UNIVERSITÀ DEGLI STUDI DI MILANO – BICOCCA

Dipartimento di Psicologia

Dottorato di Ricerca in Psicologia Sociale, Cognitiva e Clinica
XXV CICLO

RULE-GUIDED BEHAVIOUR: HOW AND WHERE RULES ARE REPRESENTED AND PROCESSED IN HUMAN BRAIN.

Tutor: Chiar.mo Prof. Paolo CHERUBINI

Tesi di Dottorato di:

Doris PISCHEDDA

Matricola N. 056882

Anno Accademico 2012/2013

To all those…

… who follow the norms and are successful.

… who break the rules and don't regret it.

… who grasp the secret laws of the world and produce science.

TABLE OF CONTENTS

ABSTRACT **1**

ABBREVIATIONS **2**

INTRODUCTION **5**

CHAPTER 1 ..... **8**

1. Theoretical framework **8**

1.1. Rule-guided behaviour **8**

1.2. Rule representation **11**

1.3. Rule processing **14**

CHAPTER 2 ..... **16**

2. Methods **16**

2.1. Physical principles of MRI **16**

2.2. Functional Magnetic Resonance Imaging **19**

2.3. Univariate analysis of fMRI data **22**

2.4. Multivariate analysis **27**

2.4.1. Feature selection and dimensionality reduction 28

2.4.2. Classification 29

2.4.3. Statistics 32

2.4.4. Applications 33

CHAPTER 3 ..... **35**

3. Study 1: Automatic processing of conditional rules **35**

3.1. Theoretical background **35**

3.2. Experiment 1 ..... **36**

3.2.1. Methods 36

3.2.2. Results 40

3.2.3. Discussion 42

3.3. Experiment 2 ..... **43**

3.3.1. Methods 44

3.3.2. Results 45

3.3.3. Discussion 46

3.4. Experiment 3 ..... **47**

3.4.1. Methods 48

3.4.2. Results 50

3.4.3. Discussion 52

3.5. Experiment 4 ..... **53**

3.5.1. Methods 53

3.5.2. Results 55

3.5.3. Discussion 56

3.6. General discussion **57**

3.7. Conclusions **61**

CHAPTER 4 ..... **62**

4. Study 2: Unattended rules cause response conflict **62**

4.1. Theoretical background **62**

4.2. Methods **64**

4.2.1. Critical contrasts 66

4.3. Analyses and results **68**

4.4. Discussion **69**

4.5. Conclusion **71**

CHAPTER 5 ..... **72**

5. Study 3: Hierarchies of control in Prefrontal Cortex **72**

5.1. Theoretical background **72**

5.2. Methods **76**

5.2.1. Participants 76

5.2.2. Stimuli and experimental procedure 76

5.2.3. Image acquisition 80

5.3. Data analyses **81**

5.3.1. Pre-processing 81

5.3.1. Multivariate Pattern Analysis 81

5.4. Results **84**

5.4.1. Behavioural results 84

5.4.2. Decoding representations of high- and low-level rules 86

5.4.3. Representations during rule integration 88

5.4.4. Comparison with results from previous studies 89

5.5. Discussion **90**

5.6. Conclusion **94**

CHAPTER 6 ..... **96**

6. Study 4: Neural basis of propositional connectives **96**

6.1. Theoretical background **96**

6.2. Methods **101**

6.2.1. Participants 101

6.2.2. Stimuli and experimental procedure 101

6.2.3. Image acquisition 104

6.3. Data analyses **105**

6.3.1. Pre-processing 105

6.3.2. Univariate analyses 105

6.3.3. Multivariate Pattern Analysis 106

6.3.4. ROI analyses 107

6.4. Results **108**

6.4.1. Behavioural results 108

6.4.2. Decoding representations of rules with logical connectives 110

6.4.3. Rule processing 112

6.5. Discussion **115**

6.6. Conclusion **121**

CONCLUSION **123**

ACKNOWLEDGEMENTS **126**

REFERENCES **129**

APPENDICES ..... **I**

A. Supplemental materials for study 1 ..... **I**

A1. Debriefing questionnaire I

B. Supplemental materials for study 3 ..... **III**

B1. Post-experiment questionnaire III

C. Supplemental materials for study 4 ..... **IV**

C1. Pre-screening questionnaire IV

C2. Post-experiment questionnaire XI

ABSTRACT

Much of our behaviour is guided by rules defining associations between meaningful stimuli and proper responses. The ability to flexibly switch between rules to adapt to a continuously changing environment is one of the main challenges for the human cognitive system. Investigating how different types and combinations of rules are encoded and implemented in human brain is crucial to understand how we select and apply rules to guide our behaviour and react flexibly to a dynamic environment. The present thesis addressed the issue of *where* in the brain different types of rules are represented and *how* they are processed. Behavioural paradigms, functional magnetic resonance imaging, and multivariate pattern classification were combined to shed light on the cognitive mechanisms underlying rule processing and to identify brain areas encoding the contents of such processes. Using a priming paradigm, the first study assessed which types of associations (conditional, disjunctive, spatial, or quantified) could be activated automatically and trigger unconscious inferences. It proved that Modus Ponens inference is carried out unconsciously. The second study demonstrated that a condition-action rule instructed on a trial-by-trial basis and immediately marked as irrelevant causes significant interference effects when involuntarily triggered by target stimuli matching the condition in the rule. In the third study, using complex rule sets, we showed that rules at different level in the hierarchy of action control are encoded in partially separate brain networks. Moreover, we found that rule information is represented in distinct brain areas when different types of rules are encoded jointly. In the fourth study, we used rules composed using different logical connectives to expand the set of associations considered and to assess possible differences in rule representation and processing between rules with distinct logical forms. We found that separate brain areas encoded task rule information during rule representation and evaluation and that the involvement of these areas depended on the specific rule active in a trial. Taken together, our results suggest that conditional rules hold a special status in the human cognitive system, contributing to our knowledge on rule-guided behaviour.

Keywords: rule representation, executive control, deduction, task set, Prefrontal Cortex

ABBREVIATIONS

ACC	Anterior Cingulate Cortex
AR	Autoregressive
BA	Brodmann Area
BOLD	Blood-Oxygenation-Level-Dependent
BR	Basic Rule
CV	Cross-Validation
DA	Decoding Accuracy
dHb	Deoxygenated Haemoglobin
DLPFC	Dorsolateral Prefrontal Cortex
EEG	Electroencephalography
EHI	Edinburgh Handedness Inventory
EIN	Excitation-Inhibition Network
EPI	Echo-Planar Imaging
FA	Flip Angle
FFX	Fixed-Effects Analyses
FIR	Finite Impulse Response
FMG	Frontal Middle Gyrus
fMRI	Functional Magnetic Resonance Imaging
FOV	Field Of View
FPC	Frontopolar Cortex
FWE	Family-Wise Error
GLM	General Linear Model
GRE	Gradient-Echo
Hb	Oxygenated Haemoglobin
HR	Hierarchical Rule

HRF	Haemodynamic Response Function
IFG	Inferior Frontal Gyrus
IFJ	Inferior Frontal Junction
IPL	Inferior Parietal Lobule
IPS	Inferior Parietal Sulcus
ISI	Inter-Stimulus Interval
ITG	Inferior Temporal Gyrus
ITI	Inter-Trial Interval
LFP	Local Field Potential
MFG	Middle Frontal Gyrus
MNI	Montreal Neurological Institute
MOG	Middle Occipital Gyrus
mPC	Medial Parietal Cortex
MRI	Magnetic Resonance Imaging
MTG	Middle Temporal Gyrus
MVPA	Multi-Voxel Pattern Analysis
OFC	Orbitofrontal Cortex
PFC	Prefrontal Cortex
PG	Precentral Gyrus
PM	Premotor
PMd	Premotor dorsal
PPC	Posterior Parietal Cortex
Pre-SMA	Pre-Supplementary Motor Area
RF	Radiofrequency
RFX	Random-Effects Analyses
ROC	Receiver Operating Characteristic
ROI	Region-Of-Interest
RT	Reaction Time

S-R	Stimulus-Response
SD	Standard Deviation
SMA	Supplementary Motor Area
SNR	Signal-to-Noise Ratio
SPL	Superior Parietal Sulcus
SVM	Support Vector Machine
TE	Echo Time
TR	Repetition Time
VLPFC	Ventrolateral Prefrontal Cortex
VMPFC	Ventromedial Prefrontal Cortex
WM	Working Memory

INTRODUCTION

> *"Ein jedes Ding der Natur wirkt nach Gesetzen. Nur ein vernünftiges Wesen hat das Vermögen, nach der Vorstellung der Gesetze, d. i. nach Prinzipien, zu handeln, oder einen Willen. Da zur Ableitung der Handlungen von Gesetzen Vernunft erfordert wird, so ist der Wille nichts anders als praktische Vernunft."*
>
> *(Everything in nature works in accordance with laws. Only a rational being has the power to act in accordance with his idea of laws – that is, in accordance with principles – and only so has he a will. Since reason is required in order to derive actions from laws, the will is nothing but practical reason.)*
>
> *(Immanuel Kant (1785), Grundlegung zur Metaphysik der Sitten, p. 80)*

One of the main challenges for the human cognitive system is to pursue and achieve its own goals in a continuously changing environment. In this context, the ability to compactly and effectively represent the variables and the relations involved in managing a specific situation is fundamental. This ability allows for switching flexibly between rules that link sensory stimuli to appropriate responses (Fuster, 2000; Miller & Cohen, 2001). In fact, associations between "meaningful stimuli […] and other stimuli, potential responses, heuristics for responding, or rewards" are usually used "to select an appropriate course of action for a given situation in which we find ourselves" (Bunge & Wallis, 2008, p. xiii). Thus, many of our actions rely on rules describing such associations. Rules that govern behaviour can be very different in nature and complexity and can be relevant for the most diverse situations we face in everyday life: we can apply simple rules linking a well-defined stimulus with a motor response or need to apply multiple rules with different levels of priority to achieve a goal; we can be aware of following a rule or we can act according to it even without noticing to do so; we can learn a rule explicitly or acquire it implicitly. Investigating how different types of rules or combinations of instructions of variable complexity are encoded and implemented in human brain is crucial to understand how we select and apply rules to guide our behaviour and react flexibly to a changing environment.

The present thesis deals with this specific issue with a twofold **aim**: first, to identify *where* in the brain different types of rules are represented (i.e., where information about rules is encoded in the brain) and, second, to investigate *how* these rules are processed (i.e., the cognitive mechanisms underlying rule retrieval, maintenance, evaluation, and implementation). Each of the studies presented in the following chapters focuses either on a specific feature of the rules (e.g., the specific type of association or the level of complexity of the rule set) or on a particular processing stage involved in rule usage (e.g., maintenance, evaluation, or implementation).

The first chapter offers an overview of the **general framework** of the projects illustrated in the present thesis and describes briefly the main currents of research and results emerged in the field. The chapter refers mainly to neuroscientific studies on rule representation and processing, since the research project focused on these two aspects of rule-guided behaviour. Despite being part of a single research line, each of the four projects deals with a separate issue, with its own literature and requiring interpretations tailored for the specific context of reference. Therefore, each study will be presented and discussed in a separate chapter, including a section describing in detail the theoretical background more relevant for the particular issue addressed.

An introduction to functional magnetic resonance imaging (fMRI) and related analysis **methods** is also given in a separate chapter (Chapter 2) for people who are not familiar with the research techniques used in the two fMRI experiments presented in this work. Specifically, the second chapter provides an overview of the basic principles of magnetic resonance imaging (MRI) and fMRI and describes the two main approaches to fMRI data analysis: the classical mass-univariate analysis and the more advanced multivariate methods. Hopefully, this will facilitate the comprehension of the analyses and the results described in the last two chapters.

The third chapter describes and argues a set of behavioural experiments investigating which types of rules can be applied **automatically** (i.e., without voluntary control). To this purpose, we tested whether, for any rule, subliminally presented information could trigger a response (i.e., whether an unconscious stimulus could trigger an inference in different deductive problems).

The fourth chapter presents and discusses a behavioural experiment testing whether representations of rules temporarily encoded in working memory (WM) result

in interference effects. Specifically, we tested whether an unpractised and irrelevant condition-action rule (i.e., an association instructed on a trial-by-trial basis, thus not encoded in long-term memory, and immediately marked as irrelevant) merely encoded in WM could be involuntary triggered by a stimulus matching the condition in the rule and thus cause **interference** when suggesting a wrong response.

The fifth chapter illustrates and debates an fMRI experiment addressing the question of whether rules at different levels in the **hierarchy** of action control are represented in distinct brain areas, in line with the main theories on the functional role of prefrontal cortex (PFC) in cognitive control (see Paragraph 1.1). It also explores rule representation when rules at different levels of control are encoded at the same time.

The sixth chapter describes and discusses an fMRI experiment that further extends the set of rules considered, investigating the encoding and evaluation of rules with different logical forms. We directly compared rules composed using distinct propositional **connectives** in order to identify possible differences in the representation and evaluation of these rules.

Finally, in the last section, the results of all studies are put together to illustrate the **contribution** of this work to our understanding of the mechanisms underlying rule representation and processing that play a role in the way we interpret and react to different stimuli in our dynamic environment.

CHAPTER 1

1. Theoretical framework

1.1. Rule-guided behaviour

As outlined in the introduction, much of our behaviour relies on rules describing associations between relevant stimuli and their appropriate responses (Bunge & Wallis, 2008). In order to adapt to a dynamic environment, we often need to select relevant associations depending on the specific context or to establish relations between multiple rules assigning different levels of priority to them. **Cognitive control** subserves such flexible behaviours by selecting actions that are coherent with our internal goals and appropriate to the demands of our environment. We can define cognitive control of action as "the ability to flexibly adapt behavior to current demands by promoting task-relevant information in the face of interference or competition." (Dreher & Berman, 2002, p. 14595) To shed light on the cognitive processes underlying the ability to flexibly react to a changing environment, it is crucial to investigate how humans represent and implement the rules that guide their courses of action.

The present thesis deals with the question of how different types of such rules or combinations of instructions of variable complexity are encoded, evaluated, and processed in human brain to carry out complex tasks. Clearly, this is a very broad question since the associations relevant for action are of diverse nature and complexity, can rely on different kind of relationships, and their use involves several processes (e.g., learning, storage, retrieval, representation, maintenance, evaluation, and application) that have to be taken into account. We can act according to simple stimulus-response (S-R) associations (e.g., if the phone rings you will answer), apply rules with response contingencies (e.g., you probably will not answer the phone if you are working and don't want to be interrupted), rely on sets of multiple rules that have to be applied in a specific order or depending on the context (e.g., the instructions to follow to set an alarm on your smartphone may not work on a different one), respond in a fixed way (e.g., you will lift the handset if the telephone at home rings) or select one of many possible

responses (e.g., if your friend texts you, you may choose to text him back, call him, send him an e-mail, or ignore him), react immediately to the stimulus (e.g., if you are waiting for a very important call you will answer promptly) or have to postpone the response (e.g., if the date on a presentation reminds you it's your mother's birthday you won't call her immediately but will wait until the conference is over to do so). However, all these situations share the same underlying chain of events that can be summarised as follows: relations are established between relevant variables in our environment and possible responses; then, when a meaningful stimulus is perceived, the relevant rules are retrieved and maintained in WM to select the appropriate response; finally, the chosen action is performed. Since many different associations may be defined for the same stimulus (e.g., it requires dissimilar responses in different contexts), in order to select the appropriate behaviour for a given situation it is crucial to represent rule information at a more abstract level. Such abstract rule knowledge is referred to as "*task set*", defined as the "configuration of perceptual, attentional, mnemonic, and motor processes necessary to perform the task" (Sakai, 2008, p. 220). Therefore, **rule-guided behaviour** can be investigated at different levels (e.g., considering concrete representations or more abstract rule information) and focusing on distinct aspects of the process (e.g., stimulus perception and encoding, rule retrieval and representation, or response selection and execution). This work presents four projects using rules varying either in quality or level of abstraction and task sets of different complexity to investigate the last two processing stages: rule representation and processing. At the most concrete level, simple S-R associations may be retrieved automatically whenever the proper stimulus is presented, even without intention (e.g., Braine & O'Brien, 1998). In study 1, we considered several types of associations to assess which one is automatic. We tested whether, for any association, an unconscious stimulus could nevertheless trigger a response. When such automatic associations are activated by a stimulus but other rules are relevant for the task to be performed, conflict in response selection may occur (Crone, Donohue, Honomichl, Wendelken, & Bunge, 2006; Gade & Koch, 2007; Liston et al., 2006; Stroop, 1935). In this context, an interesting question would be whether associations that are instructed on a trial-by-trial basis (thus not retrieved from long-term memory) and immediately marked as irrelevant before a matching stimulus is presented, may nonetheless cause conflict. Study 2 addressed this question. However, in

many situations it is not sufficient to perform an action in response to a stimulus; instead, it is necessary to integrate multiple rules at different levels of cognitive control to choose the appropriate course of action (Reverberi, Görgen, & Haynes, 2012b). In study 3 we carried out an fMRI experiment involving more complex task sets to investigate whether rules at different levels in the hierarchy of action control are represented in separate brain areas, as some recent theories hypothesised (Badre & D'Esposito, 2009; Christoff & Keramatian, 2007; Kouneiher, Charron, & Koechlin, 2009). It also explored rule representation when different types of rules were encoded at the same time (Nee & Brown, 2012b). To avoid possible confounds, we used rules with the same logical form: conditional rules describing simple S-R associations. However, further types of rules can be used to assess whether qualitatively dissimilar associations are represented in distinct brain areas or using a different neural code (e.g., compositional *vs.* independent coding, Reverberi, Görgen, & Haynes, 2012a). Recent studies found that reasoning with different types of arguments (i.e., problems in which different associations/relations have to be evaluated) activates distinct brain areas (for a meta-analysis see Prado, Chadha, & Booth, 2011). Moreover, it has also been shown that distinct logical forms are processed in different ways; for example, associations described by conditionals, but not by disjunctions, are activated automatically (Reverberi, Pischedda, Burigo, & Cherubini, 2012). Hence, if the evaluation of different relations activates separate regions and distinct logical forms are processed in a dissimilar way, it is reasonable to hypothesise that rules describing different associations will be also represented in separate brain areas. In study 4, we directly compared rules composed using various logical connectives in order to identify possible dissimilarities in the representation and evaluation of qualitatively different associations.

