sampling chains mixing.

Second, our joint modelling framework assumes a previously defined parametric representation of the multivariate model accounting for cognitive and neural parameters. This might constitute a severe constraint since a probabilistic model describing neural data may not always be consistent, especially when complex brain structural measures, such as those provided by DTI, are considered.

It is worth noticing that other neural measures might be adopted to parameterize individual structural brain connectivity, and that embedding such measures in a joint structure could be far from trivial. In our model we adopted FA as an exemplary application, due to its popularity and (relatively) ease in being reliably computed. However, a more complex and exhaustive neural measure accounting for white matter anisotropy might be the Generalized Fractional Anisotropy (GFA, (Cohen-Adad et al., 2008)). This is computed by using a more complete information on diffusion sampling directions and yields anisotropy maps with a higher angular resolution that might efficiently replace FA. Nevertheless, GFA estimates are particularly sensitive to noise and probably unreliable unless high angular sampling is available (Zhan et al., 2010). For this reason, a more accurate account of noise estimates has to be considered and formally instantiated in the joint graphical model, especially in relation to the covariance structure which directly relate neural and behavioural parameters.

In conclusion, we think that the proposed approach offers interesting insights in the development of computational models able to investigate correspondence between decision-making and brain structural connectivity. Further works are needed to investigate the potentials of the joint framework to account for BART performances and neural characteristics of individuals in clinical populations.

Chapter 7

General Discussion

In the present dissertation, the application of the PGM framework to model cognitive phenomena has been proposed at multiple levels of abstraction.

In Chapter 3 we explored the usage of a Latent Markov Model (LMM) approach for longitudinal-like data to capture dynamics in response strategy in the WCST from a group-level perspective. The modeling approach has been thought of as a tool to improve the amount of information attainable from the analysis of the scoring of response outcomes. From this perspective, the latent state process served as a proxy to synthesize distributional properties of scoring measures into latent traits, which, in general, need an a posteriori interpretation. The main advantage of this approach is that moving from a summary measures-based to a latent traits-based level of analysis, evidences for an effective dynamic in observed responses can be achieved in a principled way. However, such an approach leaves unsolved the pervasive question about how such behavioural dynamics might emerge from cognitive dynamics. Indeed, the LMM did not provide a model of cognition neither at a representational, nor at a computational, level of explanation. Therefore, a crucial requirement for such a modeling approach to be valuable for clinical and research practices is that the standard scoring measures accounting for the observation process in our LMM framework, must capture at least non-overlapping and separable psychological constructs.

The second project, depicted in Chapter 4, tried to investigate such a separability issue from a computational modeling perspective. As a natural follow-up after having highlighted the importance of accounting for behavioural outcome dynamics in a demanding environment (e.g. the WCST), one might be interested in investigating how such behaviour emerges by means of generative models. In this work, the flexibility of the PGM framework has been used to simulate interactions between cognitive sub-components. In particular, we adopted a confirmatory perspective with the aim to translate accrued knowledge from the literature in a DBN structure which resembled the main features of the cognitive (sub)system responsible for behavioural dynamics in the WCST. This provided the chance to test the capability of the scoring measures to capture the heterogeneous corpus of (parameterized) cognitive sub-component interactions. This can be conceived as an emblematic employment of PGMs in cognitive research, due to the ability of such a framework to allow structuring an artificial cognitive system which retains some of the main property of human cognition, such as the hierarchical organization of processing and the stochastic aspect of information passing.

Differently, Chapter 5 tried to embed the cognitive principles involved in set-shifting, within a consistent and overarching theory of cognitive functioning. Despite the complexity of the Bayesian computational mechanisms which are assumed to take place in a cognitive system, the Bayesian brain model proposed here was built upon a discrete state architecture with given conditional (in)dependence assumptions. In this case, the generative model was inverted in order to estimate cognitive parameters which allowed to reconstruct the process of belief formation during the fulfillment of the WCST. The main novelty of the proposed model is that the PGM representation, together with the formalization of a Bayesian cognitive computational framework, allowed to recover trial-by-trial second-order measures accounting for the interaction between the agent and the environment. More precisely, while cognitive parameters, flexibility and information loss, contribute to shape the belief updating process during the unfolding of the task, the information-theoretic measures allow to quantify the effect that an event in the environment elicits on the cognitive system (e.g. Shannon surprise following an unexpected observation).

Finally, in Chapter 6, we dedicated to a dynamic environment where participants were required to find optimal policies to maximize earning in risky conditions. Although the model adopted here was not a strict DBN, it definitely took advantage of the PGM framework to represent behavioural dynamics in a principled way. The behavioural model, in particular, is dynamic in the sense that the subjective bursting probability might change trial-by-trial, conditioned on the previous experience of the individual. Both conditional (in)dependence assumptions, and the discrete trialwise structuring of the evolving system, allowed to explicitly account for the hidden process dynamics (e.g. trial dependent pumping probabilities) driving the unfolding of behaviour in time. Therefore, the flexibility of the PGM framework allowed to embed the behavioural model within a more complex joint neural-behavioural model. In this way, conditional (in)dependence constraints were not limited to the behavioural model alone, but were also extended to to the neural model by allowing the specification of (potentially) every possible relationship between neural information and cognitive parameters.

The main purpose of the present dissertation was to show the potentials of adopting a PGM perspective to understand cognition by computational modeling. As already discussed, we provided ways to (1) obtain new and more exhaustive scoring measures from the analysis of behavioural outcomes on a set-shifting task; (2) transfer existing knowledge about a phenomenon to a probabilistic graph architecture of cognitive functioning to explore behavioural outcomes; (3) embed a complex theory of cognition within a probabilistic framework to estimate cognitive parameters and develop insight about the interaction between an agent and a demanding environment; (4) jointly model behavioural and neural information. Consistently, we argue that PGMs can be considered as an optimal candidate to provide a unified mathematical language to model cognitive phenomena which unfold in time and that can be approximated by a discrete state and time representation.

However, it is worth emphasizing that when discrete cognitive states are

taken into account as building blocks to model cognitive dynamics, we are restricting our attention to some particular snapshot of information processing, by minimizing the chance of understanding the (behavioural) data-generating process at a finer time scale (e.g. the brain dynamics time scale). To some extent, the discretization procedure is both a necessary assumption ensuring mathematical tractability and models' parsimony, and the optimal solution to the problem of mapping computational model constructs to behavioural outcomes, when the task considered is specifically designed to require a response elicitation in a single snapshot (e.g. a single trial). Note that this might be a crucial requirement for our modeling framework to succeed in building tractable and interpretable cognitive models. Future works might be interested in assessing the suitability of the PGM framework to be integrated with models accounting for local dynamics, such as diffusion process models of decisions, or to be embedded in a more complex model representation structure where both continuous and discrete time cognitive dynamics are taken into account.

Appendix A

Conditional Response Probabilities comparison

The following tables show the conditional response probabilities estimates for three models with different number of states (1-state, left; 2-state, center; 3-state, right).

	$\hat{\phi}_{y\|s}$
y	$s=1$
C	0.80
E	0.11
PE	0.09

	$\hat{\phi}_{y\|s}$	
y	$s=1$	$s=2$
C	0.92	0.67
E	0.02	0.20
PE	0.06	0.13

	$\hat{\phi}_{y\|s}$		
y	$s=1$	$s=2$	$s=3$
C	0.93	0.80	0.44
E	0.02	0.10	0.38
PE	0.05	0.10	0.18

The 1-state model can be considered as a baseline model which accounts for the absence of dynamics in the performance trend. The 2-state and the 3-state models are the candidate models in the main work. Our qualitative model selection criteria relies on comparing their conditional probabilities matrices. As can be noticed, the State 2 in the 2-state model reflects the error-related cognitive strategy. However, in our view, the error-related strategy can be decomposed in order to obtain two types of non-optimal strategies accounting for different degree of non-perseverative and perseverative components of the error. A discussion on this point can be found in the ”Discussion of Results” section in the main manuscript.