Several neuroscientific studies investigated rule-guided behaviour (Bengtsson, Haynes, Sakai, Buckley, & Passingham, 2009; Bode & Haynes, 2009; Bunge, Kahn, Wallis, Miller, & Wagner, 2003; Cavina-Pratesi et al., 2006; Petrides & Baddeley, 1996; Reverberi, Görgen, et al., 2012a, 2012b; Reverberi et al., 2007; Sakai & Passingham, 2003, 2006; Shallice, Burgess, & Robertson, 1996; Woolgar, Hampshire, Thompson, & Duncan, 2011). The following paragraphs summarise part of the findings from research

on rule representation and processing, since the studies previously outlined focused on this two aspects of rule-based behaviour.

1.2. Rule representation

The most viable way to experimentally investigate the cognitive mechanisms underlying rule-guided behaviour is to pinpoint the **neural representation** of rules governing action. Several studies, either on primates or humans, tackled this issue, using a wide variety of rules, focusing on distinct features of a task set, or investigating different aspects of the process. For example, either *conditional associative responses* (i.e., associations learned explicitly, e.g., "if stimulus A then response X, but if stimulus B then response Y"; Petrides, 1985, 1990, 1997), S-R associations (Asaad, Rainer, & Miller, 2000; Bussey, Wise, & Murray, 2002; Cavina-Pratesi et al., 2006; Murray, Bussey, & Wise, 2000; Passingham, Toni, & Rushworth, 2000), rules with *response contingencies* (i.e., responses depending on a particular condition; Bunge et al., 2003; Bunge & Zelazo, 2006; Quintana & Fuster, 1999), abstract rules (Asaad et al., 2000; Braver & Bongiolatti, 2002; Bunge et al., 2003; Bussey et al., 2002; Christoff & Gabrieli, 2000; Christoff, Keramatian, Gordon, Smith, & Mädler, 2009; Christoff, Ream, Geddes, & Gabrieli, 2003; Hoshi, Shima, & Tanji, 2000; Koechlin, Basso, Pietrini, Panzer, & Grafman, 1999; MacDonald, Cohen, Stenger, & Carter, 2000; Muhammad, Wallis, & Miller, 2006; O'Reilly, Noelle, Braver, & Cohen, 2002; Sakai & Passingham, 2003; Wallis, Anderson, & Miller, 2001), low- (Badre & D'Esposito, 2007; Crittenden & Duncan, 2012; Dias, Robbins, & Roberts, 1996; Koechlin, Ody, & Kouneiher, 2003; Nee & Brown, 2012a, 2012b; O'Reilly et al., 2002; Pasupathy & Miller, 2005; Reverberi, Görgen, et al., 2012a; Reynolds, O'Reilly, Cohen, & Braver, 2012) and high-level rules (Badre & D'Esposito, 2007; Bunge et al., 2003; Koechlin et al., 2003; Nee & Brown, 2012a, 2012b; Reynolds et al., 2012; Sakai & Passingham, 2003, 2006; Wise, Murray, & Gerfen, 1996), well-learned rules (Buckley et al., 2009; Donohue, Wendelken, Crone, & Bunge, 2005; Hadj-Bouziane & Boussaoud, 2003), or *compound rules* (i.e., rules composed of two or more single rules, e.g., rules in form of if-and-if-then statements such as "if there is a face press left and if there is a house press right"; Bunge et al., 2003; Bunge & Zelazo, 2006; Cavina-Pratesi et al., 2006; Hampshire, Thompson, Duncan, & Owen, 2010; Reverberi, Görgen, et al.,

2012a) have been used; either cue representation (Brass & von Cramon, 2002, 2004; Constantinidis, Franowicz, & Goldman-Rakic, 2001; Funahashi & Takeda, 2002; Hoshi et al., 2000; Reverberi, Görgen, et al., 2012a; Woolgar, Thompson, Bor, & Duncan, 2011), rule encoding (Bode & Haynes, 2009; Brass & von Cramon, 2004; Bunge et al., 2003; Dosenbach et al., 2006; Nee & Brown, 2012a, 2012b; Reverberi, Görgen, et al., 2012a, 2012b; Sakai & Passingham, 2003, 2006), or response planning (Funahashi & Takeda, 2002; Hoshi et al., 2000; Quintana & Fuster, 1999; Rainer, Rao, & Miller, 1999; Tanji & Hoshi, 2001) has been considered; either rule retrieval (Brass & von Cramon, 2002, 2004; Brass et al., 2003; Bunge et al., 2003; Bunge, 2004; Donohue et al., 2005; Momennejad & Haynes, 2012; Petrides, 2002; Sakai & Passingham, 2004) or maintenance (Bode & Haynes, 2009; Brass & von Cramon, 2002, 2004; Bunge et al., 2003; Bunge, 2004; Constantinidis et al., 2001; Hoshi et al., 2000; MacDonald et al., 2000; Momennejad & Haynes, 2012; Nee & Brown, 2012b, 2012a; Reverberi, Görgen, et al., 2012a, 2012b; Rowe et al., 2007; Schumacher, Cole, & D'Esposito, 2007; Toni, Schluter, Josephs, Friston, & Passingham, 1999; Wallis et al., 2001) has been investigated.

Studies on the neural correlates of rule representation have focused mainly on **lateral PFC**, a region including dorsolateral PFC (DLPFC, Brodmann Area [BA] 9/46) and ventrolateral PFC (VLPFC, BA 44/45/47) (Bunge, 2004). Recent research provided evidence that parietal cortex also plays an important role in rule representation (Amiez, Kostopoulos, Champod, & Petrides, 2006; Bode & Haynes, 2009; Brass, Ullsperger, Knoesche, von Cramon, & Phillips, 2005; Brass & von Cramon, 2002, 2004; Bunge, 2004; Corbetta & Shulman, 2002; Crone et al., 2006; Dove, Pollmann, Schubert, Wiggins, & von Cramon, 2000; Johansen-Berg et al., 2002; Kimberg, Aguirre, & D'Esposito, 2000; Reverberi, Görgen, et al., 2012a; Rushworth, Johansen-Berg, Göbel, & Devlin, 2003; Sohn, Ursu, Anderson, Stenger, & Carter, 2000). Specifically, it has been shown that the prefrontal brain regions mainly related to rule representation are: premotor dorsal (PMd) area (BA 6), pre-supplementary motor area (pre-SMA, BA 6), anterior cingulate cortex (ACC, BA 32), VLPF, DLPFC, and frontopolar cortex (FPC, BA 10). Parietal areas involved in the encoding of rules are: intraparietal sulcus (IPS, BA 40) and inferior parietal lobule (IPL, BA 40), which altogether constitute the posterior parietal cortex (PPC, BA 40). While VLPFC and DLPFC seem to be differentially activated depending on

the specific task performed (e.g., Sakai & Passingham, 2003, 2006; Yeung, Nystrom, Aronson, & Cohen, 2006) or on the complexity or the level of abstraction of the rule that is represented (e.g., Badre, 2008; Bunge et al., 2003; Koechlin & Summerfield, 2007), FPC activation is task-independent (Sakai & Passingham, 2003). Parietal cortex seems to be involved in S-R association storage and response selection (e.g., Brass & von Cramon, 2002, 2004; Bunge, 2004; Crone et al., 2006) or in translating visual cues into rules (Reverberi, Görgen, et al., 2012a), while PMd area was associated with the representation of visuomotor response rules (Brass, Derrfuss, & von Cramon, 2008; Tanji & Hoshi, 2001). Instead, ACC has been implicated in representing the values of reinforcement associated with each action rather than rules linking stimuli to actions (Rushworth, Croxson, Buckley, & Walton, 2008). Finally, pre-SMA has been shown to be particularly active when switching between sets of rules (Brass & von Cramon, 2002; Crone et al., 2006; Garavan, Ross, Kaufman, & Stein, 2003; Koechlin et al., 2003; Nachev, Wydell, O'Neill, Husain, & Kennard, 2007; Rushworth, Hadland, Paus, & Sipila, 2002; Rushworth, Walton, Kennerley, & Bannerman, 2004) or in action sequencing (Kennerley, Sakai, & Rushworth, 2004; Nakamura, Sakai, & Hikosaka, 1998; Shima, Mushiake, & Tanji, 1996).

Recently, growing interest has been devoted to pinpoint the organisation of rule representations in PFC. Several theories have been proposed hypothesising the existence of a **gradient** of control along the anterior-to-posterior axis of PFC in which abstract rules (or distal control signals) are represented in rostral areas while concrete rules (or proximal signals) are controlled by more caudal regions (Badre & D'Esposito, 2007; Badre, 2008; Christoff et al., 2009; Frank & Badre, 2012; Koechlin et al., 2003; Koechlin & Summerfield, 2007; Petrides, 2005). However, there are also positions against the existence of such a gradient (Duncan, 2001) or suggesting that the recruitment of more anterior areas in PFC depends rather on either task difficulty (Crittenden & Duncan, 2012) or task-dependent maintenance demands (Reynolds et al., 2012).

1.3. Rule processing

Recently, **task switching** has become the elective method to unravel the neural underpinnings of cognitive control and rule-guided behaviour (Monsell, 2003). In fact, "task-switching paradigms are a favorite choice for studying how humans represent and apply rules." (Stoet & Snyder, 2008, p. 227). In these paradigms, participants have to switch frequently between two tasks, each one governed by its own rules. To perform well at the task, it is necessary to use relevant information (i.e., which rules apply in a specific trial) to react properly to stimuli: in other words, to adapt action to the context. The main finding of task-switching literature is that switching between tasks is associated with longer reaction times (RTs) and higher error rates compared with repeating the same task, thus entailing a "*switch cost*" (e.g., Allport, Styles, & Hsieh, 1994; Jersild, 1927; Meiran, 2000; Rogers & Monsell, 1995). It has been argued that such switch costs are due to at least two processes: task preparation and control processes involved in task execution (e.g., Brass & von Cramon, 2002, 2004; Meiran, 1996; Rogers & Monsell, 1995). The previous paragraph (Paragraph 1.2) summarised the results related to the first aspect; the remaining part of this paragraph, instead, will focus on research on task execution.

Studies investigating the neural mechanisms underlying rule **processing** have focused primarily on PFC. This area plays a crucial role in flexibly adapting behaviour to rapid changes in the environment, by selecting appropriate actions according to external stimuli and internal goals (e.g., Miller & Cohen, 2001; Petrides & Baddeley, 1996; Petrides, 2005; Shallice et al., 1996). Distinct sub-regions within **PFC** have been associated with different aspects of rule application and task execution. For example, VLPFC seems to be involved in action selection based on conditional associations and in monitoring the rule outcome (e.g., Aron, Robbins, & Poldrack, 2004; Brass & von Cramon, 2002; Hoshi, Shima, & Tanji, 1998; Hoshi et al., 2000; Miller, 2000; Rushworth et al., 2002; Walton, Devlin, & Rushworth, 2004), especially when high-level rules are applied or information integration is required (e.g., Badre & D'Esposito, 2007; Bunge et al., 2003; Crone et al., 2006; Diamond, 2006; Dreher & Berman, 2002; Koechlin et al., 2003; Tanji & Hoshi, 2001). In studies on deduction, it was also activated when evaluating conditional rules (Monti, Osherson, Martinez, & Parsons, 2007; Noveck, Goel, & Smith, 2004; Prado, Van Der Henst, & Noveck, 2010; Reverberi et al., 2012). Moreover,

a region in posterior VLPFC, the inferior frontal junction (IFJ, BA 8), was associated with the updating of task representations depending on contextual information and with the integration of information overtime (e.g., Brass & von Cramon, 2004; Nee & Brown, 2012b). Another region in lateral PFC, DLPFC, seems to play a role in the selection of responses based on task rules and goals, especially when a strong response tendency needs to be overridden (e.g., Bunge, 2004; MacDonald et al., 2000), abstract rules or strategies are applied (e.g., Bunge & Zelazo, 2006; Bunge, 2004; Crone et al., 2006; Genovesio, Brasted, Mitz, & Wise, 2005; Wallis et al., 2001; Wallis & Miller, 2003), or the contextual information is uncertain (e.g., Badre, Doll, Long, & Frank, 2012; Koechlin & Summerfield, 2007; Nee & Brown, 2012b). The most rostral regions in PFC, such as FPC and orbitofrontal cortex (OFC, BA 10/11) have been implicated in managing the execution of multiple tasks (e.g., Badre & D'Esposito, 2009; Boorman, Behrens, Woolrich, & Rushworth, 2009; Braver & Bongiolatti, 2002; Koechlin & Hyafil, 2007) and in evaluating reward-related information (e.g., Charron & Koechlin, 2010; Koechlin & Hyafil, 2007; O'Doherty, Deichmann, & Dolan, 2003; Rushworth, Noonan, Boorman, Walton, & Behrens, 2011). Finally, ACC appears to be activated when response conflict occurs (e.g., Botvinick, Cohen, & Carter, 2004; Kerns et al., 2004; Rushworth et al., 2008; Sohn, Albert, Jung, Carter, & Anderson, 2007), when outcome information is used to guide action (e.g., Walton et al., 2004), or when information at different levels is updated (e.g., Nee & Brown, 2012a).

This paragraph and the previous one offered a summary of the most relevant results emerged from research on rule representation and processing. However, the list set forth is far from being exhaustive, since only findings potentially relevant for the issues addressed by the present thesis were considered. The general picture briefly outlined in this chapter will serve as a reference framework for evaluating the results of each study described in this work.

CHAPTER 2

2. Methods

In the last two studies described in the present thesis we used *functional magnetic resonance imaging* to measure brain activity during rule representation, processing, and application. This chapter gives an overview of the basic principles of MRI (Paragraph 2.1) and fMRI (Paragraph 2.2) and describes the two main approaches to fMRI data analysis: the mass-univariate analysis (Paragraph 2.3) and the more recent and more sensitive multivariate technique (Paragraph 2.4), since we used both of them to analyse the recorded data.

For the first two behavioural studies presented in this thesis, instead, more specific details about the methods will be given separately for each experiment in the respective chapter.

2.1. Physical principles of MRI

This paragraph describes MRI and its basics. The information reported refers mainly to two textbooks: Schild (1990) and Huettel, Song, and McCarthy (2009).

MRI is a method used for measuring brain activity in a *non-invasive* way. It exploits the magnetic properties of the hydrogen nuclei (protons) present in water, the main component of human body. These nuclei (nuclear spins) behave like tiny rotating magnets that can be aligned along a strong external magnetic field (B_0) in an MRI scanner. Protons in a magnetic field precess at a characteristic frequency that depends on the magnetic field strength: the *Larmor frequency*. This relation can be formalised as

$$\omega_0 = B_0 \gamma$$

where ω_0 [MHz] is the Larmor-frequency, B_0 [T] is the magnetic field strength, and γ [MHz/T] is the gyromagnetic ratio, which is specific for each atomic nucleus.

The protons can align parallel to the magnetic field (along the *z*-axis) or antiparallel. It requires less energy to align parallel to the magnetic field, therefore this form is preferred over the antiparallel one; as a consequence, there are more protons aligned

parallel to the magnetic field, even if the difference is slight (seven more every ten millions protons; Schild, 1990). This leads to a net magnetisation along the external magnetic field (*longitudinal magnetisation*), which however cannot be measured since it has the same direction as the external magnetic field. Therefore, a *radiofrequency* (RF) *pulse* with a frequency ω_0 is applied to the spins to disturb the alignment of the protons; this process is known as **excitation**. Only when the RF pulse frequency matches the frequency of the protons, the latter can pick up energy from the wave (resonance). Due to protons' resonance, the RF pulse has two effects: 1) it tips the magnetisation vector from the horizontal (parallel to B_0) to the transversal plane and 2) it synchronises the precession of the protons (i.e., they rotate in phase). This results in a magnetisation vector rotating along the transversal plane (*transversal magnetisation*). This vector represents a signal that can be recorded by a receiver coil placed around the head (or any part of the body you want to measure). The angle to which the application of an RF pulse tips the longitudinal magnetisation relative to the B_0 direction is called *flip angle* (FA).

After the RF pulse, some protons release their energy to the surround (lattice) and align back parallel to the magnetic field, leading to the longitudinal magnetisation re-emerging. This process is called longitudinal (or spin-lattice) **relaxation** and it is described by the time-constant T_1 (the time it takes for the longitudinal magnetisation to recover to the 67% of the original value). At the same time protons start precessing out of phase, due to random fields caused by other spins in the surround, therefore the transversal magnetisation vanishes. This process is called spin-spin relaxation and is defined by the time-constant T_2 (the time it takes for the transversal magnetisation to decay by 37% of the maximum value). However, relaxation time is faster than predicted by T_2 alone due to field inhomogeneities. The time constant T_2* takes into account also this factor. Distinct biological tissues are characterised by different structural properties that influence the duration of relaxation processes (Damadian, 1971). Therefore, different tissue properties can be imaged based on different time constants.

In order to produce a 3D **image** of the whole brain (i.e., volume or scan) it is essential to record signals from every/many different positions in the brain (voxels). A *voxel* (volumetric picture element) is the 3D analog of a pixel. To resolve spatial information it is necessary to apply different *magnetic gradients* (i.e., magnetic fields

that vary linearly across space); these cause the atoms to precess at different frequencies (Lauterbur, 1973), allowing to identify different positions in space. During the first step of the signal-acquisition process, a field gradient parallel to B_0 (along the z-axis, G_z in Figure 2.1) is applied so that protons in different positions have distinct precession frequencies. To *select* a particular *slice* within the whole imaging volume, the RF pulse has the same frequency as the Larmor frequency of the slice to be selected. In a second step, it is possible to determine the exact spatial position using a two-dimensional spatial encoding scheme within the selected slice (Huettel et al., 2009). More specifically, a first magnetic gradient (G_y in in Figure 2.1) is applied along the y-axis before the readout time (data acquisition period) to produce phase differences (*phase encoding*); then a second gradient (G_x in in Figure 2.1) is applied along the x-axis during data acquisition to modify the frequency of the spins (*frequency encoding*).

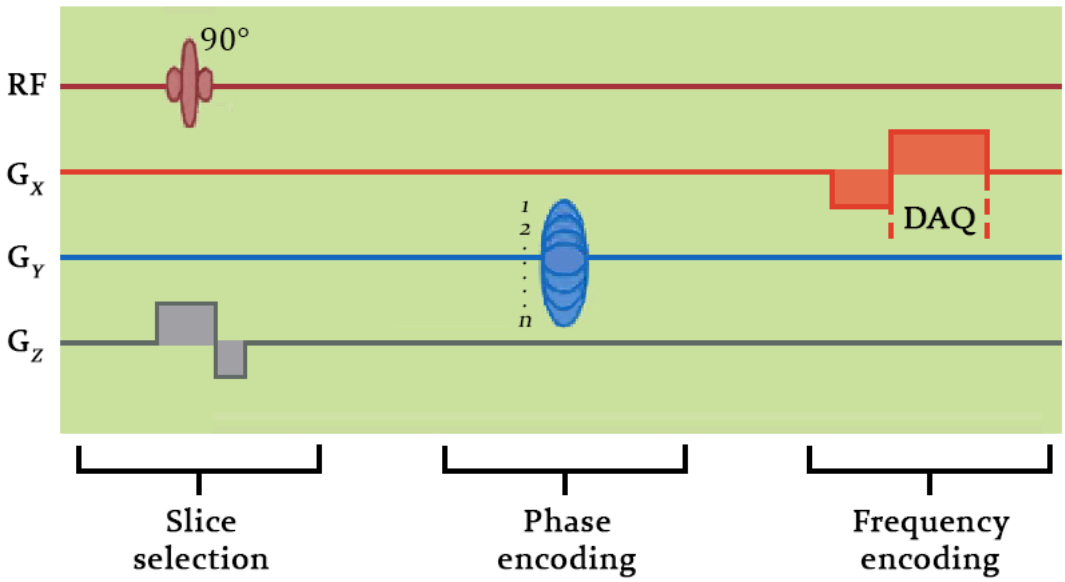


Figure 2.1 Representation of a typical gradient-echo pulse sequence. The sequence begins with an RF pulse combined with a slice selection gradient (G_z). The G_y gradient creates differences in phase between the spins, while G_x is applied during data acquisition (DAQ) to modify their frequency. Adapted from Huettel et al. (2008, p. 112).

The conventional notation schema adopted for MRI data acquisition is the *k-space* (i.e., the 2D space of a slice coded in frequency and phase). The signal information coded in the *k*-space can be converted into an image applying an inverse *Fourier transform*. Two parameters govern the acquisition of MR images: 1) *repetition time* (TR), i.e., the time interval between consecutive RF pulses, and 2) *echo time* (TE), i.e., the time interval between excitation and measurement. These parameters can be varied to obtain

different types of contrast images (T_1- or T_2-weighted, respectively). Several image sequences can be used affecting T_1- T_2-, and T_2*- weighting. Currently the most common method in fMRI (Buxton, 2002) is *echo-planar imaging* (EPI). The introduction of this technique (Mansfield, 1977) greatly improved the efficiency of data acquisition, making it possible to generate a complete 2D image at one time. In general, in EPI, a transmitter coil sends an RF pulse and then rapidly changing magnetic field gradients (sometimes in combination with additional RF pulses) are introduced to produce an **echo** (i.e., the emission of an electromagnetic resonance signal) that the receiver coil records (for details see, e.g., Moonen & Bandettini, 1999). The echo signal can be produced in different ways. *Gradient-echo* (GRE) sequences (Figure 2.1) use a bipolar frequency-encoding gradient: first a negative gradient is applied to dephase the spins, then a gradient with opposite polarity rephases the spins and generates the echo signal. In a *spin-echo* (SE) sequence, instead, the echo signal is produced by a second 180° RF pulse (refocusing pulse). All the fMRI studies discussed in this thesis used GRE EPI sequences.

2.2. Functional Magnetic Resonance Imaging

When investigating a cognitive task, brain activity has to be related to different aspects of the task. To this purpose **fMRI** can be used, because it allows for measuring neural activity indirectly by recording the *blood-oxygenation-level-dependent* (BOLD) signal (Ogawa & Lee, 1990). BOLD fMRI exploits the difference in magnetic susceptibility (i.e., degree of magnetisation of a material placed in a magnetic field) between oxygenated haemoglobin (Hb) and deoxygenated haemoglobin (dHb). Hb and dHb are diamagnetic and paramagnetic, respectively. Since paramagnetic substances produce field distortions, neighbouring protons precess at different frequencies due to the different field strengths experienced; thus, the decay of transversal magnetisation is faster (i.e., T_2* is shorter) resulting in a decreased MR signal in T_2*-weighted images (Ogawa, Lee, Nayak, & Glynn, 1990). The difference in magnetic susceptibility between the blood vessels and the enclosing tissue induced by dHb can therefore be used as an *endogenous contrast*, because it relies on the intrinsic qualities of the biological tissue.

Moreover, BOLD fMRI is an *indirect* measure of brain activity based on evidence that the BOLD signal reflects the neural response evoked by an event. Neuronal processes

require energy to be carried out; this energy is obtained from glucose metabolism, mainly via aerobic (i.e., requiring oxygen) mechanisms. Thus, when a brain region is active, there is an increase in oxygen consumption that should result in a higher level of dHb, since oxygen is provided by Hb. However, this process is paralleled by an increase in regional blood flow, producing an *oversupply* of Hb, therefore the resulting BOLD signal is enhanced (Moseley & Glover, 1995). The shape of the **haemodynamic response function** (HRF – i.e., the function describing the changes in BOLD signal following a period of neural activity) is characteristic (Figure 2.2): after an initial *dip*, probably reflecting the increased proportion of dHb due to glucose metabolism, it reaches the *peak* about 3-8s after a short-duration event, reflecting the oversupply in Hb, and then the signal decreases producing an extended *undershoot* (i.e., signal amplitude below baseline), due to changes both in blood flow and volume (Buxton, Wong, & Frank, 1998) and in oxygen metabolism (Harshbarger & Song, 2008). The latency of the peak depends on the specific process being measured and on the brain region involved (for details see, e.g., Heeger & Ress, 2002).

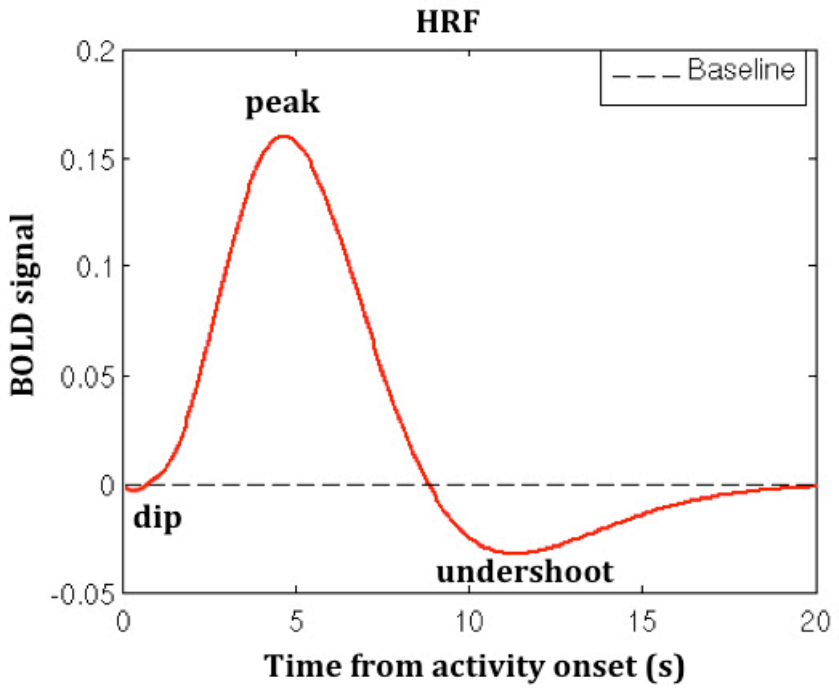


Figure 2.2 Schematic representation of the BOLD haemodynamic response and its components. Adapted from http://theclevermachine.files.wordpress.com/2012/11/fmribasics-1.png

fMRI can well reconstruct the spatial position of the signal because of the high spatial *resolution* (1-5mm). However, the temporal resolution is rather poor (in the order of seconds) compared to other neuroimaging techniques such as electroencephalography (EEG), due to the delay of the hemodynamic response.

The firing rate of neurons sensitive to a stimulus or a task condition (*action potential*) constitutes a *direct* measure of the neural response to that stimulus or condition. A relation between this measure and the fMRI activation has to be proven to consider the latter a reliable measure of neuronal activity. However, it is difficult to ascertain the exact nature of this relationship because the two processes have different dynamics. In fact, neuronal spikes occur tens of milliseconds after a sensory stimulation, while changes in BOLD signal are visible only some seconds later (Figure 2.2). Moreover, the HRF depends not only on the actual oxygen consumption by active nerve cells but also (and for a greater extent) on the increase in regional blood flow (Huettel et al., 2009), as explained above in more detail. Simultaneous fMRI and electrophysiological recordings indeed confirmed the notable correspondence between BOLD fMRI and neuronal activity (e.g., Huettel et al., 2004; Mukamel et al., 2005), measured either as *local field potentials* (LFPs – i.e., the integrated electrical activity of nearby neurons within a few millimetres) or *single-unit* activity (i.e., the firing rate of a single neuron next to the electrode tip). BOLD signal changes seem to reflect synaptic activity, measured as LFP, rather than spiking itself (Niessing et al., 2005; Viswanathan & Freeman, 2007). For example, Logothetis and colleagues measured LFPs, single-unit, and *multi-unit* (i.e., the aggregate firing rate of neurons in a few hundreds of microns) activity jointly with BOLD signal, showing that the latter is better predicted by LFPs (Logothetis, Pauls, Augath, Trinath, & Oeltermann, 2001). Since LFPs reflect not just post-synaptic potentials, but also soma-dendrite interactions, the BOLD signal may account mainly for incoming perisynaptic activity (Logothetis & Wandell, 2004). Moreover, cortical areas receive massive feedback and comprise inner interconnections among micro-units of neurons characterised by strong excitatory and inhibitory recurrence (*excitation-inhibition networks*, EIN) (Logothetis, 2008). Therefore, it is still unclear whether the BOLD signal recorded from a specific brain region actually reflects neural activity within that region rather than incoming inputs from other areas or EIN activity not directly related to the stimulus/condition.

Given all these **limitations**, BOLD fMRI cannot be used to draw conclusions about processes involving single voxels, to describe with high precision how cognitive processes exactly work, nor to accurately quantify differences between brain areas or tasks based on the amplitude of the fMRI signal (Bandettini, 2009; Logothetis, 2008). Nevertheless, BOLD fMRI is one of the best tools currently available to investigate the neural basis of human cognitive functions (Logothetis, 2008). Alternative methods present serious shortcomings as well. Hence, as long as its drawbacks are taken into account and appropriate experimental protocols are devised, fMRI can be used to formulate interesting hypotheses about brain functions.

2.3. Univariate analysis of fMRI data

This section illustrates the univariate approach to the analysis of fMRI data. More detailed descriptions of each step of the procedure are available in many textbooks (e.g., Friston, Ashburner, Kiebel, Nichols, & Penny, 2006; Poldrack, Mumford, & Nichols, 2011).

fMRI data collection produces **time series** of voxels, storing information about the intensity of the BOLD signal at each time point. In fMRI data analysis, images from different experimental conditions are compared in order to evaluate whether the differences in voxel intensity between the conditions are statistically significant. However, the change in BOLD signal in the fMRI time series is a very small proportion of the total intensity of the MR signal (e.g., from 0.5 to 3% at 1.5 T). In addition, it depends both on the neural activity that reflects the change in the net magnetisation due to the RF pulse and on uncorrelated noise. The *signal-to-noise ratio* (SNR) defines how much signal is ascribable to the experimental condition; SNR is usually low in fMRI data, since the signal due to task-related activity is very small compared to unsystematic noise inducing variability in the images. More specifically, **noise** can reflect fluctuations due to either the *thermal energy* in the sample or in the scanner (which however are highly random and unrelated to the experimental manipulations); *physiological effects*, such as head motion, changes in heart pulsation or breath rate (which likely correlate with the experimental conditions); specific characteristics of the scanner, as for *scanner drift* (i.e., changes in voxel intensity over time); *brain processes unrelated* to the task; or the effects

of learning, tiredness or the use of strategies (Huettel et al., 2009). Recent studies demonstrated that the main source of variability in fMRI experiments is physiological noise (e.g., Triantafyllou et al., 2005). SNR can be increased either by enhancing the signal of interest, for example by using higher-field scanners (Krüger, Kastrup, & Glover, 2001) or by improving the power of the experimental design, or by reducing the level of noise, for example by applying pre-processing correction algorithms.

Different **pre-processing** steps are necessary before performing fMRI analysis. Since subjects naturally move during data acquisition, it is necessary to correct for head motion to improve the spatial alignment between images. In a *motion correction* step, subsequent volumes are spatially aligned to a reference volume. Since brain size and shape are fixed, rigid-body transformations (Friston, Ashburner, et al., 1995) are applied and six motion parameters (translational and rotational movements in each of the three spatial dimensions – X, Y and Z) are computed by a realignment algorithm.

The different slices of functional brain volumes are acquired at slightly distinct time points. *Slice acquisition time correction* compensates for this time delay using temporal interpolation (usually relative to the first or the middle slice), so that it appears as if the whole brain image was acquired at a single time point.

Magnetic field inhomogeneities, instead, can cause geometric or intensity distortions. *Magnetic field mapping* can correct for geometric distortions. A field map of B_0 is created by acquiring a GRE image with two distinct echoes. Two magnitude images, one for each TE, and a phase map showing the difference in phase between these two echoes are produced. These images can be used to correct distortions in EPI images.

The resolution (i.e., the level of detail of an image) of functional volumes is often relatively low. Therefore it is difficult to precisely identify anatomical regions of activation. To improve the discriminability of anatomical boundaries, EPI images can be *co-registered* to a high-resolution structural image.

Inferences about the population can be made when brain images from different subjects are compared. Since there is high variability in brain shape and size between different subjects, it is necessary to transform the images of each subject's brain into a common space. During the *normalisation* step, individual images are warped spatially to a standard brain (see, e.g., Aguirre, 2011), such as the template of the Montreal

Neurological Institute (MNI) or the stereotactic brain atlas by Talairach and Tournoux (1988).