Appendix B

Full joint probability distribution for all the individuals

In what follows, full joint posterior densities are provided for the two cognitive parameters, namely, flexibility (λ) and information loss (δ), together with the mean coordinates (dotted lines) used as point estimates to recover information-theoretic quantities in the main text.

Figure B.1 shows posterior densities for the healthy group.

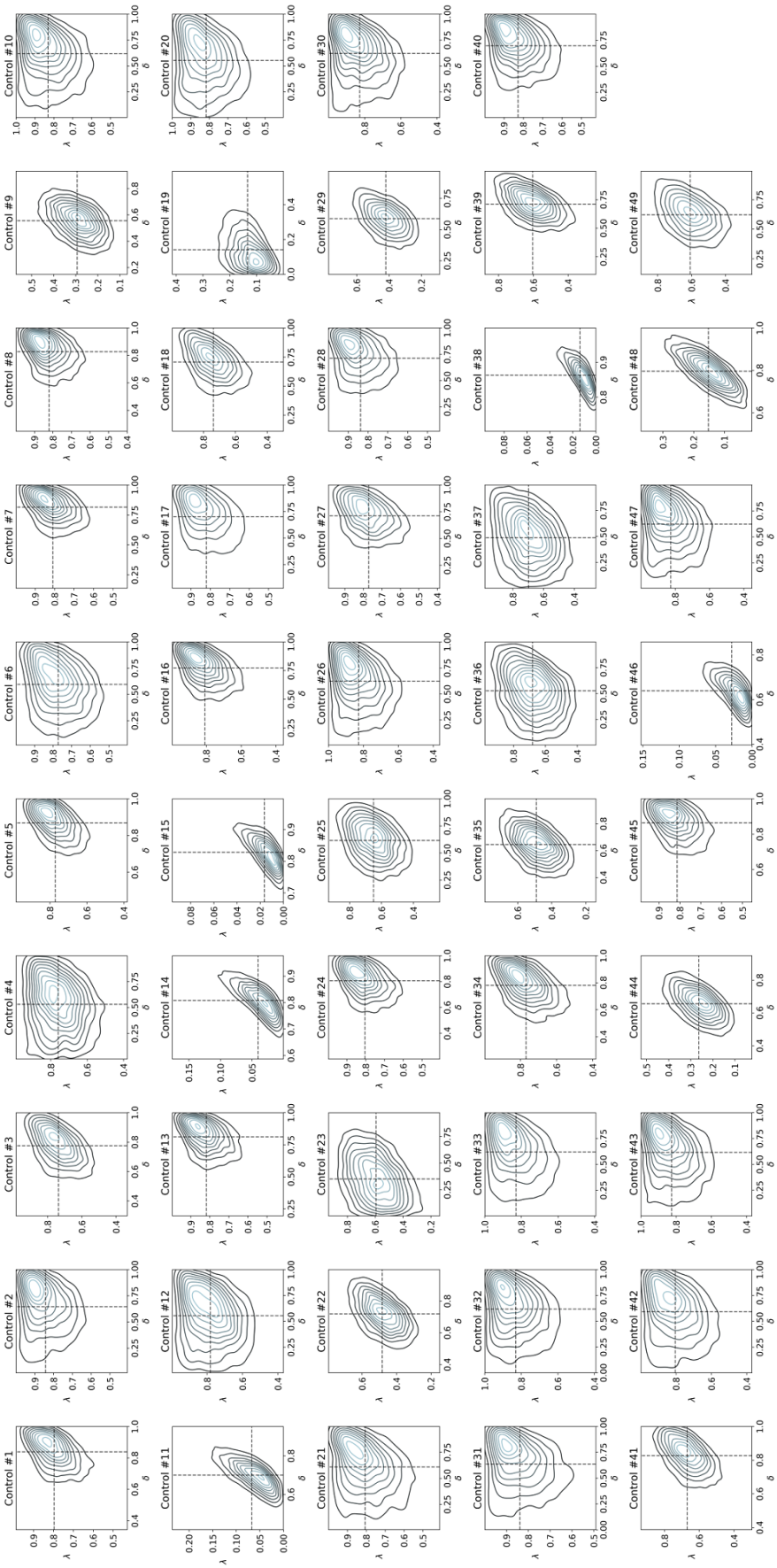


FIGURE B.1: Posterior densities for healthy individuals. Dotted lines show the distribution mean.

Figure B.2 shows posterior densities for the substance dependent individuals.

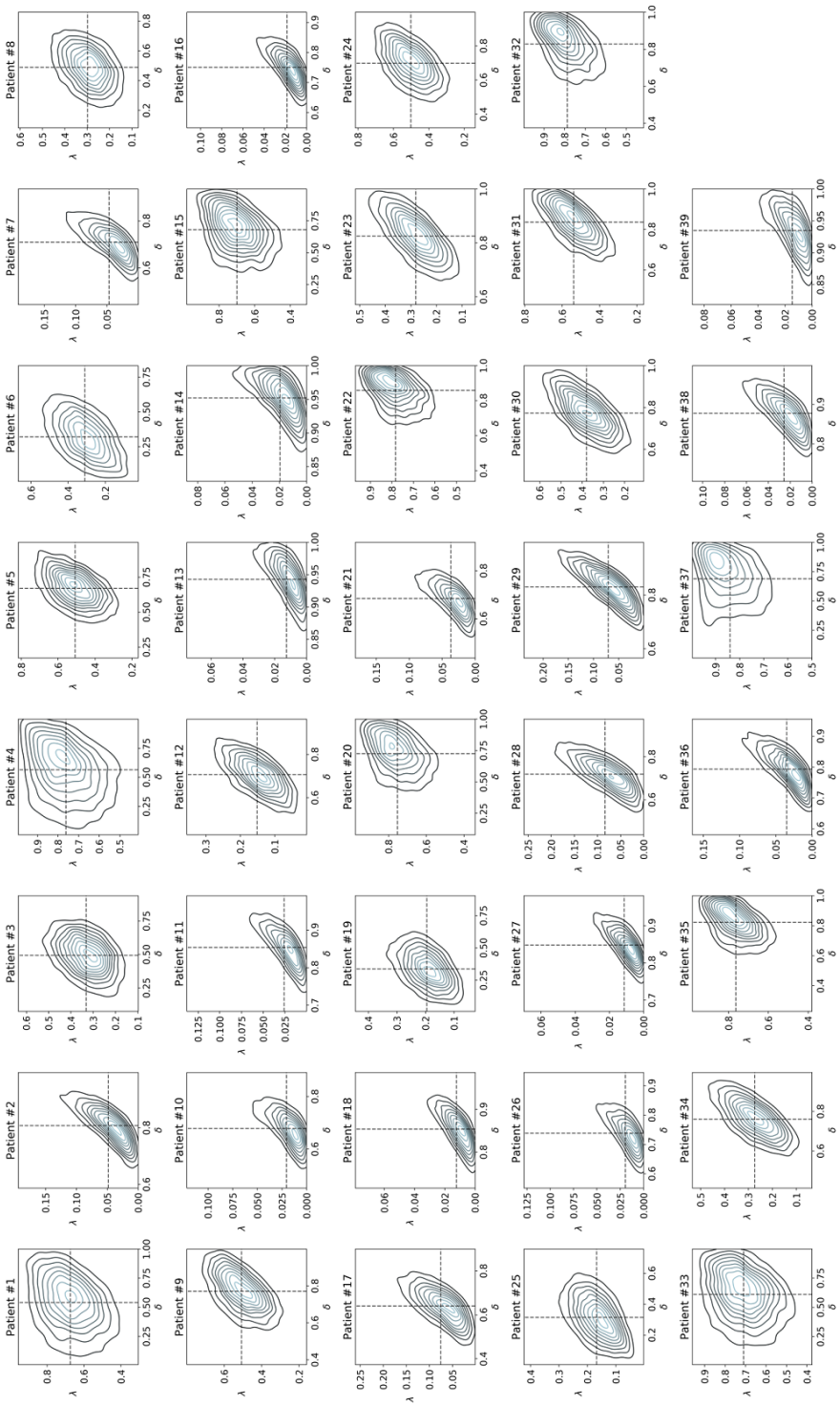


FIGURE B.2: Posterior densities for substance dependent individuals. Dotted lines show the distribution mean.

Appendix C

Multivariate Model tuning

C.1 Multivariate Student's t-distribution specification

Here we discuss some computational details related to the tuning of the Multivariate Student's t-distribution. More precisely, we focus on the estimation of the correlation parameters in the covariance matrix of the multivariate model.

In our model representation, we assumed a sparse covariance matrix in which some correlation parameters were fixed to zero since our confirmatory approach focused on testing specific meaningful relations between cognitive and neural parameters. Thus, prior distributions should be considered for each correlation coefficient in the decomposed covariance matrix.

The JAGS probabilistic programming framework allows to embed prior distributions in the hierarchical model by considering the Multivariate Student's t-distribution as modelled according to a precision matrix instead of a covariance matrix. Matrix inversion is then needed when correlation coefficients have to be obtained. However, matrix inversion problems may arise when the assumption of positive-definite matrix is violated, and it is often the case in which this happens.

Differently, prior distributions for the covariance matrix parameters can be specified when the Multivariate Normal distribution is considered. This allows to estimate correlation coefficients when the Multivariate Student's t-distribution is taken into account by overcoming the limitations related to matrix inversion.

Consider the vector-valued random variable $\boldsymbol{X}^* = [\gamma^*, \beta^*, \alpha_0^*, \alpha_1^*]$. We assume $\boldsymbol{X}^*$ to have a Standard Multivariate Normal distribution:

$$\boldsymbol{X}^* \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

with 4-dimensional zero mean vector and covariance matrix $\boldsymbol{\Sigma}$ as structured according to the main text. Consider now the vector-valued random variable $\boldsymbol{X} = [\gamma, \beta, \alpha_0, \alpha_1]$ containing the individual level cognitive and neural parameters. We model $\boldsymbol{X}$ as a Multivariate Student's t distributed random variable as follows:

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{X}^* \left(\sqrt{\frac{\zeta}{V}} \right)$$

where $\boldsymbol{\mu} = [\mu_\gamma, \mu_\beta, \mu_{\delta_1}, \mu_{\delta_2}]$ and V is a Chi-square distributed random variable with parameter ρ denoting degrees of freedom. Note that parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are estimated in a hierarchical bayesian framework in the main text.

Here, ζ was fixed to a default value such that $\zeta = 5$. In general, values of ζ in the range [5-30] do not compromise posterior sampled chains mixing in our application. However, when ρ is treated as a parameter to estimate and left free to vary within a broader range, chains mixing is not ensured and parameter estimates are unreliable.

C.2 Simulation study and tuning parameter

In this section we provide a simulation study aimed to explore model performance in meaningful scenarios. In particular, a Monte Carlo 3-factorial design is employed to recover Effective Sample Size of posterior MCMC samples and Computation Time across levels of three factors, namely, Number of subjects, Number of ROI-to-ROI connections (which we refer to as ROIs), and the tuning scenarios. In particular, in each cell of the factorial design, parameters are sampled from the prior, synthetic neural and behavioural data are simulated based on the sampled parameters, and the simulated data pattern is used to fit the joint model. For each cell, the process of data simulation and model fitting is repeated for 10 times. Number of subjects are allowed to vary across three levels, that is, $[10, 30, 50]$, whilst the Number of ROIs across two levels, that is, $[5, 10]$. Simulating neural data consists in directly sampling an FA measure related to a specific ROI-to-ROI connection. The tuning factor consists in two levels in which the degrees of freedom of Multivariate t-distribution is fixed to a default value such that $\zeta = 5$, or is treated as a free parameter with an exponential prior such that $\zeta \sim \text{Exp}(1/30)$. Such a factorial structure entails a configuration of 12 cells, and a total of 120 data simulations and model fitting. To make the computations feasible we adopt a simplified version of the joint model in which the behavioural model consists only in parameters γ and β, risk taking and response variability, respectively. The remaining nodes of the graphical model remain unchanged.

As in the main work, the calculations are performed in R, and MCMC samples are obtain via Gibbs Sampling. For computational convenience 6 chains of 5000 iterations each are used, with a burn-in period of 500 iterations and a thinning size of 1. Computations are parallelized on an Intel i7 6 cores CPU. The total factorial design simulation time was 14 hours.

The mean computation time to estimate parameters of the joint model within each cell of the factorial design is reported in Figure C.1.

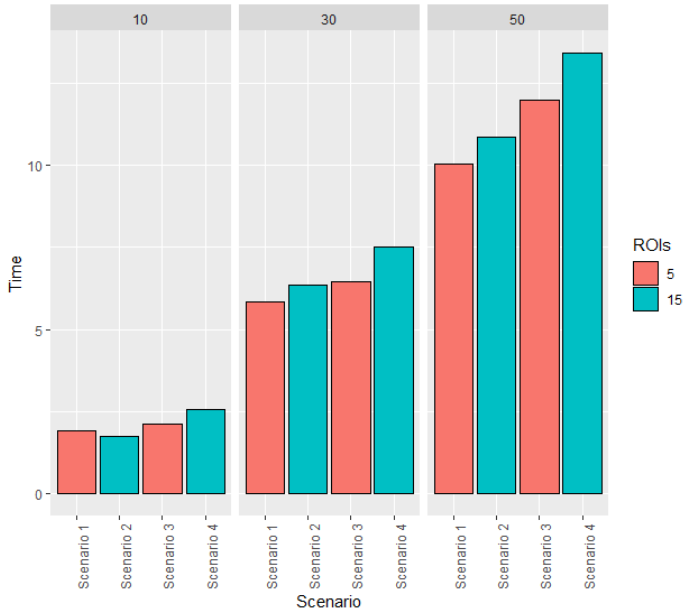


FIGURE C.1: Mean computation time (in minutes) for each cell of the factorial design. Here, scenario 1 and 2 refer to the case in which ζ is fixed, whilst scenario 3 and 4 refer to the case in which ζ is treated as a free parameter.

As can be noticed, computation time increases slightly linearly, based on the number of individuals. Increasing the number of ROIs seems to contribute to extend the computation time especially for higher sample sizes. Scenarios in which ζ is treated as a free parameter (scenario 3 and 4) are the most computationally expensive. It is worth noticing that, in general, computations are rather cheap and this is due to the simplified joint model adopted. The increase of the computation time might not be linear in case the full model is employed.

Figure C.2 shows the distribution of the Effective Sample Size across the MCMC joint posterior samples for each cell of the factorial design. Such a metric helps assessing the convergence of the MCMC sampels, but it also quantifies how much independent information there is in autocorrelated chains. Having non-autocorrelated chains ensure to decrease the uncertainty of the estimation of posterior quantities of interest, such as credible intervals, which are useful for empirical research. As a substantiated heuristic, the higher the Effective Sample size the better.

It is computed according to:

$$\mathrm{ESS} = \frac{N}{1 + 2\sum_{k=1}^{\infty} \mathrm{ACF}(k)} \tag{C.1}$$

where N is the number of samples for a given chain, and $\mathrm{ACF}(\cdot)$ is the autocorrelation function at lag k.