Depending on the purpose of the research and the analysis approach, the functional images are also spatially smoothed with Gaussian kernels. Since fMRI data are spatially correlated, *smoothing* increases the SNR, given the matched filters theorem (Rosenfeld & Kak, 1982). It also improves the validity and power of statistical tests (Friston et al., 2000) because it increases the normality of the data and reduces the false-positive rates in multiple comparisons (see below). In addition, time series can be *temporally filtered* to remove slow signal drifts (Lund, Madsen, Sidaros, Luo, & Nichols, 2006). *Pre-whitening* algorithms, instead, eliminate the portion of the signal that is predictable based on past time points to control for temporal autocorrelation (Bullmore et al., 1996). As an alternative, autocorrelation can be purposely introduced in the dataset (*pre-colouring*), so that the time series will possess known statistical properties (Woolrich, Ripley, Brady, & Smith, 2001).

After the pre-processing steps described above, the images are suitable for statistical analysis. Different experimental conditions are compared directly or contrasted with a baseline condition. In most of the studies, a **general linear model** (GLM) is calculated (Friston et al., 1994; Friston, Holmes, et al., 1995). The events grouped in different conditions (or different experimental bocks) are used as *regressors* to explain variance in the data. More specifically, onset time vectors for each condition are typically *convolved* with the HRF to generate regressors (Henson & Friston, 2006). Additional regressors, such as temporal or dispersion derivatives, can be added to shape the BOLD response in a more precise way, accounting for differences in HRF onset or shape, respectively. As an alternative, onset time vectors can be convolved with a combination of many basis functions modelling single time units instead of with a single HRF (Henson, Rugg, & Friston, 2001). *Finite impulse response* (FIR) functions use a distinct stick function to model each time unit; thus, it is possible to detect changes in the signal related to the experimental conditions, irrespective of the timing and the shape of the elicited haemodynamic response.

In the **mass-univariate approach**, the model is fitted to the time course of each voxel and parameter weights (*β-value*) are estimated for each voxel. The model is formalised as

$$Y(t) = \beta X(t) + \varepsilon(t)$$

where Y(t) indicates the observed data (i.e., the signal intensity for each voxel at each time point), X(t) is the *design matrix* with the regressors and ε(t) is the vector of the errors (residuals). The parameter weights are then combined into a *contrast* to select a specific effect of interest in order to test experimental hypotheses. The aim of **single-subject** analyses is to evaluate whether the regressors contribute to the variability observed in the voxel's time course. Generally, the statistic employed to perform this analysis is the *t-test*, calculated by the formula

$$T = \frac{c^T \hat{\beta}}{\sqrt{\mathrm{var}(c^T \hat{\beta})}} = \frac{c^T \hat{\beta}}{\sqrt{\hat{\sigma}^2 c^T \left(X^T X\right)^{-1} c}} \sim t_{N-p}$$

where $c^T\beta$ is the contrast (defined as a linear combination of parameter estimates β), σ^2 is the variance of the residuals, N is the number of scans, and p is the number of regressors. The values of the statistic calculated for each voxel in the brain can be combined into a *statistical parametric map* and displayed through a colour code. A *threshold value* (α) is set, specifying the p-value below which the results of the test will be considered as significant. Since an experimental hypothesis is tested independently for each voxel and fMRI data sets contain several thousands of voxels, a substantial issue for whole-brain data analysis is that of *multiple comparisons*. In fact, when the number of statistical tests to be performed increases, also the Type I error rate (i.e., the chance to identify a voxel as activated by an experimental condition when there is actually no difference in activation) rises (e.g., Smith, 2004). To overcome this problem, it is necessary to adopt a correction for multiple comparisons.

Among *single-voxel* thresholding methods, controlling for *family-wise error* (FWE) rate using Bonferroni correction is the most conservative. The α value is divided by the number of independent statistical tests in order to minimise Type I error rate (see, e.g., Nichols & Hayasaka, 2003). A less conservative method is controlling for *false discovery rate* (FDR), that is the expected proportion of false-positive results (for details about how this method is implemented see, e.g., Genovese, Lazar, & Nichols, 2002).

At *cluster level*, less conservative approaches to control the FWE rate can be used. In fact, neighbouring voxels are not activated independently to each other but more likely they are activated together in clusters. Worsley and colleagues applied Gaussian *random*

field theory (RFT) to estimate the number of possible independent tests, taking into account the smoothness of the experimental data (Worsley, Evans, Marrett, & Neelin, 1992). This method has to be used on images smoothed beforehand in order to fit the theory (the level of smoothness should be at least twice or three times the voxel size). An alternative method is *cluster-size thresholding*. A threshold is defined as minimum size (in voxels) based on the results of less stringent voxelwise comparisons; only clusters as large as the threshold are regarded as significant (Forman et al., 1995).

Voxelwise and cluster-level analyses are typically performed on the whole brain to identify which regions display a specific pattern of fMRI activation. As an alternative, *region-of-interest* (ROI) analyses are carried out on specific brain areas to test hypotheses about which activation profile characterises those regions. ROIs are defined according to *a priori* expectations concerning which regions are more likely to be involved in a specific task. The selection may rely on anatomical as well as functional criteria, such as selecting regions that were activated by the same task in previous studies.

Group-level analyses combine results from multiple subjects to assess commonality and steadiness of the effects within or across groups of interest. Different approaches are possible. *Fixed-effects analyses* (FFX) assume that the effect of the experimental manipulations on the BOLD signal is the same in every subject (Monti, 2011). Since FFX take into account only the within-subject variability, these methods allow inferences only about the subjects used in a specific study. *Random-effects analyses* (RFX), instead, allow the possibility of differential effects in different subjects (Penny & Holmes, 2004) and can be used to make inferences about the population from which the subjects come from. Specifically, statistical tests assess whether the magnitude of the effect (i.e., activation) is significant with respect to the between-subject variability. Usually the two effects are taken into account jointly and a *mixed-effects analysis* (MFX) results. For example, in a first-level analysis a single GLM is computed for each subject using a FFX approach. Then the single-subject parameter estimates are used in a second-level analysis to perform a group-level test that takes into account both the within- and the between-subject variability (Friston, Stephan, Lund, Morcom, & Kiebel, 2005).

2.4. Multivariate analysis

This paragraph provides an overview of the methodological principles of **multi-voxel pattern analysis** (MVPA) and a description of the different steps involved. More detailed explanations of the method and of its application to neuroimaging data can be found in the literature (e.g., Formisano, De Martino, & Valente, 2008; Mahmoudi, Takerkart, Regragui, Boussaoud, & Brovelli, 2012; Norman, Polyn, Detre, & Haxby, 2006; Pereira, Mitchell, & Botvinick, 2009).

Univariate analysis assesses, separately for each single voxel, if the intensity of the signal in that voxel is significantly different between distinct experimental conditions. This approach has been successfully used to identify brain regions involved in specific cognitive processes. One of the basic assumptions of this method is that the covariance across neighbouring voxels encodes no information on the cognitive processes of interest (Mahmoudi et al., 2012). In other words, fine-grained information in activation patterns is not taken into account or even purposely removed; for example, spatial smoothing is usually applied to increase the SNR (see Paragraph 2.3). However, *fine-grained* spatial activation *patterns* may carry relevant information about the experimental conditions (Haynes & Rees, 2006). In contrast, MVPA is considered to be sensitive to differences at a finer scale (Norman et al., 2006). It has been argued that MVPA exploits spatial activation patterns to identify unique features that allow to distinguish between different classes of stimuli or conditions (Cox & Savoy, 2003; Haxby et al., 2001; Haynes & Rees, 2005; Kamitani & Tong, 2005; Norman et al., 2006). Therefore, often fMRI data are not normalised nor smoothed to preserve as much information as possible for MVPA (Formisano et al., 2008; Haynes & Rees, 2006; Kamitani & Sawahata, 2010; but see, e.g., Op de Beeck, 2010). However, it is still unclear whether 1) this finer-scale information really contributes, 2) the combination of voxels is the only informative factor when discriminating between conditions, or 3) something else drives the classification (e.g., relations reflecting the correlation between different conditions or mental states). In any case, MVPA methods have the advantage to create optimal *spatial filters* (i.e., models used to extract spatial information in the data), which might be more sensitive to detect the signal of interest. In fact, these filters explicitly enhance the signal while, at the same time, suppress part of the signal due to uninteresting factors (Haufe et al., in press).

In general, in MVPA, an algorithm (**classifier**) is applied to capture the relationship between spatial patterns of brain activation related to different task conditions. During MVPA, fMRI data are split into two subsets: training and test sets, to avoid *overfitting* (Mitchell, 2010). One subset of data is used to *train* the classifier to distinguish between different experimental conditions (*labels*) based on specific "*features*". During this step, a weight is estimated for each feature and used to predict the labels of the data in the *test set* (i.e., to which condition each example belongs), either by *classification* (discrete variables, classes) or *regression* (continuous variables). To assess whether the classifier found differences between the two classes that are not only present in the training dataset but also are caused by the condition and thus generalise over datasets, an estimate of the generalised *decoding accuracy* (DA) onto new datasets is calculated. MVPA is carried out in three successive steps: 1) feature selection, 2) classification and 3) statistical evaluation of the accuracy of the classification. Each step will be described in more detail in the following sections. The last paragraph outlines the main applications of MVPA.

2.4.1. Feature selection and dimensionality reduction

MVPA begins with the selection of **features**. In this step, it is necessary to define *what* represents a relevant feature and *where* in the brain the information we use is located. The first matter concerns how the patterns are defined. For example, the classification could be performed on raw fMRI data (e.g., Weygandt et al., 2012). However, better results might be achieved if less noisy measures of the conditions are used, such as, for example, the average of the time points within each block (e.g., Mourão-Miranda, Reynaud, McGlone, Calvert, & Brammer, 2006) or the beta parameters estimated by the GLM (e.g., Bogler, Bode, & Haynes, 2011). The second issue pertains to choosing which voxels will be included in the analysis. MVPA can be performed on the whole brain (or on large ROIs). However, this may be problematic when analysing fine-grained patterns to discriminate between cognitive states, since the proportion of informative voxels (features) is small compared to the total number of voxels considered (De Martino et al., 2008). In fact, classifier performance deteriorates when dealing with many irrelevant features (Guyon & Elisseeff, 2003; Kohavi & John, 1997).

Additionally, we might be able to extract the signal more effectively if we consider only informative voxels (i.e., voxels containing little noise).

Univariate feature selection strategies, based on either the activation level or the discrimination ability of the voxels, may be used to reduce the number of features (e.g., Mitchell et al., 2004; Mourão-Miranda, Bokde, Born, Hampel, & Stetter, 2005). However, such methods neglect the inherent multivariate structure of fMRI data. In fact, a voxel that, taken individually, is uninformative might nevertheless carry information about a condition of interest when considered jointly with other voxels (Haufe et al., in press; Haynes & Rees, 2006). It is also possible to select subsets of voxels on the basis of their anatomic or functional similarity, using pre-defined ROIs. A special instance of ROI analysis is the *searchlight decoding approach* (Haynes & Rees, 2006; Kamitani & Tong, 2005; Kriegeskorte, Goebel, & Bandettini, 2006; Norman et al., 2006). A searchlight sphere of radius *r* is centred on each of the N voxels in the brain. MVPA is performed N times, separately on each searchlight cluster (see Figure 5.2 A). This procedure results in accuracy maps showing how well the multivariate signal in the sphere differentiates the experimental conditions (i.e., the classification accuracy for each cluster of voxels in the brain).

An alternative method to reduce the number of features is *dimensionality reduction*. Techniques such as recursive feature elimination (RFE) (e.g., Guyon, Weston, Barnhill, & Vapnik, 2002), principal component analysis (PCA) (e.g., Brouwer & Heeger, 2009), or independent component analysis (ICA) (e.g., Hyvärinen & Oja, 2000) transform the original feature space into a low-dimensional space. However, better results are not guaranteed because most of these methods ignore the conditions they are trying to predict when identifying the features (Pereira & Gordon, 2006; Pereira et al., 2009).

2.4.2. Classification

After the features have been selected, the data are split into two subsets: *training* and *test dataset*. The first is used to train a **classifier** to distinguish between different experimental conditions. Then the trained classifier is used to predict to which condition (class) the data of the test dataset belong. A popular type of classifier in fMRI data analysis are *support vector machines* (SVMs) (Hanson & Halchenko, 2008; Mourão-Miranda et al., 2005; Muller, Mika, Ratsch, Tsuda, & Scholkopf, 2001), since they perform

well in high dimensional datasets with few examples. SVMs were used for the analyses performed in the fMRI experiments described in the present thesis and are briefly outlined here. For an overview of other algorithms potentially suitable for MVPA studies refer to the reviews available in literature (e.g., Norman et al., 2006; Pereira et al., 2009; Yang, Fang, & Weng, 2012). An extensive description of the classifiers can be found, for example, in Duda, Hart, and Stork's book on pattern classification (2001), instead.

Let's consider the simplest case of classification where only two categories are present and a linear SVM is used (for non-linear SVMs see, e.g., Buhmann, 2009; Vapnik, 1998). During feature selection, fMRI time series are reduced to a set of vectors. Each vector contains BOLD values at specific time points during the experiment associated with one example from a given category. The SVM takes those vectors and estimates the combination of voxel weights (*w*) defining the *hyperplane* that optimally (i.e., with maximum margin) separates the examples belonging to category A from the examples of category B within the space (Figure 2.3 A). The *margin* can be either "hard" (no training error is allowed) or "soft" (some misclassifications are possible), reflecting a tradeoff between the SVM *performance* and its ability to *generalise* to unseen examples. This tradeoff is defined by a *regularisation factor* (C) (Figure 2.3 B). With a hard margin (C = 0), the SVM performance would be perfect but the SVM would overfit the data and thus it would hardly generalise to a different dataset; with a softer margin (e.g., C = 1), the SVM performance would be worse but the SVM would generalise better to new data. *Multi-class classification* can be reduced to multiple binary problems performing the classification individually for every pair of categories and then averaging the results of the single classifications (Allwein, Schapire, & Singer, 2001).

To obtain a more reliable measure of the classification accuracy, the process is usually repeated many times and the results are averaged. In the leave-one-out *cross-validation* (CV) procedure (Efron, 1983; Lachenbruch & Mickey, 1968), for example, the data are split into K subsets. The classifier is trained on all subsets but one; the left-out subset is used to test the classifier performance. It is important to make sure that the training and the test dataset in each cross-validation step do not overlap to avoid overfitting. The procedure is repeated K times until each of the subsets has been used as test set once. The accuracy estimates from each cross-validation step can be averaged (or otherwise combined) to obtain a single *measure of performance*.

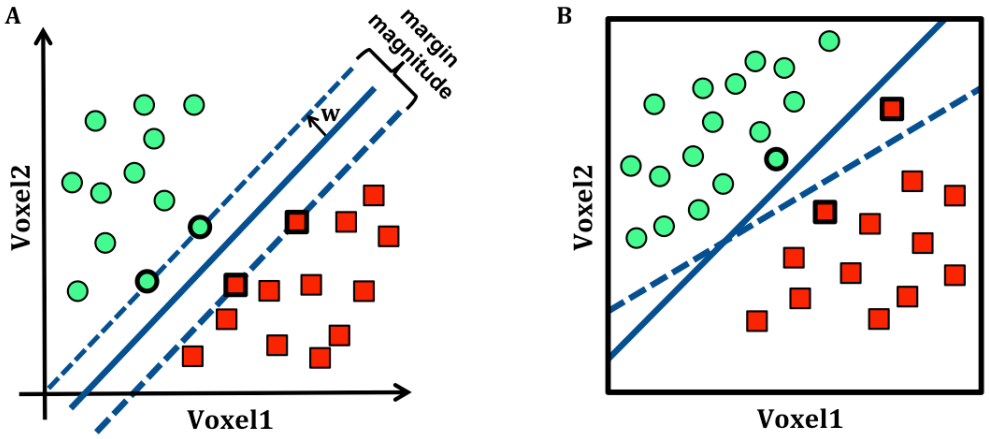


Figure 2.3 Illustration of the decision boundary of the linear SVM in the simplest case with only two voxels. Example of hard margin where no training error is allowed (A). Effect of C on the decision boundary (B). The solid line (C = 0) does not allow any training error; the dashed line (C = 1) allows some training errors (the red square between the two lines). Examples with thick borders are the support vectors (i.e., the vectors that define the margin that separates the two classes).