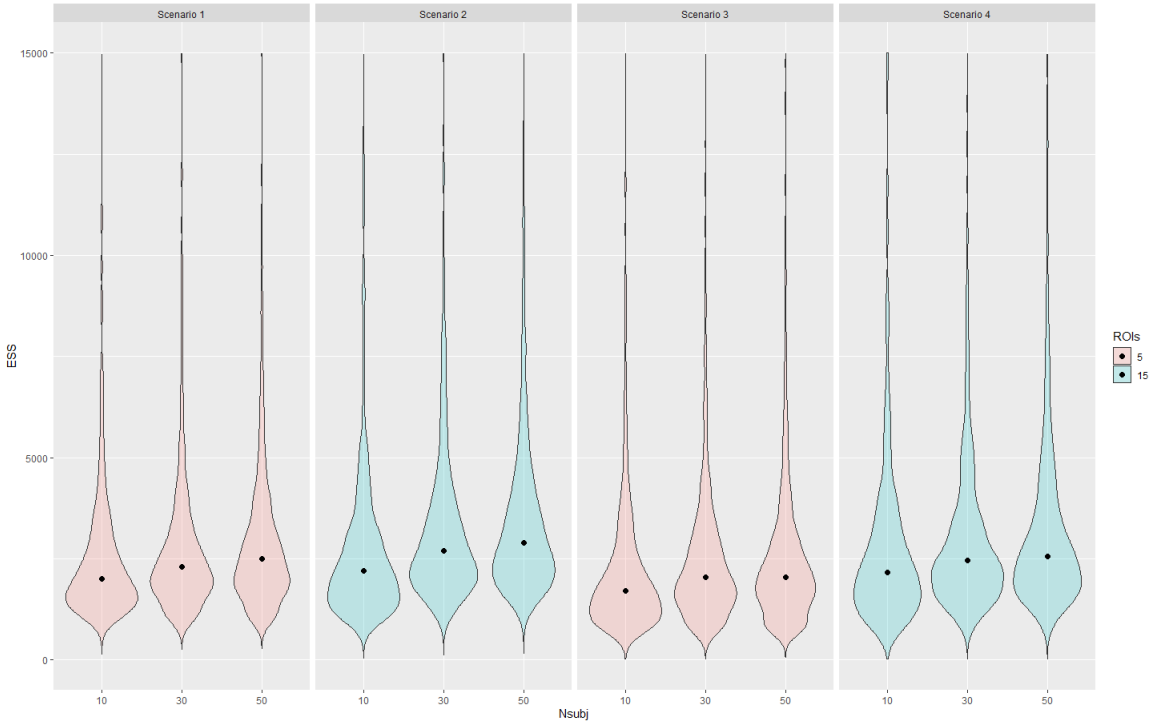


FIGURE C.2: Effective Sample Size distribution for each cell of the factorial design. Scenarios 1 and 2 refer to fixed ζ and scenarios 3 and 4 to ζ as a free parameter.

In general, scenarios in which ζ is fixed to a default value show a higher Effective Sample Size compared to that of the scenarios in which ζ is sampled and included in the joint posterior. Enriching information yielded by the data, by increasing both sample size and ROIs, seems to ensure a more reliable posterior distributions. This provides an advantage when more data are available, due the fact that computation time seems to be affected by larger data structure to a lesser extent. In a similar way, reliability of posterior samples when ζ is assumed as a free parameter seems to be affected by the amount of data available. On the other hand, ζ seems to be generally unreliable even when larger datasets are considered. In our case, mean Effective Sample Size estimates for ζ samples were 143(SD=18) (resp. 272(SD=34)) for 50 synthetic subjects and 5 ROIs (resp. 15 ROIs).

Bibliography

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Aklin, W. M., Lejuez, C., Zvolensky, M. J., Kahler, C. W., and Gwadz, M. (2005). Evaluation of behavioral measures of risk taking propensity with inner city adolescents. *Behaviour research and therapy*, 43(2):215–228.

Alexander, W. H. and Brown, J. W. (2018). Frontal cortex function as derived from hierarchical predictive coding. *Scientific reports*, 8(1):1–11.

Alvarez, J. A. and Emory, E. (2006). Executive function and the frontal lobes: a meta-analytic review. *Neuropsychology review*, 16(1):17–42.

Amos, A. (2000). A computational model of information processing in the frontal cortex and basal ganglia. *Journal of cognitive neuroscience*, 12(3):505–519.

Anderson, P. J. (2008). Towards a developmental model of executive function. *Executive functions and the frontal lobes: A lifespan perspective*, 3:21.

Anderson, P. J. (2010). Towards a developmental model of executive function. In *Executive functions and the frontal lobes*, pages 37–56. Psychology Press.

Badcock, P. B., Friston, K. J., Ramstead, M. J., Ploeger, A., and Hohwy, J. (2019). The hierarchically mechanistic mind: an evolutionary systems theory of the human brain, cognition, and behavior. *Cognitive, Affective, & Behavioral Neuroscience*, 19(6):1319–1351.

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

Barceló, F. (2001). Does the wisconsin card sorting test measure prefrontal function. *The Spanish journal of psychology*, 4(1):79–100.

Barcelo, F., Escera, C., Corral, M. J., and Periáñez, J. A. (2006). Task switching and novelty processing activate a common neural network for cognitive control. *Journal of cognitive neuroscience*, 18(10):1734–1748.

Barceló, F. and Knight, R. T. (2002). Both random and perseverative errors underlie wcst deficits in prefrontal patients. *Neuropsychologia*, 40(3):349–356.

Barceló, F. and Rubia, F. J. (1998). Non-frontal p3b-like activity evoked by the wisconsin card sorting test. *Neuroreport*, 9(4):747–751.

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2012). *Latent Markov models for longitudinal data*. CRC Press.

Bartolucci, F. and Solis-Trapala, I. L. (2010). Multidimensional latent markov models in a developmental study of inhibitory control and attentional flexibility in early childhood. *Psychometrika*, 75(4):725–743.

Basser, P. J. and Pierpaoli, C. (1996). Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor mri. *Journal of magnetic resonance, Series B*, 111(3):209–219.

Batchelder, W. H. (2010). Cognitive psychometrics: Using multinomial processing tree models as measurement tools.

Bechara, A., Damasio, A. R., and Damasio, H. (2001). Insensitivity to future consequences following damage to human prefrontal. *The Science of Mental Health: Personality and personality disorder*, 50:287.

Bechara, A. and Damasio, H. (2002). Decision-making and addiction (part i): impaired activation of somatic states in substance dependent individuals when pondering decisions with negative future consequences. *Neuropsychologia*, 40(10):1675–1689.

Beppu, T., Inoue, T., Shibata, Y., Kurose, A., Arai, H., Ogasawara, K., Ogawa, A., Nakamura, S., and Kabasawa, H. (2003). Measurement of fractional anisotropy using diffusion tensor mri in supratentorial astrocytic tumors. *Journal of neuro-oncology*, 63(2):109–116.

Berg, E. A. (1948). A simple objective technique for measuring flexibility in thinking. *The Journal of general psychology*, 39(1):15–22.

Bestmann, S., Ruge, D., Rothwell, J., and Galea, J. M. (2014). The role of dopamine in motor flexibility. *Journal of cognitive neuroscience*, 27(2):365–376.

Bishara, A. J., Kruschke, J. K., Stout, J. C., Bechara, A., McCabe, D. P., and Busemeyer, J. R. (2010). Sequential learning models for the wisconsin card sort task: Assessing processes in substance dependent individuals. *Journal of mathematical psychology*, 54(1):5–13.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Bornovalova, M. A., Daughters, S. B., Hernandez, G. D., Richards, J. B., and Lejuez, C. (2005). Differences in impulsivity and risk-taking propensity between primary users of crack cocaine and primary users of heroin in a

residential substance-use program. *Experimental and clinical psychopharmacology*, 13(4):311.

Braff, D. L., Heaton, R., Kuck, J., Cullum, M., Moranville, J., Grant, I., and Zisook, S. (1991). The generalized pattern of neuropsychological deficits in outpatients with chronic schizophrenia with heterogeneous wisconsin card sorting test results. *Archives of general psychiatry*, 48(10):891–898.

Bridwell, D. A., Cavanagh, J. F., Collins, A. G., Nunez, M. D., Srinivasan, R., Stober, S., and Calhoun, V. D. (2018). Moving beyond erp components: a selective review of approaches to integrate eeg and behavior. *Frontiers in human neuroscience*, 12:106.

Brouwer, G. J. and Heeger, D. J. (2011). Cross-orientation suppression in human visual cortex. *Journal of neurophysiology*, 106(5):2108–2119.

Buchsbaum, B. R., Greer, S., Chang, W.-L., and Berman, K. F. (2005). Meta-analysis of neuroimaging studies of the wisconsin card-sorting task and component processes. *Human brain mapping*, 25(1):35–45.

Buckley, C. L., Kim, C. S., McGregor, S., and Seth, A. K. (2017). The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, 81:55–79.

Buckner, R. L., Andrews-Hanna, J. R., and Schacter, D. L. (2008). The brain's default network: anatomy, function, and relevance to disease.

Buesing, L., Bill, J., Nessler, B., and Maass, W. (2011). Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS Comput Biol*, 7(11):e1002211.

Burgess, P. W., Alderman, N., Evans, J., Emslie, H., and Wilson, B. (1998). The ecological validity of tests of executive function. *Journal of the international neuropsychological society*, 4(6):547–558.

Busemeyer, J. R. and Stout, J. C. (2002). A contribution of cognitive decision models to clinical assessment: decomposing performance on the bechara gambling task. *Psychological assessment*, 14(3):253.

Busemeyer, J. R., Stout, J. C., and Finn, P. (2003). Using computational models to help explain decision making processes of substance abusers. *Cognitive and affective neuroscience of psychopathology*, pages 1–41.

Carlson, S. M. and Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child development*, 72(4):1032–1053.

Casella, G. and George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174.

Cazzell, M., Li, L., Lin, Z.-J., Patel, S. J., and Liu, H. (2012). Comparison of neural correlates of risk decision making between genders: an exploratory fnirs study of the balloon analogue risk task (bart). *Neuroimage*, 62(3):1896–1911.

Cella, M., Bishara, A. J., Medin, E., Swan, S., Reeder, C., and Wykes, T. (2014). Identifying cognitive remediation change through computational modelling—effects on reinforcement learning in schizophrenia. *Schizophrenia bulletin*, 40(6):1422–1432.

Chater, N., Tenenbaum, J. B., and Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations.

Chevalier, N. and Blaye, A. (2008). Cognitive flexibility in preschoolers: The role of representation activation and maintenance. *Developmental science*, 11(3):339–353.

Christidi, F., Karavasilis, E., Samiotis, K., Bisdas, S., and Papanikolaou, N. (2016). Fiber tracking: A qualitative and quantitative comparison between four different software tools on the reconstruction of major white matter tracts. *European journal of radiology open*, 3:153–161.

Christopoulos, G. I., Tobler, P. N., Bossaerts, P., Dolan, R. J., and Schultz, W. (2009). Neural correlates of value, risk, and risk aversion contributing to decision making under risk. *Journal of Neuroscience*, 29(40):12574–12583.

Cohen, J. R. and Poldrack, R. A. (2014). Materials and methods for openfmri ds009: The generality of self control.

Cohen-Adad, J., Descoteaux, M., Rossignol, S., Hoge, R. D., Deriche, R., and Benali, H. (2008). Detection of multiple pathways in the spinal cord using q-ball imaging. *Neuroimage*, 42(2):739–749.

Collell, G. and Fauquet, J. (2015). Brain activity and cognition: a connection from thermodynamics and information theory. *Frontiers in psychology*, 6:818.

Cooper, R., Fox, J., Farringdon, J., and Shallice, T. (1996). A systematic methodology for cognitive modelling. *Artificial Intelligence*, 85(1-2):3–44.

Coulacoglou, C. and Saklofske, D. H. (2017). *Psychometrics and psychological assessment: Principles and applications*. Academic Press.

Crone, E. A., Richard Ridderinkhof, K., Worm, M., Somsen, R. J., and Van Der Molen, M. W. (2004). Switching between spatial stimulus–response mappings: a developmental study of cognitive flexibility. *Developmental science*, 7(4):443–455.

Dagum, P., Galper, A., and Horvitz, E. J. (1991). Temporal probabilistic reasoning: Dynamic network models for forecasting. knowledge systems laboratory, medical computer science.

Dai, J., Pleskac, T. J., and Pachur, T. (2018). Dynamic cognitive models of intertemporal choice. *Cognitive psychology*, 104:29–56.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215.

de Haan-Rietdijk, S., Kuppens, P., Bergeman, C. S., Sheeber, L., Allen, N., and Hamaker, E. (2017). On the use of mixed markov models for intensive longitudinal data. *Multivariate behavioral research*, 52(6):747–767.

Dean, A. C., Sugar, C. A., Hellemann, G., and London, E. D. (2011). Is all risk bad? young adult cigarette smokers fail to take adaptive risk in a laboratory decision-making test. *Psychopharmacology*, 215(4):801–811.

Dehaene, S. and Changeux, J.-P. (1991). The wisconsin card sorting test: Theoretical analysis and modeling in a neuronal network. *Cerebral cortex*, 1(1):62–79.

Demakis, G. J. (2003). A meta-analytic review of the sensitivity of the wisconsin card sorting test to frontal and lateralized frontal brain damage. *Neuropsychology*, 17(2):255.

Dempster, F. N. (1992). The rise and fall of the inhibitory mechanism: Toward a unified theory of cognitive development and aging. *Developmental review*, 12(1):45–75.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980.

Dobrow, R. P. (2016). *Introduction to stochastic processes with R*. John Wiley & Sons.

Duffy, J. D. and Campbell III, J. J. (2001). Regional prefrontal syndromes: A theoretical and clinical overview.

Erdfelder, E., Hu, X., Rouder, J. N., and Wagenmakers, E.-J. (2020). Cognitive psychometrics: The scientific legacy of william h. batchelder (1940-2018). *Journal of Mathematical Psychology*, 99(Article 102468):1–7.

Farrell, S. and Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.

Feldman, H. and Friston, K. (2010). Attention, uncertainty, and free-energy. *Frontiers in human neuroscience*, 4:215.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of applied statistics*, 31(7):799–815.

Figueroa, I. J. and Youmans, R. J. (2013). Failure to maintain set: A measure of distractibility or cognitive flexibility? In *Proceedings of the human factors and ergonomics society annual meeting*, volume 57, pages 828–832. Sage Publications Sage CA: Los Angeles, CA.

First, M. B. (1997). Structured clinical interview for dsm-iv axis i disorders. *Biometrics Research Department*.

Flashman, L. A., Homer, M. D., and Freides, D. (1991). Note on scoring perseveration on the wisconsin card sorting test. *The Clinical Neuropsychologist*, 5(2):190–194.

Forstmann, B. U., Wagenmakers, E.-J., Eichele, T., Brown, S., and Serences, J. T. (2011). Reciprocal relations between cognitive neuroscience and formal cognitive models: opposites attract? *Trends in cognitive sciences*, 15(6):272–279.