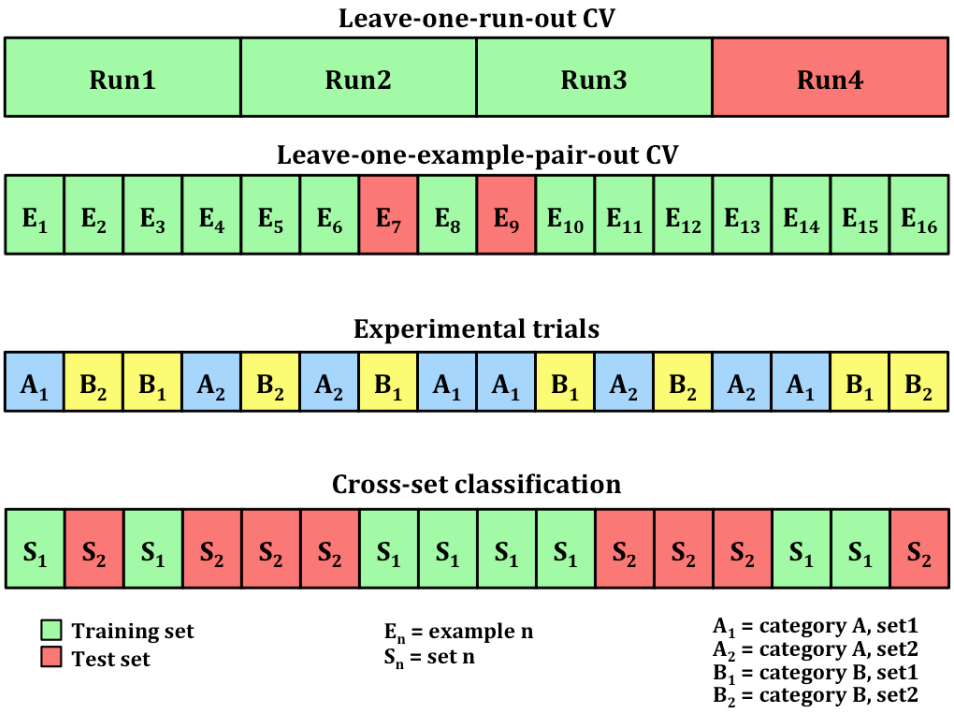


Figure 2.4 Some possible classification schemes. A classifier is trained on the training set (in green) and tested on the test set (in red) to get a measure of performance. The procedure is repeated so that each run in the leave-one-run-out CV, each example pair in the leave-one-example-pair-out CV, and each set in the cross-set classification is used as test set once. The measures obtained from each step of the procedure are averaged to get a single accuracy value.

Two of the possible schemes in single-subject MVPA are: *leave-one-run-out* CV, where data from one run are used in turn as test dataset, and *leave-one-example-pair-out* CV, where one example from each class is used as test sample in each CV step (Figure

2.4). While in the former method the training and the test set belong to different runs all containing stimuli from both classes, in the latter, training and test set contain stimuli from different classes but in both sets stimuli belonging to the same run are present. Mahmoudi and colleagues (2012) report that leave-one-example-pair-out CV produces higher performance but is computationally more demanding. However, with this method the data in the training and in the test set are not completely independent since both sets contain data belonging to the same runs, thus reflecting overlapping haemodynamic responses (Misaki, Kim, Bandettini, & Kriegeskorte, 2010). Moreover, the accuracies estimated using the two methods are conceptually different: while the leave-one-run-out CV tests the classifier's ability to generalise to new runs, the leave-one-example-pair-out CV tests its ability to generalise to new stimuli.

An alternative to CV, is *cross-set classification*. In this case, the aim of the classification is to assess whether a classifier that is trained to distinguish between category A and B (e.g., faces and houses) using one set (e.g., stimuli presented on the left of the screen) can generalise to a different set (e.g., stimuli presented on the right of the screen) instead of, for example, run. Data from one set are used to train the classifier and data from the alternative set serve as test sample and vice versa (Figure 2.4).

2.4.3. Statistics

In the last step of MVPA, the accuracies of the predictions are **statistically** tested at the group-level against the null hypothesis that the mean accuracy is at chance level (1/number of conditions). For example, the *accuracy maps* obtained for each subject using a searchlight approach are normalised to a standard space and a statistic (typically a t-test) is calculated to assess if the prediction accuracies are significantly above chance at the group level (Haynes et al., 2007). The assumptions of the t-test are not perfectly satisfied for accuracy measures (Pereira & Botvinick, 2011; Stelzer, Chen, & Turner, 2013). Mahmoudi and colleagues (2012) suggest a procedure that represents a better measure than accuracy. This method evaluates the classifier performance using *receiver operating characteristic* (ROC) curves. The ROC curves are obtained by plotting the true positive rate (TPR, i.e., the proportion of examples correctly classified as belonging to one class out of the total number of examples) against the false positive rate (FPR, i.e., the proportion of examples erroneously classified as belonging to one class among all

the examples) and each point in space represents the performance of a single classifier. The area under the ROC curve can be used to evaluate the classifier performance and represents a better measure than accuracy (Smith & Nichols, 2009). It is also possible to use *permutation tests* (Nichols & Holmes, 2002) as a parameter-free alternative to group-level t-test. The development of permutation statistics for neuroscience is an active field of research; in the following, only the very basics of the procedure will be outlined. In this class of methods, the classification performance is evaluated by comparing the actual accuracy values with a distribution of values obtained performing permutation tests. Specifically, first the labels are permuted, then the accuracy is calculated; the procedure is repeated over thousands of permutations to obtain a distribution of accuracies for the comparison. Finally a p-value threshold is defined to test the hypothesis (Golland & Fischl, 2003).

2.4.4. Applications

The main **applications** of MVPA are: pattern discrimination, pattern localisation, and pattern characterisation (Pereira et al., 2009). So far classifiers have been used mainly for pattern *discrimination*, i.e., to evaluate whether fMRI data contain information about a given experimental condition (e.g., Cox & Savoy, 2003; Haxby et al., 2001). However, they can be successfully applied also for pattern *localisation*, i.e., to identify where information about a specific cognitive process is represented in the brain. It has been argued that localisation can be achieved determining which voxels in the brain mostly contribute to the success of the classifier (e.g., Mourão-Miranda et al., 2005; Pereira et al., 2009). However, this statement is based on the "widespread misconception about multivariate classifier weight vectors […] that (the brain regions corresponding to) measurement channels with large weights are strongly related to the experimental condition" (Haufe et al., in press, p. 2). Haufe and colleagues propose a method to retrieve the activation patterns from the extraction filter and the covariance matrix of the data. An alternative method that can localise neural sources is the searchlight approach. In fact, significantly above-chance accuracy with which specific brain states can be decoded within a particular searchlight sphere is interpreted as indicating the presence of class-specific information in that region. The last application of MVPA is pattern *characterisation*, i.e. to define how different classes (e.g., conditions,

stimuli, responses) are encoded in the brain, and how this code is linked to known relationships among stimuli (Kamitani & Tong, 2005; Kriegeskorte et al., 2006). For example, *representation similarity analysis* (RSA) aims to identify more abstract similarity patterns across different conditions or stimuli (Kriegeskorte, Mur, & Bandettini, 2008). This last application, in particular, is a very peculiar aspect of the method that makes MVPA a powerful tool for fMRI data analysis (Pereira et al., 2009).

CHAPTER 3

3. Study 1: Automatic processing of conditional rules

This chapter describes a research project published on *Acta Psychologica* (Reverberi, Pischedda, et al., 2012) and testing whether deductive inferences can be triggered unconsciously.

3.1. Theoretical background

Deduction "is the process of drawing conclusions that are guaranteed to follow from given premises" (Prado et al., 2011, p. 3483). The ability to recognise simple deductive arguments is a core property of rational thought (Rips, 1988). Two of the most intuitively straightforward deductive inferences are the *Modus Ponens* and the *Disjunctive Syllogism*:

1. Modus Ponens: If A then B; A; Therefore, B;
2. Disjunctive Syllogism: A or B or both; Not A; Therefore, B.

Acceptance of the Modus Ponens inference is nearly ubiquitous, with individuals correctly applying the inference in roughly 90% to 100% of the cases. Disjunctive Syllogism inferences are slightly more difficult and are, on average, correctly applied 80% of the times (Braine & O'Brien, 1998; Evans, Newstead, & Byrne, 1993). Because these two deductive processes are so fundamental to human rationality, some scholars contend that Modus Ponens and Disjunctive Syllogism inferences occur automatically and are triggered whenever a set of premises matches the corresponding logical schema, even without intention (Braine & O'Brien, 1998). Thus far, this issue has not been definitively settled by empirical studies. Nevertheless, some methodologically interesting strategies have been proposed, which capitalise on "priming effects" or the effects of pre-activated information on the processing of target stimuli (Meyer & Schvaneveldt, 1971; Neely, 1977). In the present case, if, for example, Modus Ponens were fully automatic, given the premise pair: "If there is a 3 then there is an 8" (major premise, P_1) and "there is a 3" (minor premise, P_2), individuals should activate the

conclusion “there is an 8”, even if they did not intentionally try to complete this premise pair. Similarly, if the Disjunctive Syllogism is automatic, the pair of premises “there is not a 3 or there is an 8” (P_1) and “there is a 3” (P_2) should unavoidably activate “there is an 8”. If those inferences are automatically triggered, there should be a priming effect when participants are shown, after the premise pair, a target number (T) that matches the conclusion (“8” in the previous examples), even if the participant has not been required to deduce a conclusion from the premises. In this way it would be possible to examine the automaticity of deductive reasoning across different types of inferences. However, all previous investigations of priming effects in processing of Modus Ponens or Disjunctive Syllogisms have used text comprehension tasks, in which the premises are embedded in text that participants are explicitly required to understand (Lea, 1995; Rader & Sloutsky, 2002). Under these circumstances, it is difficult to argue that participants are not voluntarily reasoning, even though they are not explicitly told to do so. Subliminally presented stimuli can be processed at a semantic level, as demonstrated by Naccache and Dehaene (2001b). Hence, it is plausible that subliminally presented stimuli may trigger inferences, if another premise is already encoded in WM and if that inference can be made without voluntary control. The finding that valid conclusions in Modus Ponens or Disjunctive Syllogism problems prime subsequent targets when one of the problem premises is not consciously available to the participant may indicate that the inference was triggered and carried out without voluntary control.

3.2. Experiment 1

3.2.1. Methods

Fifty-four (44 females) graduate and undergraduate students from the University of Milano-Bicocca participated in the experiment in exchange for course credit.

The experiment involved two tasks: a *number evaluation task* (512 trials), designed to detect priming effects on target numbers, and a *sentence recognition task* (128 trials). The aim of the **sentence recognition** task was to check whether the subjects correctly encoded the first premise (P_1) in memory (Figure 3.1), as explicitly instructed (see below), and to exclude from the analyses those who did not. All trials had two premises, P_1 and P_2. In the number evaluation trials, the second premise (P_2) was followed by a

target number (T) and participants were required to judge whether the number was odd or even. In the sentence recognition trials, P_2 was followed by two sentences, one of them identical to the P_1. Participants were required to choose the sentence identical to P_1.

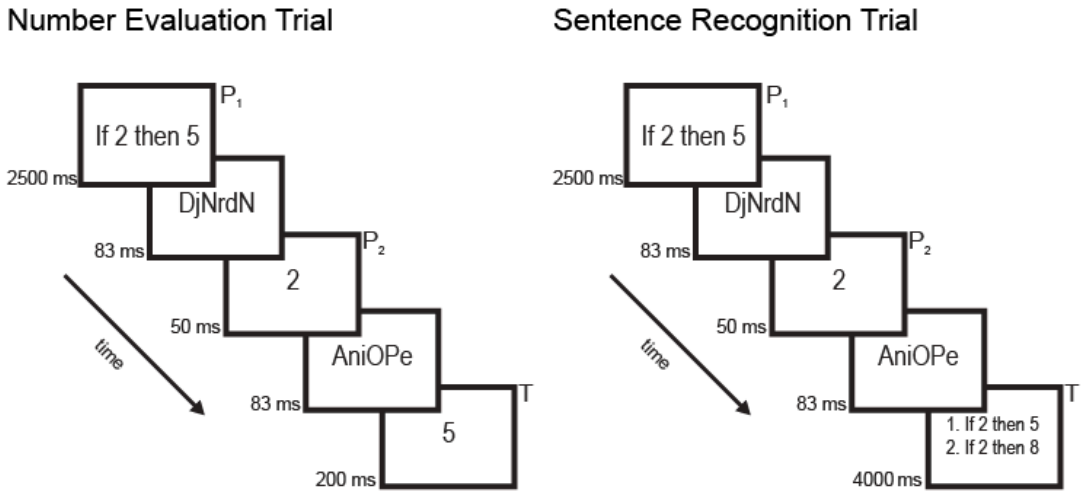


Figure 3.1 An example of two trial types used in Experiments 1-3. On the left, we show an example of a number evaluation trial. The participants had to read and remember the first premise (P_1). Following P_1, P_2 was presented very briefly (50ms), preceded and followed by random string of ten characters. This masking procedure prevented most participants (see Experiments 2 and 3) to consciously access the number presented on P_2. Finally, a target number (T) was presented briefly (200ms). The participants had to judge within 2000ms from target presentation whether the number was odd or even. Sentence recognition trials (example on the right) were very similar to number evaluation trials. The only difference was at the target phase: instead of a number, two sentences were presented. The participant had to recognise the sentence that was identical to P_1. The target sentences remained on the screen for at most 4000ms.

More specifically, in **number evaluation** trials, P_1 was either a conditional (e.g., "if there is a 2 then there is a 4") or a disjunctive (e.g., "there is a 2 or there is a 4") statement. P_2 and T were numbers (e.g., "2"). As a short code for identifying each experimental trial in the following paragraphs, we state the P_1, P_2, and target number in a synthetic form. For instance, "if p then not q; r; p" stands for "if there is a p, then there is not a q (P_1); r (P_2); p (T)". The literals p, q, r, and s stand for Arabic numerals. Matching literals correspond to matching numbers. During the experiment, the specific numbers used as p, q, r, or s were randomly chosen from numbers between 1 and 8. Eight different types of P_1 were administered, corresponding to the cells of a 2 (logical connective: if *vs.* or) × 4 (negation: none, first argument, second argument, both arguments) design. P_2 was either identical to one of the numbers in P_1 (i.e., p, q) or different from both of them (r). The target (T) could be p, q, r, or s, being s a number

different from all the preceding ones. Overall, 64 trial types were considered. The trial types relevant for the assessment of the automaticity of inferences were 32 (Table 3.1). The other trials were used as fillers. In the filler trials either T was identical to P_2 (e.g., "p or q; p; p") or the numbers involved in the whole trial were different (e.g., "p or q; r; s").

Table 3.1 Trial types used for assessing inference making.

	Inference Problem	**Baseline**
Conditional Problems		
1.	If p then q; p; q	If p then q; r; q
2.	If not p then q; p; q	If not p then q; r; q
3.	If p then not q; p; q	If p then not q; r; q
4.	If not p then not q; p; q	If not p then not q; r; q
5.	If p then q; q; p	If p then q; r; p
6.	If not p then q; q; p	If not p then q; r; p
7.	If p then not q; q; p	If p then not q; r; p
8.	If not p then not q; q; p	If not p then not q; r; p
Disjunctive Problems		
9.	p or q; p; q	p or q; r; q
10.	not p or q; p; q	not p or q; r; q
11.	p or not q; p; q	p or not q; r; q
12.	not p or not q; p; q	not p or not q; r; q
13.	p or q; q; p	p or q; r; p
14.	not p or q; q; p	not p or q; r; p
15.	p or not q; q; p	p or not q; r; p
16.	not p or not q; q; p	not p or not q; r; p

The literals p, q, r, and s stand for Arabic numerals between 1 and 8. Matching literals correspond to matching numbers.

Overall, the **design** allowed that the first premises (P_1) could or could not endorse a valid conclusion, depending on the presence of negations, the type of connective, and the type of P_2. Beyond Modus Ponens and Disjunctive Syllogism inferences, the problems comprised a wide array of other inferences, such as valid *Modus Tollens* ("if A then B; not B; therefore, not A", tested in trials such as "if not p then not q; q; p") and four invalid inferences that are often endorsed by naïve reasoners: "*Affirmation of the Consequent*"

("if A then B; B; therefore, A"), "*Denial of the Antecedent*" ("if A then B; not A; therefore, not B") and the corresponding invalid disjunctive inferences "A or B; A; therefore, not B" and "A or B; B; therefore, not A". Each experimental condition was matched with a corresponding *baseline* condition in which all the elements visible to the subject (P_1 and T) were identical to those in the experimental condition but with a different P_2, not visible to the participants (see Table 3.1). Thus, for example, in the case of the Modus Ponens trials ("if p then q; p; q"), the relevant baseline was ("if p then q; r; q"). Given the critical difference in P_2 between experimental and baseline conditions, any inference possibly triggered in the experimental trials would not be triggered in the corresponding baseline trials. Importantly, however, any effect related only to P_1 and T (the visible part of the problems) would be equally present in both the experimental and baseline conditions. In this way, effects related only to the interaction between P_1 and T would be cancelled out.

Participants were informed that they would be presented with sentences about the possible presence of digits on an imaginary blackboard. There were no explicit limits on the number of digits that the blackboard might contain. Following this, stating "There is a 2" did not imply the absence of other digits than 2. Each trial started with a fixation point (1000ms), followed by the presentation of P_1 for 2500ms. Participants had to read and remember P_1. After P_1, a mask of 10 random characters was displayed for 83ms (Figure 3.1), followed by the presentation of P_2 (50ms), and then by another 10-character mask (83ms). After the second mask, there was a delay of 750ms in the long inter-stimulus interval (ISI) group and no delay for the short ISI group. This method was done to sample priming effects across different time frames (Draine & Greenwald, 1998). In number evaluation trials, the target number was presented for 200ms and the participant had to determine whether the number was odd or even. Up to 2000ms were allowed for parity judgment on the target. If no response was produced within the time limit, the trial ended and was scored as incorrect. In sentence recognition trials, a pair of statements was presented and the participants had to identify the one that exactly matched P_1. The statements disappeared as soon as the participants responded or after 4000ms if no response was given. The two response keys were the same in both tasks, one of them meaning "first sentence" or "odd number" and the other meaning "second

sentence" or "even number". The response mapping was counterbalanced across subjects.

Before the experiment, participants were given written instructions describing the task. The presence of a number (P_2) between the masks was not mentioned. The importance of answering both quickly and correctly was stressed in the number evaluation trials. For sentence recognition trials, accuracy was emphasised more than speed. A 40-trial training session with accuracy feedback was administered before the experimental session. RTs on the number evaluation task were used to assess whether processing of targets was affected by the deductive conclusions entailed by the premises. Subjects with an average score lower than 90% on the sentence recognition trials were discarded from further analyses because the triggering of an inference – if any – strictly depended on an adequate representation of P_1.

The experiment was programmed and administered using Presentation(R) (Neurobehavioral Systems, Inc., San Francisco, CA, US). The CRT screen had a 60Hz refresh rate and the resolution was set at 1024 × 768 pixels. The premises, masks, and target stimuli were written in white, 22-point Arial characters and were presented against a black background. Two buttons on a serial mouse were used as response keys. Participants responded with the index fingers of both hands. Participants sat 60cm from the screen in a quiet, dimly lit lab.

3.2.2. Results

Thirteen participants were discarded from our analyses because of low accuracy in the sentence recognition task. The remaining 41 participants had accuracies > 90% in both tasks. After the experiment, participants were asked whether they ever noticed a number (the masked P_2) appearing during the presentation of the two random strings of characters. Nobody reported noticing P_2.

In analysing number evaluation trials, we considered only responses that were 250ms < RT < 1000ms (92.5% of all responses). This procedure is consistent with existing priming research (e.g, Kouider & Dehaene, 2009; Naccache & Dehaene, 2001a, 2001b). We adopted it to remove possible outliers (i.e., either anticipatory or delayed responses) and to keep our findings comparable with those already available in literature. As a first step, we assessed the presence of **priming effects** in the Modus

Ponens ("if p then q; p; q") and Disjunctive Syllogism ("not p or q; p; q"; "p or not q; q; p") problems by comparing the RTs of the critical problems with the RTs of the relevant baselines (which we refer to as ΔRT). Then, we tested for priming effects in ΔRT in the other relevant problems to evaluate whether the effects possibly observed for Modus Ponens or Disjunctive Syllogism were specific to the type of problem. The additional control conditions involved: (*i*) all other problems that validly or invalidly entailed T (e.g., Affirmation of the Consequent) and (*ii*) all problems that did not support T, but bore a strong similarity to Modus Ponens or Disjunctive Syllogism problems, differing from them in one feature only (e.g., Modus Ponens *vs.* "p or q; p; q"; Modus Ponens *vs.* "if p then not q; p; q").

A preliminary ANOVA factoring the two critical problems (Modus Ponens and Disjunctive Syllogism) and the ISI condition (long and short) with ΔRT as the dependent variable showed that the interaction was not significant ($F(1,39) = .35$; $p = .55$). Data from the short and long ISI conditions were therefore pooled together.

Next, we tested for priming effects in Modus Ponens and Disjunctive Syllogism problems. We found a priming effect for Modus Ponens (ΔRT = -20.6, $\eta_p^2 = .162$, $t(40) = 2.77$, $p = .008$) but not for Disjunctive Syllogisms (ΔRT = 9.7, $\eta_p^2 = .059$, $t(40) = 1.58$, $p = .121$). Subjects were *faster* at answering "odd" or "even" when the target represented the conclusion of a Modus Ponens inference. Given this priming effect for Modus Ponens, we ran planned control analyses. The comparisons and results are detailed in Table 3.2. Overall, no priming effect was observable in any of the control problems suggesting that the effect we found for Modus Ponens is highly specific to the type of problem. To further test this conclusion, we directly compared the ΔRT of Modus Ponens with the ΔRT of all control problems. The ΔRT was significantly smaller for Modus Ponens than for all other control trials (ΔRT = -21 *vs.* -1; $t(40) = 2.31$, $p = .026$). The ΔRT of the Modus Ponens was also smaller than the mean ΔRT related to all other types of trials (ΔRT = -21 *vs.* 2; $t(40) = 3.06$, $p = .004$). Furthermore, we compared the RT on Modus Ponens with a more "general" baseline, consisting in the average RT over all types of conditional baseline trials (i.e., any conditional trial using "r" as P_2). The RT on Modus Ponens was faster than the general baseline ($t(40) = 3.94$, $p < .001$). By contrast, none of the other control problems differed from the general baseline (all *ps* > .1).

Table 3.2 Test of priming effects on the Modus Ponens inference and the relevant control conditions.

P_1	P_2	T	RT exp trial (SD)	ΔRT (SD)	$t(40)$	p	η_p^2	Description of the experimental trial
If p then q	p	q	601 (100)	- 21(47)	2.78	.008	.162	Target matches Modus Ponens inference
If p then q	p	p	621 (104)	- 11 (71)	0.96	.34	.022	Target does not match Modus Ponens inference
If p then not q	p	q	619 (102)	2 (54)	0.26	.79	.002	Target does not match Modus Ponens inference
If not p then q	p	q	617 (103)	-12 (61)	1.24	.22	.037	Target does not match any inference from the premises
If not p then not q	q	p	635 (101)	6 (59)	0.63	.53	.010	Target matches Modus Tollens inference
If p then q	q	p	622 (100)	-10 (60)	1.02	.32	.025	Target matches Affirmation of the Consequent fallacious inference
If not p then not q	p	q	630 (106)	9 (54)	1.13	.27	.031	Target matches Denial of the Antecedent fallacious inference
p or q	p	q	626 (105)	16 (57)	1.77	.08	.073	Target matches a Disjunctive Fallacy

ΔRT: Difference in RT between the experimental problems and the relevant baseline problems.

Finally, no priming effects for Modus Ponens or Disjunctive Syllogisms were found when separate analyses were carried out on the data from the 13 participants who were discarded due to low performance in the sentence recognition task. This finding suggests that an adequate representation of the first premise in WM is a necessary condition for the observation of the priming effect in Modus Ponens trials.

3.2.3. Discussion

The findings of Experiment 1 are interesting in several respects. First, we showed that judging parity is faster in trials in which the number to be evaluated is the conclusion of a single step Modus Ponens inference ("if p then q; p; q") than in baseline trials, that were identical except for the non-matching second premise ("if p then q; **r**; q"). Critically, this was the case even though, in our paradigm, the second premise was reportedly *invisible* to the subjects. Thus, as far as the subjective experience of the participants was concerned, the Modus Ponens problems and the relevant baseline problems were indistinguishable.

Second, the observed advantage in RTs was *only* found for the Modus Ponens problems (Table 3.2). The effect was not present in the Affirmation of the Consequent fallacy ("if p then q; q; p"), thus showing that the effect was not based on a bidirectional association of the antecedent ("p") and consequent ("q") of the conditional premise. Furthermore, the effect was not present in trials with a sequence of numbers *identical* to Modus Ponens ("p.q.p.q") but containing negations (e.g., "if not p then q; p; q") or containing a disjunction instead of a conditional connective ("p or q; p; q"). Thus, the effect observed for Modus Ponens cannot be merely considered an association of the two arguments (p and q) of the conditional rule independent of the presence of negations and the type of logical connective involved.

Third, trials in which Modus Ponens would be triggered by an indirect match between P_2 and P_1 did not show the effect. This suggests that the match between P_2 and the antecedent in P_1 needs to be straightforward (e.g., an identity between the antecedent in P_1 and P_2) to automatically trigger the Modus Ponens inference.

Fourth, the Disjunctive Syllogisms ("not p or q; p; q"; "p or not q; q; p") did not demonstrate a similar priming effect relative to their baseline trials ("not p or q; r; q"; "p or not q; r; p"). Similar to preceding studies arguing for the automaticity of some basic inferential schemata (Lea, 1995), we expected that both the Modus Ponens and the Disjunctive Syllogism would be triggered automatically by an unconscious premise or that neither would. Instead, the results strongly suggest that Modus Ponens is a "special" type of inference, one that can be triggered even by invisible premises, seemingly in the absence of conscious willingness to reason, whereas the Disjunctive Syllogism – apparently – is not drawn if the minor premise is not seen.

3.3. Experiment 2

No participant in Experiment 1 reported noticing the presentation of any number after P_1. They were aware only of having seen a string of flickering letters (the masks). Even though this *subjective* measure is valid for establishing that P_2 was subjectively non-conscious according to some authors and theories (e.g., Seth, Dienes, Cleeremans, Overgaard, & Pessoa, 2008), according to others, it is not exhaustive (e.g., Kouider & Dupoux, 2001). Thus, the **aim** of this control experiment was twofold. First, we wanted

to replicate the finding that, in the same experimental conditions as in Experiment 1, participants do not report noticing P_2 when they are explicitly questioned about it after the experiment. Second, and more importantly, we wanted to check whether participants who were alerted to the presence of P_2 between the two masks could access information about it.

3.3.1. Methods

Twenty-three (9 females) graduate and undergraduate students from the University of Milano-Bicocca participated in the experiment in exchange for course credit.

Experiment 2 was composed of two independent tasks, administered in the same order.

Task 1. The first task was a shortened version of Experiment 1. The stimuli were the same, but only 128 *number evaluation trials* (i.e., two trials per problem type) and 32 *sentence recognition trials* were used. After that, a *debriefing questionnaire* was administered, asking whether the participants saw anything between the two strings of random characters, and – if so – what they saw.

Task 2. Task 2 was administered immediately after Task 1. The same stimuli as in Experiment 1 were used, with the exception that half the time P_2 was a number and half the time a letter. Subjects were alerted to the presence of a number or a letter between the two strings of random characters and were explicitly asked to both *(i)* read and retain P_1 in memory and *(ii)* try as hard as possible to detect whether P_2 was a number or a letter. Sixty-four *sentence recognition trials*, identical to those in Experiment 1, were administered to assure that the WM load in Task 2 was similar to that in Experiment 1. Additionally, 256 *detection trials* were administered, 128 displaying a number in P_2 and 128 displaying a letter. These trials were a substitute for the number evaluation trials in Experiment 1. In detection trials, the question "Was there a number or a letter between the two strings of characters?" was displayed on the screen after the second mask. Participants were told to choose a category, even if they were uncertain or had the impression of having seen nothing. Sentence recognition trials and detection trials differed only during question/target presentation. Participants could know whether it was a sentence recognition trial or a detection trial only at the end of a trial. Thus, the

processing of both kinds of trials was identical until the final phase. Detection trials and sentence recognition trials were randomly interspersed.

The timeline for both tasks was the same as in Experiment 1, except that the ISI after the second mask was always set at 0ms. There were no time limits for responding to detection trials. The first task was preceded by a training session of 40 trials with feedback on accuracy. The main dependent variable of the first task was the response to the debriefing questionnaire. A 10-trial training session with accuracy feedback preceded the second task. The main dependent variable of the second task was the *d'* (sensitivity index, i.e., a measure comparing the *hit rate* and the *false-alarm rate*) in the detection trials. In the current study, we lowered the recognition-accuracy threshold in the sentence recognition trials from 90% to 80% accuracy because participants performed worse on this measure than in the previous study due to the high cognitive demand required to detect P_2 (using a 90% threshold would exclude 61% of participants). This choice implies that participants in this task had – on average – a slightly reduced WM load compared to those in Experiment 1. Nevertheless, please notice that a lower exclusion threshold should, if anything, decrease priming effects.

The experiment was programmed and administered using Presentation(R) (Neurobehavioral Systems, Inc., San Francisco, CA, US). The CRT screen had a 60Hz refresh rate and the resolution was set at 1024 × 768 pixels. The premises, masks, and target stimuli were written in white, 22-point Arial characters and were presented against a black background. Two buttons of a serial mouse were used as response keys. Participants responded with the index fingers of both hands. Participants sat 60cm from the screen in a quiet, dimly lit lab.

3.3.2. Results

Task 1. None of the participants reported having seen a number (P_2) between the masks. Thus, the task confirmed that participants did not notice P_2, when they were not alerted to the presence of P_2 and were not explicitly told to detect it.

Notwithstanding the low power of Experiment 2, we also checked whether the main findings of Experiment 1 replicated in Experiment 2. Seven participants had to be excluded because of accuracy scores < 80% in the sentence recognition task and three subjects had to be excluded because they had less than two trials per condition after

removal of incorrect trials. This left 13 participants for the analyses. Given that Experiment 1 established the directionality of the priming effect for Modus Ponens, we used one-tailed statistical tests. A **priming effect** specific for Modus Ponens was found also in Experiment 2. The Modus Ponens trials were faster compared to the relevant baseline ("If p then q; p; q" *vs.* "If p then q; r; q"; $t(12) = 2.43$, $\eta_p^2 = .33$, $p = .016$). By contrast, the priming effect was not present in the Affirmation of the Consequent fallacy ("If p then q; q; p" *vs.*" If p then q; r; p"; $t(12) = 1.09$, $\eta_p^2 = .09$, $p = .15$) and the Disjunctive problem similar to Modus Ponens ("p or q; p; q" *vs.* "p or q; r; q"; $t(12) = .13$, $\eta_p^2 = .001$, $p = .45$). As in Experiment 1, we directly compared the ΔRT of Modus Ponens with the ΔRT of all control problems. The ΔRT was significantly smaller for Modus Ponens than for all other control trials ($t(12) = 2.32$, $p = .019$). The ΔRT of the Modus Ponens was also smaller than the mean ΔRT related to all other types of trials ($t(12) = 2.60$, $p = .012$).

Task 2. Nine participants were excluded from the following analyses because of accuracy scores < 80% in the sentence recognition task. Of the remaining 14 participants, one was strongly biased toward the "letter" response: she was 100% correct when P_2 was a letter (not allowing us to compute her d'), but 80% of the time confused numbers for letters. The mean **d'** of the remaining 13 participants was .46 (median = .37), which is significantly greater than zero ($t(12) = 4.58$, $p = .001$). This finding corresponds to an overall accuracy rate of 56% with 11% more hits than false alarms, meaning that – in aggregate form – they were moderately able to discriminate whether the second premise was a letter or a number.

3.3.3. Discussion

The first task used a subjective measure of consciousness (Seth et al., 2008) and confirmed that in an experimental setting similar to Experiment 1 people did not report having seen P_2. Furthermore, Experiment 2 replicated the main findings of Experiment 1. Namely, a priming effect was found for Modus Ponens ("if p then q; p; q") but not for the Affirmation of the Consequent ("if p then q; q; p") and for the disjunctive trial similar to Modus Ponens ("p or q; p; q"). Two facts rule out the possibility that the effect we observed was learned during our experimental procedure: 1) The effect could be found even after the administration of only two Modus Ponens trials with the target matching the conclusion and 2) No feedback was provided during Experiment 1 or 2.

In the second task, participants were informed of the possible presence of a number in P_2 and were asked to keep track of it. This instruction produced an expected reallocation of attention to P_2, as indexed by a poorer overall performance in the sentence recognition trials. In these more favourable conditions, the subjects were able to judge whether P_2 was a letter or a number with better-than-chance accuracy. It must be noted, however, that – even though alerted to the presence of P_2 – participants attained a rather low d'. Overall, our findings suggest that the P_2 premises in Experiment 1 were probably "preconsciously" processed, which, according to a distinction proposed by Dehaene and collaborators, means that they were only weakly processed but could be consciously detected if enough attention were devoted to them (Dehaene, Changeux, Naccache, Sackur, & Sergent, 2006; Kouider & Dehaene, 2007). Accordingly, it is possible that, in a few trials, some participants in Experiment 1 were conscious of P_2 even though they might have considered the experience too weak to be reported. It should be noted that this result does not change the fact that parity judgments in Experiments 1 and 2 were selectively enhanced when the target numbers matched the valid conclusion of a Modus Ponens inference. However, the findings in Experiment 2 weaken the claim that a Modus Ponens inference can be triggered even if its minor premise is presented *subliminally*. It is possible that the advantage on Modus Ponens in Experiments 1 and 2 depends on the few trials in which the participants might have accessed some information about P_2.