Freeman, J. B. and Ambady, N. (2010). Mousetracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior research methods*, 42(1):226–241.

Friston, K. (2003). Learning and inference in the brain. *Neural Networks*, 16(9):1325–1352.

Friston, K. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456):815–836.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138.

Friston, K., Adams, R., Perrinet, L., and Breakspear, M. (2012). Perceptions as hypotheses: saccades as experiments. *Frontiers in psychology*, 3:151.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017a). Active inference: a process theory. *Neural computation*, 29(1):1–49.

Friston, K. and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1211–1221.

Friston, K. J., Daunizeau, J., Kilner, J., and Kiebel, S. J. (2010). Action and behavior: a free-energy formulation. *Biological cybernetics*, 102(3):227–260.

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., and Ondobaka, S. (2017b). Active inference, curiosity and insight. *Neural computation*, 29(10):2633–2683.

Fukunaga, R., Brown, J. W., and Bogg, T. (2012). Decision making in the balloon analogue risk task (bart): anterior cingulate cortex signals loss aversion but not the infrequency of risky choices. *Cognitive, Affective, & Behavioral Neuroscience*, 12(3):479–490.

Furman, D. J., Hamilton, J. P., and Gotlib, I. H. (2011). Frontostriatal functional connectivity in major depressive disorder. *Biology of mood & anxiety disorders*, 1(1):11.

Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Gelman, A. and Shalizi, C. R. (2013). Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38.

Gershman, S. J. (2015). A unifying probabilistic view of associative learning. *PLoS Comput Biol*, 11(11):e1004567.

Gershman, S. J. (2018). The successor representation: its computational logic and neural substrates. *Journal of Neuroscience*, 38(33):7193–7200.

Ghahramani, Z. (1997). Learning dynamic bayesian networks. In *International School on Neural Networks, Initiated by IIASS and EMFCSC*, pages 168–197. Springer.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. Chapman and Hall/CRC.

Gläscher, J., Adolphs, R., and Tranel, D. (2019). Model-based lesion mapping of cognitive control using the wisconsin card sorting test. *Nature communications*, 10(1):1–12.

Glaser, J. I., Perich, M. G., Ramkumar, P., Miller, L. E., and Kording, K. P. (2018). Population coding of conditional probability distributions in dorsal premotor cortex. *Nature communications*, 9(1):1–14.

Goldenberg, D., Telzer, E. H., Lieberman, M. D., Fuligni, A. J., and Galván, A. (2017). Greater response variability in adolescents is associated with increased white matter development. *Social cognitive and affective neuroscience*, 12(3):436–444.

Greve, K. W., Stickle, T. R., Love, J. M., Bianchini, K. J., and Stanford, M. S. (2005). Latent structure of the wisconsin card sorting test: a confirmatory factor analytic study. *Archives of Clinical Neuropsychology*, 20(3):355–364.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., and Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences*, 14(8):357–364.

Haker, H., Schneebeli, M., and Stephan, K. E. (2016). Can bayesian theories of autism spectrum disorder help improve clinical practice? *Frontiers in psychiatry*, 7:107.

Harrison, L., David, O., and Friston, K. (2005). Stochastic models of neuronal dynamics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1457):1075–1091.

Hawkins, G. E., Mittner, M., Forstmann, B. U., and Heathcote, A. (2017). On the efficiency of neurally-informed cognitive models to identify latent cognitive states. *Journal of Mathematical Psychology*, 76:142–155.

Heaton, R. (1981). Wisconsin card sorting test manual; revised and expanded. *Psychological Assessment Resources*, pages 5–57.

Hirsh, J. B., Mar, R. A., and Peterson, J. B. (2012). Psychological entropy: A framework for understanding uncertainty-related anxiety. *Psychological review*, 119(2):304.

Hsieh, P. C., Yeh, T. L., Lee, I. H., Huang, H. C., Chen, P. S., Yang, Y. K., Chiu, N. T., Lu, R. B., and Liao, M.-H. (2010). Correlation between errors on the wisconsin card sorting test and the availability of striatal dopamine transporters in healthy volunteers. *Journal of psychiatry & neuroscience: JPN*, 35(2):90.

Hull, R., Martin, R. C., Beier, M. E., Lane, D., and Hamilton, A. C. (2008). Executive function in older adults: a structural equation modeling approach. *Neuropsychology*, 22(4):508.

Humphries, M. D., Khamassi, M., and Gurney, K. (2012). Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Frontiers in neuroscience*, 6:9.

Hyndman, R. J., Koehler, A. B., Snyder, R. D., and Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of forecasting*, 18(3):439–454.

Jimura, K., Konishi, S., and Miyashita, Y. (2004). Dissociable concurrent activity of lateral and medial frontal lobe during negative feedback processing. *Neuroimage*, 22(4):1578–1586.

Kam, J. W., Solbakk, A.-K., Endestad, T., Meling, T. R., and Knight, R. T. (2018). Lateral prefrontal cortex lesion impairs regulation of internally and externally directed attention. *NeuroImage*, 175:91–99.

Kersten, D., Mamassian, P., and Yuille, A. (2004). Object perception as bayesian inference. *Annu. Rev. Psychol.*, 55:271–304.

Kibby, M. Y., Schmitter-Edgecombe, M., and Long, C. J. (1998). Ecological validity of neuropsychological tests: focus on the california verbal learning test and the wisconsin card sorting test. *Archives of Clinical Neuropsychology*, 13(6):523–534.

Knill, D. C. and Pouget, A. (2004). The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719.

Koechlin, E. and Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in cognitive sciences*, 11(6):229–235.

Kohno, M., Morales, A. M., Guttman, Z., and London, E. D. (2017). A neural network that links brain function, white-matter structure and risky behavior. *Neuroimage*, 149:15–22.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Kopp, B. (2012). A simple hypothesis of executive function. *Frontiers in Human Neuroscience*, 6:159.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., and Shams, L. (2007). Causal inference in multisensory perception. *PLoS one*, 2(9):e943.

Kragel, J. E., Morton, N. W., and Polyn, S. M. (2015). Neural activity in the medial temporal lobe reveals the fidelity of mental time travel. *Journal of Neuroscience*, 35(7):2914–2926.

Krain, A. L., Wilson, A. M., Arbuckle, R., Castellanos, F. X., and Milham, M. P. (2006). Distinct neural mechanisms of risk and ambiguity: a meta-analysis of decision-making. *Neuroimage*, 32(1):477–484.

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

Krawitz, A., Fukunaga, R., and Brown, J. W. (2010). Anterior insula activity predicts the influence of positively framed messages on decision making. *Cognitive, Affective, & Behavioral Neuroscience*, 10(3):392–405.

Kruschke, J. K. (2008). Models of categorization. *The Cambridge handbook of computational psychology*, pages 267–301.

Kübler, A., Murphy, K., and Garavan, H. (2005). Cocaine dependence and attention switching within and between verbal and visuospatial working memory. *European Journal of Neuroscience*, 21(7):1984–1992.

Kwon, M. S., Vorobyev, V., Moe, D., Parkkola, R., and Hämäläinen, H. (2014). Brain structural correlates of risk-taking behavior and effects of peer influence in adolescents. *PloS one*, 9(11).

Landry, O. and Al-Taie, S. (2016). A meta-analysis of the wisconsin card sort task in autism. *Journal of autism and developmental disorders*, 46(4):1220–1235.

Lane, S. D., Steinberg, J. L., Ma, L., Hasan, K. M., Kramer, L. A., Zuniga, E. A., Narayana, P. A., and Moeller, F. G. (2010). Diffusion tensor imaging and decision making in cocaine dependence. *PLoS one*, 5(7).