3.4. Experiment 3

Experiment 3 had two **goals**. First and foremost – following Experiment 2 – we wanted to check whether the priming effect observed for Modus Ponens in Experiment 1 was present in subjects for whom processing of P_2 could be considered *subliminal*. Thus, in Experiment 3, we considered only those subjects who showed chance performance in identifying P_2 in a forced-choice detection task. Furthermore, the procedure adopted in Experiment 3 for assessing conscious access to P_2 improved compared with Experiment 2. In Experiment 2 we only checked whether the participants could access relatively superficial information, i.e., whether a stimulus was a number or a letter. However, it is possible that the participants had conscious access only to *some* information about P_2

(e.g., “number or letter?”) but not to other (e.g., “which number was presented?”). Given that the critical information to trigger Modus Ponens or other inferences is indeed the identity of P_2, in Experiment 3 we tested more specifically for the availability of this information (see also Naccache & Dehaene, 2001b for a similar line of reasoning).

The second goal was to test whether the advantage for Modus Ponens could also be observed when the conditional major premise P_1 was presented in an inverted form, namely with the logical consequent (in this case p) preceding – linguistically – the logical antecedent (q). To accomplish this, we added a subset of “inverted Modus Ponens” trials, whose form was “p if q; q; p” (e.g., P_1: “There is an 8 if there is a 3”; P_2: “3”; T: “8”), and the corresponding baseline trials (“p if q; r; p”) to the main task.

3.4.1. Methods

Forty-three (35 females) graduate and undergraduate students from the University of Milano-Bicocca participated in the experiment in exchange for course credit.

Task 1. Task 1 consisted of a shorter version of Experiment 1. Specifically, for the *number evaluation trials*, negations of P_1 in the experimental trials and some of the original filler trials were not used, leaving only four experimental problems and their corresponding baselines. To these, we added two experimental problems with an inverted conditional in P_1 of the form “p if q” and their baselines. To avoid always presenting target numbers that were also mentioned in P_1, we introduced six filler conditions in which a different number was used as the target. The full list of the types of trials in the number evaluation task is reported in Table 3.3. The numbers used as p, q, or r were randomly chosen from numbers between 1 and 8.

Finally, to reduce the similarity of the visible sentences in the task, we also included some trials where we randomly inserted negations of P_1. Overall, we administered 32 replications for each experimental cell and 96 trials with random negations, corresponding to $32 \times 18 + 96 = 672$ trials in the number evaluation condition. We also administered 130 *sentence recognition trials*, in which the features of P_1 (type of logical connective and presence or absence of negations) and P_2 (the numbers p, q, or r) were independently and randomly manipulated.

Table 3.3 Types of problems used in Experiment 3.

Type	Problem	Baseline	Filler
Standard Modus Ponens	If p then q; p; q	If p then q; r; q	If p then q; p; r
Standard Affirmation of the Consequent	If p then q; q; p	If p then q; r; p	If p then q; q; r
Disjunctive fallacy 1	p or q; p; q	p or q; r; q	p or q; p; r
Disjunctive fallacy 2	p or q; q; p	p or q; r; p	p or q; q; r
Inverted Modus Ponens	p if q; q; p	p if q; r; p	p if q; q; r
Inverted Affirmation of the Consequent	p if q; p; q	p if q; r; q	p if q; p; r

Task 2. Task 2 was administered immediately after Task 1. In Task 2, we used a set of stimuli similar to Task 1, with the exception that P_2 was always a random number between 0 and 9. Subjects were alerted to the presence of a number between the two strings of random characters and were *explicitly* asked: (*i)* to read and retain P_1 in memory and (*ii)* to try as hard as possible to identify P_2. Ninety-six *detection trials* were administered. In addition, 48 *sentence recognition trials*, identical to those in Task 1, were presented to make sure that the WM load in this task was the same as in Task 1. In *detection trials*, the question "Which number was there between the two strings of random characters?" was displayed on the screen after the second mask. Participants were told to choose a number and press the corresponding key on the PC keyboard, even in trials in which they were highly uncertain about the identity of P_2. Sentence recognition trials and detection trials differed only during question/target presentation. Participants could know whether it was a sentence recognition trial or a detection trial only at the end of a trial. Thus the processing of both kinds of trials was identical until the final phase. Detection trials and sentence recognition trials were randomly interspersed. There were no number evaluation trials in Task 2.

The timeline of both tasks is the same as in Experiment 1, except that the ISI after the second mask varied across four levels: 0, 250, 500 or 750ms. In Task 2, subjects had no time limit to respond to detection trials. The first task was preceded by a training session of 50 trials with feedback on accuracy. The second task was preceded by a 10-trial training session with accuracy feedback.

The experiment was programmed and administered using Presentation(R) (Neurobehavioral Systems, Inc., San Francisco, CA, US). The CRT screen had a 60Hz

refresh rate and the resolution was set at 1024 × 768 pixels. The premises, masks, and target stimuli were written in white, 22-point Arial characters and were presented against a black background. Two buttons on the Cedrus™ Response Box RB-730 were used as response keys in Task 1. Participants responded with the index fingers of both hands. In Task 2, responses were collected via the keyboard. Participants sat 60cm from the screen in a dimly lit lab. To limit distraction due to ambient phasic noise, participants wore Sennheiser HD280 headsets, providing 32db of attenuation in external noise. Finally, white noise was played in the headset to further isolate the subject. A photodiode was connected to an oscilloscope to check exactly how long the short P_2 stimuli were displayed on the screen. The oscilloscope confirmed that the stimuli were displayed for, at most, 50ms as requested. More specifically, as expected on CRT monitors, the brightness of the stimuli started to decay slightly before 50ms so that at 40ms the brightness was about 1lux, at 45ms about 0.5lux, and at 50ms less than 0.1lux.

3.4.2. Results

Task 2. Average accuracy on the sentence recognition trials in Task 2 was 91% and no participant responded at chance. The average **accuracy** on the detection task was 21%. This was greater than the accuracy level expected by chance alone (10%, $t(42) = 5.19$, $p < .001$) showing that, on average, subjects were able to get some information about the identity of P_2, even though performance was far from optimal, closely replicating the results of Experiment 2. The correlation between accuracy on the sentence recognition task and on the detection task was negligible ($r = - .11$, $p > .1$); this assures that the differences in performance in the detection task were not due to a general disengagement from the task. The ability to **identify P_2** varied widely across subjects, from chance-level to a maximum of 61% accuracy. In particular, a subset of 18 subjects performed at chance (< 16% accuracy) as determined by a binomial test. The average accuracy of these subjects was 10.7% (Standard Deviation, SD = 3%) and was not significantly different from chance ($t(17) = .99$, $p > .1$).

Task 1. The main objective in Experiment 3 was to check whether the **priming effect** observed in the Modus Ponens trials could be replicated for participants who did not have conscious access to the identity of P_2. Accordingly, we applied our main

analyses to the 18 subjects with chance performance on Task 2. Given that Experiment 1 established the directionality of the priming effect for Modus Ponens, we used one-tailed statistical tests. As in Experiment 1, in analysing number evaluation trials we considered only responses that were 250ms < RT < 1000ms. The pattern of results closely replicated the main findings of Experiment 1. In all trials and conditions, subjects were accurate and fast in judging whether the target number presented was odd or even: average accuracy was 96% (SD = 3%) and the average RT was 660ms (SD = 108ms). Most importantly, however, subjects were faster in judging whether a target number was odd or even when the target was the conclusion of a Modus Ponens: trials "if p then q; p; q" were faster than trials "if p then q; r; q" (ΔRT = -14.0ms, η_p^2 = .242, $t(17)$ = 2.33, p = .016). As in Experiment 1, this advantage was specific to the Modus Ponens problems; the effect was not present in any of the other conditions. In particular, we did not find any advantage for the Affirmation of the Consequent fallacy "if p then q; q; p" over the corresponding baseline "if p then q; r; p" (ΔRT = -4.2ms, η_p^2 = .014, $t(17)$ = .484, p > .1) and for the relevant disjunctive fallacy "p or q; p; q" *vs.* "p or q; r; q" (ΔRT = 1.0ms, η_p^2 = .001, $t(17)$ = .11, p > .1). To further test this conclusion, we directly compared the ΔRT of Modus Ponens with the ΔRT of all other problems, but inverted Modus Ponens. The ΔRT of the Modus Ponens showed a robust statistical trend in being smaller than the mean ΔRT related to all other types of trials (ΔRT = -14.0 *vs.* -4.3; $t(17)$ = 1.67, p = .056). The inverted Modus Ponens "p if q; q; p" trials did not reach a significant ΔRT with respect to their baseline, but they showed a trend in the expected direction (ΔRT = -10.2, η_p^2 = .126 $t(17)$ = 1.56, p = .068). Conversely, the inverted Affirmation of the Consequent "p if q; p; q" showed no advantage over its control baseline (ΔRT = -6.4ms, η_p^2 = .037, $t(17)$ = .81, p > .1).

Subjects in this subgroup had an average accuracy on the sentence recognition task of 88% (SD = 6%). All subjects had an accuracy level higher than chance. Given the relatively small size of the subgroup, we could not apply a selection criterion based on the sentence recognition performance as stringent as in Experiment 1, as it would have left only eight subjects for analysis. However, we applied a more moderate threshold (performance on sentence recognition trials higher than 80%) and found that the remaining 16 subjects showed the same pattern of performance as the whole sample, but more pronounced. The Modus Ponens condition was faster than baseline (ΔRT =-

17.1ms, η_p^2 = .437, $t(15)$ = 3.41, p = .002), while the other control conditions were not (Affirmation of the Consequent: ΔRT = -5.4ms, η_p^2 = .023, $t(15)$ = .59, p > .1; disjunctive fallacy: ΔRT = 1.0ms, η_p^2 = .001, $t(15)$ = .12, p > .1). Finally, the inverted Modus Ponens condition still showed a statistical trend towards being faster than the baseline (ΔRT =-10.3ms, η_p^2 = .133, $t(15)$ = 1.51, p = .075), whereas the inverted Affirmation of the Consequent condition did not show any advantage (ΔRT =-5.1ms, η_p^2 = .032, $t(15)$ = .70, p > .1).

3.4.3. Discussion

Experiment 3 replicated and extended the main finding of Experiment 1, namely the positive priming effect caused by the automatic triggering of a Modus Ponens inference. We tested whether the priming effect for Modus Ponens observed in Experiment 1 could be found also in subjects who processed P_2 *subliminally*. To do this, we selected participants that were at chance in a forced-choice detection task on P_2. Even though these subjects did not have either conscious or pre-conscious access to P_2, the priming effect for the standard Modus Ponens was still observed. Furthermore, as in Experiment 1, this effect was specific: it was absent in other highly similar conditions, namely the Affirmation of the Consequent ("if then q; q; p") and the disjunctive problem with a sequence of arguments paralleling the standard Modus Ponens argument ("p or q; p; q").

A further objective of Experiment 3 was to test whether the effect we found for Modus Ponens inferences in Experiment 1 would generalise to Modus Ponens inferences relying on an alternative, non-standard representation of the problem. We considered an "inverted" version of the Modus Ponens in which the logical consequent linguistically preceded the logical antecedent (i.e., the trials in which P_1 had the form "there is p if there is q"). In addition to Modus Ponens, we also considered an "inverted" version of the Affirmation of the Consequent fallacy. Notice that the sequence of literals in the inverted fallacy is identical to the sequence of literals in the standard Modus Ponens ("p.q.p.q") and, analogously, the sequence of literals in the inverted Modus Ponens is identical to the standard fallacy. The findings were in the expected direction. The inverted fallacy did not show any effect, while the inverted Modus Ponens showed a statistical trend. The fact that the inverted fallacy was not significant corroborates evidence from Experiment 1 showing that the mere consideration of the sequence of the

literals ("p.q.p.q") is *not* enough to justify the observed effect for Modus Ponens. Nevertheless, the weaker effect shown for the inverted Modus Ponens suggests that the two alternative representations of P_1 are not equivalent (see also Grosset & Barrouillet, 2003).

3.5. Experiment 4

In Experiments 1-3 we showed that subjects judged faster whether a digit was odd or even when it was the conclusion of a Modus Ponens. We also showed that this effect was present *specifically* in Modus Ponens but not in a wide set of other conditional and disjunctive problems. In Modus Ponens, the conditional statement ("if p then q") instantiates a unidirectional linkage from the antecedent "p" to the consequent "q". Thus, it may be possible that other kinds of relations able to produce a "unidirectional linkage" between antecedent and consequent would generate a priming effect, as we observed for Modus Ponens. In Experiment 4, we **aimed** to explore whether the priming effect observed for Modus Ponens was present in problems instantiating other kinds of relations inducing a unidirectional linkage between the two terms. In particular, we considered spatial relations (e.g., "*a* before *b*") and sentences with quantifiers (e.g., "all *a* with *b*"). Besides these new conditions, we also administered the Modus Ponens problem to check for the presence of the effect in another subject sample.

3.5.1. Methods

Twenty-seven (18 females) graduate and undergraduate students from the University of Milano-Bicocca participated in the experiment in exchange for course credit.

The administered **task** was a modified version of Experiment 1. For the *number evaluation trials*, we used eight experimental conditions and their corresponding baselines (Table 3.4). Two of the eight experimental problems were a replication of two conditions already administered in all preceding experiments: the Modus Ponens and the Affirmation of the Consequent. The other six experimental problems were new conditions exploring the possible priming effects related to *spatial relations* (e.g., "There is a 2 before the 4") and *quantified propositions* (e.g., "Every 2 is together with a 4") that

suggested a directional association between the two terms. To avoid always presenting target numbers that were also mentioned in P_1, we introduced eight filler conditions in which a different number was used as a target (Table 3.4). Overall, we administered 16 replications for each experimental condition, corresponding to 384 trials in the number evaluation condition. We also administered 192 *sentence recognition trials*, in which the features of P_1 (type of logical connective and presence or absence of negations) and P_2 (the numbers p, q, or r) were independently and randomly manipulated. Thus, overall, 576 experimental trials were administered. A *debriefing questionnaire* (see Appendix A1) was administered after completion of the main task, exploring whether the participants saw digits between the two strings of random characters and – if so – how many times.

Table 3.4 Types of problems used in Experiment 4.

Type	Problem	Baseline	Filler
Modus Ponens	If p then q; p; q	If p then q; r; q	If p then q; p; r
Affirmation of the Consequent	If p then q; q; p	If p then q; r; p	If p then q; q; r
Spatial relation 1	p before q; p; q	p before q; r; q	p before q; p; r
Spatial relation 2	p before q; q; p	p before q; r; p	p before q; q; r
Quantified relation 1	Every p with q; p; q	Every p with q; r; q	Every p with q; p; r
Quantified relation 2	Every p with q; q; p	Every p with q; r; p	Every p with q; q; r
Quantified relation 3	No p with q; p; q	No p with q; r; q	No p with q; p; r
Quantified relation 4	No p with q; q; p	No p with q; r; p	No p with q; q; r

The timeline of the task was the same as in Experiment 1, except that the ISI between the second mask and the target was always set at 0ms. The task was preceded by a training session of 50 trials with feedback on accuracy.

The experiment was programmed and administered using Presentation(R) (Neurobehavioral Systems, Inc., San Francisco, CA, US). The CRT screen had a 60Hz refresh rate and the resolution was set at 1024 × 768 pixels. The premises, masks, and target stimuli were written in white, 22-point Arial characters and were presented against a black background. Two buttons on the Cedrus™ Response Box RB-730 were used as response keys. Participants responded with the index fingers of both hands. To limit distraction due to ambient phasic noise, participants wore Sennheiser HD280

headsets, providing 32db of attenuation in external noise. Finally, white noise was played in the headset to further isolate the subject.

3.5.2. Results

Average **accuracy** in the sentence recognition trials was 96% (SD = 3%). All subjects had an accuracy level above 90% but one, who had an accuracy level between 80% and 90%. Subjects were accurate in judging whether the target numbers presented were odd or even: average accuracy was 95% (SD = 4%). All subjects had an average accuracy above 90% but one, who had an average accuracy between 80% and 90%. The subjects were fast at answering the number evaluation trials: the average RT on correct trials was 635ms (SD = 106ms). As in Experiment 1, in analysing number evaluation trials we considered only responses that were 250ms < RT < 1000ms and subject who had a high accuracy (> 90%). When specifically asked, four subjects reported having noticed the presence of digits between the two strings of random characters *less* than ten times out of 576 trials. All the other subjects did not notice that a digit was presented.