Lawson, R. P., Rees, G., and Friston, K. J. (2014). An aberrant precision account of autism. *Frontiers in human neuroscience*, 8:302.

Lee, M. (2011a). Special issue on hierarchical bayesian models. *Journal of Mathematical Psychology*, 55:1–118.

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

Lee, M. D. and Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge university press.

Lee, T. S. and Mumford, D. (2003). Hierarchical bayesian inference in the visual cortex. *JOSA A*, 20(7):1434–1448.

Lejuez, C., Aklin, W. M., Jones, H. A., Richards, J. B., Strong, D. R., Kahler, C. W., and Read, J. P. (2003). The balloon analogue risk task (bart) differentiates smokers and nonsmokers. *Experimental and clinical psychopharmacology*, 11(1):26.

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., Strong, D. R., and Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: the balloon analogue risk task (bart). *Journal of Experimental Psychology: Applied*, 8(2):75.

Levine, D. S., Parks, R. W., and Prueitt, P. S. (1993). Methodological and theoretical issues in neural network models of frontal cognitive functions. *International Journal of Neuroscience*, 72(3-4):209–233.

Lewandowsky, S. and Farrell, S. (2010). *Computational modeling in cognition: Principles and practice*. SAGE publications.

Lie, C.-H., Specht, K., Marshall, J. C., and Fink, G. R. (2006). Using fmri to decompose the neural processes underlying the wisconsin card sorting test. *Neuroimage*, 30(3):1038–1049.

Lu, Z.-L., Li, X., Tjan, B. S., Dosher, B. A., and Chu, W. (2011). Attention extracts signal in external noise: a bold fmri study. *Journal of Cognitive Neuroscience*, 23(5):1148–1159.

Ma, W. J. (2012). Organizing probabilistic models of perception. *Trends in cognitive sciences*, 16(10):511–518.

Mack, M. L., Preston, A. R., and Love, B. C. (2013). Decoding the brain's algorithm for categorization from its neural implementation. *Current Biology*, 23(20):2023–2027.

Maes, J., Damen, M., and Eling, P. (2004). More learned irrelevance than perseveration errors in rule shifting in healthy subjects. *Brain and cognition*, 54(3):201–211.

Malloy, P. F. and Richardson, E. D. (1994). Assessment of frontal lobe functions. *The Journal of Neuropsychiatry and Clinical Neurosciences*.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information, henry holt and co. *Inc., New York, NY*, 2(4.2).

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., and Stephan, K. E. (2014). Uncertainty in perception and the hierarchical gaussian filter. *Frontiers in human neuroscience*, 8:825.

Miller, E. K. and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24(1):167–202.

Miller, H. L., Ragozzino, M. E., Cook, E. H., Sweeney, J. A., and Mosconi, M. W. (2015). Cognitive set shifting deficits and their relationship to repetitive behaviors in autism spectrum disorder. *Journal of autism and developmental disorders*, 45(3):805–815.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., and Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive psychology*, 41(1):49–100.

Monchi, O., Petrides, M., Petre, V., Worsley, K., and Dagher, A. (2001). Wisconsin card sorting revisited: distinct neural circuits participating in different stages of the task identified by event-related functional magnetic resonance imaging. *Journal of Neuroscience*, 21(19):7733–7741.

Monchi, O., Taylor, J. G., and Dagher, A. (2000). A neural model of working memory processes in normal subjects, parkinson's disease and schizophrenia for fmri design and predictions. *Neural Networks*, 13(8-9):953–973.

Moriguchi, Y., Okanda, M., and Itakura, S. (2008). Young children's yes bias: How does it relate to verbal ability, inhibitory control, and theory of mind? *First Language*, 28(4):431–442.

Murphy, K. P. and Russell, S. (2002). Dynamic bayesian networks: representation, inference and learning.

Nagahama, Y., Okina, T., Suzuki, N., Nabatame, H., and Matsuda, M. (2005). The cerebral correlates of different types of perseveration in the wisconsin card sorting test. *Journal of Neurology, Neurosurgery & Psychiatry*, 76(2):169–175.

Nour, M. M., Dahoun, T., Schwartenbeck, P., Adams, R. A., FitzGerald, T. H., Coello, C., Wall, M. B., Dolan, R. J., and Howes, O. D. (2018). Dopaminergic basis for signaling belief updates, but not surprise, and the link to paranoia. *Proceedings of the National Academy of Sciences*, 115(43):E10167–E10176.

Nunez, M. D., Vandekerckhove, J., and Srinivasan, R. (2017). How attention influences perceptual decision making: Single-trial eeg correlates of drift-diffusion model parameters. *Journal of mathematical psychology*, 76:117–130.

Nyhus, E. and Barceló, F. (2009). The wisconsin card sorting test and the cognitive assessment of prefrontal executive functions: a critical update. *Brain and cognition*, 71(3):437–451.

Ott, T. and Nieder, A. (2019). Dopamine and cognitive control in prefrontal cortex. *Trends in Cognitive Sciences*.

O'Reilly, J. X., Schüffelgen, U., Cuell, S. F., Behrens, T. E., Mars, R. B., and Rushworth, M. F. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proceedings of the National Academy of Sciences*, 110(38):E3660–E3669.

Palestro, J. J., Bahg, G., Sederberg, P. B., Lu, Z.-L., Steyvers, M., and Turner, B. M. (2018). A tutorial on joint models of neural and behavioral measures of cognition. *Journal of Mathematical Psychology*, 84:20–48.

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier.

Pennoni, F. (2014). *Issues on the estimation of latent variable and latent class models*. Scholars' Press.

Petzschner, F. H., Glasauer, S., and Stephan, K. E. (2015). A bayesian perspective on magnitude estimation. *Trends in cognitive sciences*, 19(5):285–293.

Pierpaoli, C. and Basser, P. J. (1996). Toward a quantitative assessment of diffusion anisotropy. *Magnetic resonance in Medicine*, 36(6):893–906.

Pinheiro, J. C., Liu, C., and Wu, Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution.

Plummer, M. et al. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, pages 1–10. Vienna, Austria.

Poldrack, R. A., Barch, D. M., Mitchell, J., Wager, T., Wagner, A. D., Devlin, J. T., Cumba, C., Koyejo, O., and Milham, M. (2013). Toward open sharing of task-based fmri data: the openfmri project. *Frontiers in neuroinformatics*, 7:12.

Pouget, A., Beck, J. M., Ma, W. J., and Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature neuroscience*, 16(9):1170–1178.

Radev, S. T., Mertens, U. K., Voss, A., Ardizzone, L., and Köthe, U. (2020). Bayesflow: Learning complex stochastic models with invertible neural networks.

Rolls, E. T. and Deco, G. (2010). *The noisy brain: stochastic dynamics as a principle of brain function*, volume 34. Oxford university press Oxford.

Rolls, E. T., Joliot, M., and Tzourio-Mazoyer, N. (2015). Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. *Neuroimage*, 122:1–5.

Rushworth, M. F. and Behrens, T. E. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature neuroscience*, 11(4):389.

Rybakowski, J., Borkowska, A., Czerski, P., Kapelski, P., Dmitrzak-Weglarz, M., and Hauser, J. (2005). An association study of dopamine receptors polymorphisms and the wisconsin card sorting test in schizophrenia. *Journal of neural transmission*, 112(11):1575–1582.

Sayood, K. (2018). Information theory and cognition: A review. *Entropy*, 20(9):706.

Schwartenbeck, P., FitzGerald, T. H., and Dolan, R. (2016). Neural signals encoding shifts in beliefs. *Neuroimage*, 125:578–586.

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

Scott, W. A. (1962). Cognitive complexity and cognitive flexibility. *Sociometry*, pages 405–414.

Singh, S., Aich, T. K., and Bhattarai, R. (2017). Wisconsin card sorting test performance impairment in schizophrenia: An indian study report. *Indian journal of psychiatry*, 59(1):88.

Sisson, S. A. and Fan, Y. (2011). *Likelihood-free MCMC*. Chapman & Hall/CRC, New York.[839].

Smallwood, J. and Schooler, J. W. (2015). The science of mind wandering: empirically navigating the stream of consciousness. *Annual review of psychology*, 66:487–518.

Speekenbrink, M., Lagnado, D. A., Wilkinson, L., Jahanshahi, M., and Shanks, D. R. (2010). Models of probabilistic category learning in parkinson's disease: Strategy use and the effects of l-dopa. *Journal of Mathematical Psychology*, 54(1):123–136.

Steinke, A., Lange, F., Seer, C., Hendel, M. K., and Kopp, B. (2020). Computational modeling for neuropsychological assessment of bradyphrenia in parkinson's disease. *Journal of Clinical Medicine*, 9(4):1158.

Stelzel, C., Basten, U., Montag, C., Reuter, M., and Fiebach, C. J. (2010). Frontostriatal involvement in task switching depends on genetic differences in d2 receptor density. *Journal of Neuroscience*, 30(42):14205–14212.

Stephan, K. E., Baldeweg, T., and Friston, K. J. (2006). Synaptic plasticity and dysconnection in schizophrenia. *Biological psychiatry*, 59(10):929–939.

Stoianov, I., Genovesio, A., and Pezzulo, G. (2016). Prefrontal goal codes emerge as latent states in probabilistic value learning. *Journal of Cognitive Neuroscience*, 28(1):140–157.

Stoianov, I. and Zorzi, M. (2012). Emergence of a'visual number sense'in hierarchical generative models. *Nature neuroscience*, 15(2):194–196.

Strange, B. A., Duggins, A., Penny, W., Dolan, R. J., and Friston, K. J. (2005). Information theory, novelty and hippocampal responses: unpredicted or unpredictable? *Neural Networks*, 18(3):225–230.

Stuss, D., Levine, B., Alexander, M., Hong, J., Palumbo, C., Hamer, L., Murphy, K., and Izukawa, D. (2000). Wisconsin card sorting test performance in patients with focal frontal and posterior brain damage: effects of lesion location and test structure on separable cognitive processes. *Neuropsychologia*, 38(4):388–402.

Su, Y.-S., Yajima, M., Su, M. Y.-S., and SystemRequirements, J. (2015). Package 'r2jags'. *R package version 0.03-08, URL http://CRAN. R-project. org/package= R2jags*.

Sun, R. (2008). *The Cambridge handbook of computational psychology*. Cambridge University Press.

Sun, R. (2009). Theoretical status of computational cognitive modeling. *Cognitive Systems Research*, 10(2):124–140.

Taghia, J., Cai, W., Ryali, S., Kochalka, J., Nicholas, J., Chen, T., and Menon, V. (2018). Uncovering hidden brain state dynamics that regulate performance and decision-making during cognition. *Nature communications*, 9(1):1–19.

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. (2018). Validating bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*.

Tarter, R. E. (1973). An analysis of cognitive deficits in chronic alcoholics. *Journal of Nervous and Mental Disease*.

Team, R. C. et al. (2013). R: A language and environment for statistical computing.

Tenenbaum, J. B., Griffiths, T. L., and Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, 10(7):309–318.

Thulasiraman, K. and Swamy, M. N. (2011). *Graphs: theory and algorithms*. John Wiley & Sons.

Turner, B. M., Forstmann, B. U., Love, B. C., Palmeri, T. J., and Van Maanen, L. (2017). Approaches to analysis in model-based cognitive neuroscience. *Journal of Mathematical Psychology*, 76:65–79.

Turner, B. M., Forstmann, B. U., Wagenmakers, E.-J., Brown, S. D., Sederberg, P. B., and Steyvers, M. (2013). A bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, 72:193–206.

Van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and brain sciences*, 21(5):615–628.

van Ravenzwaaij, D., Dutilh, G., and Wagenmakers, E.-J. (2011). Cognitive model decomposition of the bart: Assessment and application. *Journal of Mathematical Psychology*, 55(1):94–105.

van Ravenzwaaij, D., Provost, A., and Brown, S. D. (2017). A confirmatory approach for integrating neural and behavioral data into a single model. *Journal of Mathematical Psychology*, 76:131–141.

Visser, I. (2011). Seven things to remember about hidden markov models: A tutorial on markovian models for time series. *Journal of Mathematical Psychology*, 55(6):403–415.

Wallsten, T. S., Pleskac, T. J., and Lejuez, C. W. (2005). Modeling behavior in a clinically diagnostic sequential risk-taking task. *Psychological review*, 112(4):862.

Welsh, A. and Richardson, A. (1997). Approaches to the robust estimation of mixed models handbook of statistics.

Wiggins, L. M. (1973). Panel analysis: Latent probability models for attitude and behavior processes.

Willuhn, I., Sun, W., and Steiner, H. (2003). Topography of cocaine-induced gene regulation in the rat striatum: relationship to cortical inputs and role of behavioural context. *European Journal of Neuroscience*, 17(5):1053–1066.

Yechiam, E., Goodnight, J., Bates, J. E., Busemeyer, J. R., Dodge, K. A., Pettit, G. S., and Newman, J. P. (2006). A formal cognitive model of the go/no-go discrimination task: Evaluation and implications. *Psychological assessment*, 18(3):239.

Yeh, F.-C., Panesar, S., Fernandes, D., Meola, A., Yoshino, M., Fernandez-Miranda, J. C., Vettel, J. M., and Verstynen, T. (2017). A population-based atlas of the macroscale structural connectome in the human brain. *bioRxiv*, page 136473.

Yeh, F.-C., Verstynen, T. D., Wang, Y., Fernández-Miranda, J. C., and Tseng, W.-Y. I. (2013). Deterministic diffusion fiber tracking improved by quantitative anisotropy. *PloS one*, 8(11).

Yousefi, A., Basu, I., Paulk, A. C., Peled, N., Eskandar, E. N., Dougherty, D. D., Cash, S. S., Widge, A. S., and Eden, U. T. (2019). Decoding hidden cognitive states from behavior and physiology using a bayesian approach. *Neural computation*, 31(9):1751–1788.

Yuille, A. and Kersten, D. (2006). Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, 10(7):301–308.

Zabelina, D. L. and Robinson, M. D. (2010). Creativity as flexible cognitive control. *Psychology of Aesthetics, Creativity, and the Arts*, 4(3):136.

Zakzanis, K. K. (1998). The subcortical dementia of huntington's disease. *Journal of Clinical and Experimental Neuropsychology*, 20(4):565–578.

Zelazo, P. D., Carter, A., Reznick, J. S., and Frye, D. (1997). Early development of executive function: A problem-solving framework. *Review of general psychology*, 1(2):198–226.

Zelazo, P. D., Müller, U., Frye, D., Marcovitch, S., Argitis, G., Boseovski, J., Chiang, J. K., Hongwanishkul, D., Schuster, B. V., Sutherland, A., et al. (2003). The development of executive function in early childhood. *Monographs of the society for research in child development*, pages i–151.

Zhan, L., Leow, A. D., Jahanshad, N., Chiang, M.-C., Barysheva, M., Lee, A. D., Toga, A. W., McMahon, K. L., De Zubicaray, G. I., Wright, M. J., et al. (2010). How does angular resolution affect diffusion imaging measures? *Neuroimage*, 49(2):1357–1371.

Zucchini, W., Raubenheimer, D., and MacDonald, I. L. (2008). Modeling time series of animal behavior by means of a latent-state model with feedback. *Biometrics*, 64(3):807–815.