As in Experiments 1-3, subjects were faster in judging whether a target number was odd or even when it was the conclusion of a Modus Ponens: The subjects were faster in trials "if p then q; p; q" than in "if p then q; r; q" (ΔRT = -16.9ms, η_p^2 = .125, $t(24) = 1.86$, p = .035). This advantage was specific to the Modus Ponens problems: The **effect** was not present in any of the other conditions. In particular, we did not find any advantage for the Affirmation of the Consequent fallacy "if p then q; q; p" over the corresponding baseline (ΔRT = 4.70ms, η_p^2 = .016, $t(24) = -.625$, $p > .1$). This was also the case for the two spatial relations, "p before q; p; q" (ΔRT = 10.31ms, η_p^2 = .078, $t(24) = -1.428$, $p > .1$) and "p before q; q; p" (ΔRT = 4.07ms, η_p^2 = .012, $t(24) = -.539$, $p > .1$), and for all the four quantified relations: "Every p with q; p; q" (ΔRT = -3.13ms, η_p^2 = .009, $t(24) = 0.464$, $p >$ .1); "Every p with q; q; p" (ΔRT = 14.4ms, η_p^2 = .12, $t(24) = -1.82$, $p > .1$); "No p with q; p; q" (ΔRT = 11.94ms, η_p^2 = .079, $t(24) = -1.43$, $p > .1$); "No p with q; q; p" (ΔRT = -6.60ms, η_p^2 = .016, $t(24) = .62$, $p > .1$). To further test this conclusion, we directly compared the ΔRT of Modus Ponens with the ΔRT of all other problems. The ΔRT of the Modus Ponens was also smaller than the mean ΔRT related to all other types of trials (ΔRT = -18.7 *vs.* 2.8; $t(26) = 2.27$, p = .016). The overall pattern of performance did not change when we

excluded from the analyses the four subjects who noticed (less than 10 times) the presence of P_2 (Modus Ponens: ΔRT = -21.4ms, η_p^2 = .176, $t(20)$ = 2.069, p = .026). The pattern did not change also when the two participants who had an accuracy level between 80% and 90% in one of the two tasks (sentence recognition and number evaluation) were considered in the analyses as in Experiment 2 and Experiment 3. (Modus Ponens: ΔRT = -18.7ms, η_p^2 = .149, $t(26)$ = 2.13, p = .021).

3.5.3. Discussion

The findings of Experiment 4 confirmed the existence and the specificity of the priming effect in the Modus Ponens problems.

First, Experiment 4 replicated the main findings of Experiments 1-3: People were faster in evaluating the parity of a number when the number was the conclusion of a Modus Ponens inference, even though they were not required to make any inference and they were not subjectively aware that the second premise was presented.

Second, we further confirmed that the priming effect for Modus Ponens is highly specific. The priming effect, as in Experiments 1-3, was absent in the Affirmation of the Consequent problems. More interestingly, the priming effect was also absent in problems exploring spatial and quantified relations, suggesting that the observed effect is specific to conditional sentences. Furthermore, it should be noticed that a subset of the spatial and quantified problems shared with Modus Ponens the same sequence of presented digits (compare "If p then q; p; q" with "p before q; p; q" or "Every p with q; p; q" or "No p with q; p; q"). The absence of a priming effect in these cases corroborates the conclusion that presenting a sequence of numbers similar to those involved in Modus Ponens ("p.q.p.q") is *not* a sufficient condition to induce a faster processing of the target number. Finally, it should be pointed out that *neither* is logical validity. This had already been shown in Experiment 1 with the Modus Tollens and the Disjunctive Syllogism. Experiment 4 added a further valid problem to the list: "Every p with q; p; q".

Overall, the effect observed for Modus Ponens problems seems to depend on the specific representation induced by the conditional sentence "If p then q".

3.6. General discussion

In the present study, we investigated whether two basic deductive inferences at the core of human rational thought (Rips, 1988, 1994), the Modus Ponens and the Disjunctive Syllogism, are automatically and unconsciously processed. Furthermore, we investigated whether these inferences require a fully explicit representation of both premises or they can also be triggered by minor premises that are not consciously perceived. In Experiment 1, the major premises (P_1) of various valid or invalid inferential schemata were explicitly displayed and retained in memory, but the corresponding minor premises (P_2) were very rapidly displayed and masked (see also Naccache & Dehaene, 2001b). Participants were not alerted to the presence of a minor premise or to the possibility that deductive inferences could be drawn during the task. In a post-experimental debriefing session, subjects reported that they did not notice any P_2. Thus, according to their self reports, two trials sharing the same first premise (for instance, "if there is a 3 then there is an 8") and the same target (for instance, "8") but with different second premises (for instance, "3" and "5") were perceived as identical trials, since P_2 was not noticed. Surprisingly, when the unnoticed second premise corresponded to the antecedent of the conditional ("3" in the example above), participants pre-activated the number "8", as shown by a positive priming effect on an ensuing number evaluation task displaying the same number ("8"). This finding suggests that Modus Ponens inferences are not only automatic but can be carried out unconsciously and can be triggered by unnoticed premises. Intriguingly, the Modus Ponens was the only schema that showed these properties, when triggered by a direct match on the rule's antecedent. No priming effects were observed neither for the Disjunctive Syllogism nor for any other set of premises allowing valid or invalid inferences (e.g., Affirmation of the Consequent, Denial of the Antecedent, disjunctive fallacies, Modus Tollens). Thus, we suggest that the observed priming effect did not depend exclusively on non-logical associations between the arguments of the major premise, but on some peculiarity in the way the brain deals with the Modus Ponens logical structure.

Findings from Experiment 1 were replicated and extended in three follow-up experiments. First, we checked whether the participants were, indeed, unable to access the minor premises as they reported in Experiment 1. Experiment 2 suggested that

some information about P_2 was actually available to some participants, at least when their attention had been directed to P_2 by telling them about the masked prime. Given the results of Experiment 2, it was not possible to fully establish whether the effect on Modus Ponens required *pre-conscious* access to P_2 or, alternatively, whether it would also have been present in the case of a *subliminal* processing of P_2 (Dehaene et al., 2006). Accordingly, we ran a third Experiment in which we focused on the subset of participants for whom subliminal processing of P_2 had been demonstrated. The priming effect for Modus Ponens was also confirmed in Experiment 3. Finally, in Experiment 4, we explored whether relations other than those involved in propositional logic would produce a priming effect similar to the one observed for Modus Ponens. In particular, we explored spatial relations (e.g., "*a* before *b*") and sentences with quantifiers (e.g., "all *a* with *b*"). By contrast with the priming effect for Modus Ponens, which was also replicated in Experiment 4, we could not observe a priming effect for any of the other types of relations.

These findings shed new light on our understanding of human deduction. First, they suggest that Disjunctive Syllogism is not automatic, or is less automatic than the Modus Ponens, a conclusion that is partially at odds with the predictions made according to Mental Logic Theory (Braine & O'Brien, 1998). This theory proposes that both inferences are part of a direct routine of reasoning, each requiring the activation of only one direct reasoning schema. Consequently, the theory correctly predicts the automaticity of the Modus Ponens, but also the theory wrongly predicts the automaticity of the Disjunctive Syllogism.

Second, the results suggest that subliminal processing of conditional premises is unidirectional, going from the antecedent to the consequent. The Affirmation of the Consequent fallacy (i.e., "If p then q; q; therefore p") is not automatic, or is less automatic than the Modus Ponens. This is at odds with the predictions made by the Mental Model Theory (Johnson-Laird & Byrne, 1991; Rader & Sloutsky, 2002), which assumes that the initial understanding of a conditional rule consists in the mental representation of a possibility, in which the antecedent and consequent coexist, with no directionality assumptions. If confirmed, these results would prompt a revision of those influential cognitive theories of human deductive reasoning.

Third, many theories have suggested that a unique reasoning machinery deals with all basic deductive inferences in essentially the same way (Braine & O'Brien, 1998; Johnson-Laird, 2006). Instead, it seems that one important and very common inference, the Modus Ponens, behaves in a *qualitatively different* way from all the other inferences we tested. Contrary to other inferences, Modus Ponens is arguably automatic and, as long as the conditional major premise is encoded in WM, it can be triggered unconsciously by a minor premise that is below the threshold of detection. This finding is in keeping with other recent findings focused on the neural basis of human deduction that suggest that human reasoning might be a diversified, componential ability, where apparently similar logical steps can be carried out in different ways and by different neural circuits (Goel, 2007; Prado et al., 2010; Reverberi et al., 2010, 2012; Reverberi, Shallice, D'Agostini, Skrap, & Bonatti, 2009). Similarly, our findings suggest that early stages of processing do not trigger fallacious arguments, such as the Affirmation of the Consequent, even though the same fallacious arguments might be triggered by later, and explicit, stages of reasoning. This observation is consistent with recent studies showing that even when people give illogical, belief-biased responses, for example when solving categorical syllogisms, they are nevertheless able to detect that their response is not fully warranted (De Neys & Franssens, 2009).

The present findings finally suggest a straightforward way to reconcile the contrasting pieces of evidence concerning the automaticity of the Modus Ponens. On one hand, all the studies supporting the automaticity of the Modus Ponens (including Lea (1995), Rader & Sloutsky (2002), and the present one) tested Modus Ponens indirectly. The participants were required neither to draw nor to evaluate a Modus Ponens argument: The dependent variables were indirect, e.g., a recognition memory task or a priming effect on a secondary tasks. On the other hand, the effects suggesting that Modus Ponens is *not* automatic (including the modulation of Modus Ponens by additional premises, e.g., Byrne (1989), Stevenson & Over (1995), and its varying rates in accordance to cognitive ability, e.g., Newstead, Handley, Harley, Wright, & Farrelly (2004)) concern *explicit* Modus Ponens inferences: Participants in those studies were explicitly required to evaluate whether or not a given conclusions followed from a given set of premises. This pattern of apparently contrasting pieces of evidence hints at the possibility that two mechanisms contribute to the Modus Ponens inference: a fast one

automatically triggered by the logical structure alone and a slow non-automatic one that integrates other processes and that is sensitive to previous knowledge about the truth of the premises. The process ending first cues a response, which, depending upon task instructions, available resources, and time, may need to be inhibited in favour of an alternative, less rapidly cued response (see also the "parallel competitive account" in Handley, Newstead, & Trippas, 2011). This view is also consistent with a previous interpretation of the modulation of explicit Modus Ponens by "disabling" premises proposed by Stevenson and Over (1995): additional premises and background knowledge reduce the believability of one or both of the premises of the Modus Ponens. Because the participants doubt the premises, then they also doubt the correctness of the conclusion cued by those premises, and they sometimes do not endorse it. In so doing, they confound the validity of an argument with its soundness[1]. In the present perspective, the fast, automatic, implicit mechanism that cues the acceptance of the Modus Ponens inference is sensitive to its validity alone; the "slow lane" mechanism – by integrating previous knowledge – is also sensitive to soundness, in keeping with that *belief-driven* responses are slower than some *logic-driven* responses reported by (Handley et al., 2011).

Beyond the importance of these results for further improving current theories of human deduction, they also extend our understanding of the cognitive activities that can be performed without consciousness. The existence of subliminal perception has been accepted by the scientific community. What is still a matter of controversy is both the depth of processing of subliminal stimuli and the type of cognitive processes that subliminal stimuli can trigger. Thus, although it is largely accepted that subliminal stimuli can activate lower-level processing (Greenwald, 1992), the possibility that subliminal stimuli can also trigger semantic (Dell'Acqua & Grainger, 1999; Naccache & Dehaene, 2001b) or inferential (e.g., deduction, as in the present study) computations is debated (for a review see Kouider & Dehaene, 2007). However, in recent years, an increasing number of studies have provided evidence of subliminal activation in high-level processing, including in decision making (Soon, Brass, Heinze, & Haynes, 2008), motivation (Pessiglione et al., 2007), emotion (Etkin et al., 2004), and in the extraction of formal features from audio streams (Peña, Bonatti, Nespor, & Mehler, 2002). The

[1] In logic a *valid* conclusion is a conclusion that necessarily follows from a given set of premises (disregarding the factual truth of the premises). A *sound* conclusion is a valid conclusion from factually true premises.

present study indicates that a critical component of deductive reasoning – the Modus Ponens – can be triggered by a non-conscious stimulus and performed without voluntary control. Thanks to this inferential tool, unconscious processing - already proven to be able to extract structural rules (e.g., Peña et al., 2002) - may also be able to use these rules by activating their predictions (i.e., the consequents) whenever the triggering conditions (i.e., the antecedents) are satisfied. In this way, unconscious processing of Modus Ponens may play the same role in unconscious cognition as deduction in explicit thinking, i.e., it may highlight the consequences that follow from previously held pieces of knowledge.

3.7. Conclusions

The present study provides evidence that *Modus Ponens* ("if p then q; p; therefore, q"), unlike any other valid or invalid inference that we tested (including Modus Tollens and Disjunctive Syllogism), is an *automatic* information processing step that can be performed even when individuals are not aware of a subliminal second premise ("p"). These results suggest that this form of deductive reasoning can be included in the group of high-level cognitive tasks that are performed unconsciously. Second, these findings are not fully consistent with current cognitive theories of human deductive reasoning and, thus, may prompt further developments and revisions.

CHAPTER 4

4. Study 2: Unattended rules cause response conflict

4.1. Theoretical background

The study of selective attention dates back to Wundt and the very beginnings of experimental psychology (Danziger, 1980). Focusing on task-relevant information while ignoring task-irrelevant information is a process that is neither automatic nor effortless (Lu & Proctor, 1995; MacLeod, 1991). Several studies on attention (Navon, 1977), WM (Frank, Loughry, & O'Reilly, 2001), and executive control (Burgess & Braver, 2010) have shown that irrelevant information may interfere with the execution of a task and produce an increase in response time and/or error rates. **Interference** due to task-irrelevant features can occur, for example, during *incongruent* experimental trials, where the irrelevant features suggest responses that are different from those required according to the imperative features. This type of interference is measured by comparing incongruent trials with *congruent* trials, or trials in which the irrelevant features suggest the same responses as the imperative features. Alternatively, interference can be measured by comparing incongruent trials with *neutral* trials, or trials in which no irrelevant features are associated with the responses (Jonides & Mack, 1984). Neutral trials also allow for measuring facilitation effects in the form of decreased latencies and/or increased accuracy by comparing them with congruent trials.

Interference has been demonstrated to have effects in a variety of tasks, the most common being the *Stroop tasks* (Stroop, 1935). In Stroop tasks, participants must name the ink colour while ignoring a written word that spells the name of a different colour. In a single-word Stroop task (Dalrymple-Alford & Budayr, 1966), incongruent trials consist of a word that names a colour different from the colour in which the word is written ("RED" written in blue). During congruent trials, the two features match ("RED" written in red). During neutral trials, the words are either unrelated to colours ("TEN" written in red), non-words ("ORT" written in red), or strings of characters ("XXX" written in red).

Stroop interference is the increase in response latency and errors during incongruent trials compared with neutral trials (and, in some studies, with congruent trials, e.g., Hamers & Lambert, 1972). Some studies have also reported either Stroop facilitation (Dalrymple-Alford, 1972) or reduced interference (Sichel & Chandler, 1969) when comparing congruent trials with neutral trials.

The Stroop effect arises from an irrelevant feature (the meaning of the word) triggering a well-established S-R association (word-reading). Other well-known interference effects (e.g., Navon effect, Simon effect, flanker effects) occur in response to information conveyed by perceptual stimuli that simultaneously or shortly precede the imperative stimulus.

Some studies have also found interference effects due to internal representations of previously practised task rules. For example, in studies involving *task switching*, a drop in performance is observed during trials in which the task changes compared with trials in which the task repeats (Allport et al., 1994). Finally, in *reasoning tasks*, well-established previous knowledge concerning a putative valid or invalid conclusion can strongly interfere with whether a participant accepts or rejects a conclusion (e.g., Klauer, Musch, & Naumer, 2000).

In all of these effects, well-learned associations, rules, concepts firmly encoded in long-term memory, or information that subjects are explicitly instructed to represent, cause interference. To the best of our knowledge, there are no studies to date that address the ability to focus on a relevant task rule while ignoring or inhibiting an irrelevant one when the rules have just been encoded in WM and the relevant rule is promptly specified (i.e., when a rule is encoded for the current trial only but it is immediately marked as irrelevant and subjects are encouraged to forget it).

We devised a simple reasoning task dealing with biconditional antecedent-consequent (or condition-action) rules ("if X occurs, then do Y; otherwise, do Z."). The **aim** of the present study was to test whether an irrelevant, unpractised, and unattended rule encoded in WM could be triggered involuntarily by target stimuli matching the condition in the rule and thus cause interference effects in reasoning.

4.2. Methods

Twenty right-handed graduate and undergraduate students from the University of Milano-Bicocca (13 females; mean age 23.5 years; age range 19–27) participated in this experiment. They completed the procedure in approximately 30 minutes, including the time it took to give instructions.

Figure 4.1 illustrates the structure of the task, with examples of the six experimental conditions originating from a fully orthogonal 2 × 3 within-participants factorial **design**. Factors included the consequents of the rules (same consequent *vs.* different consequent) and the type of minor premise (from now on, *trigger*: a figure that either matched the antecedent of the cued rule [baseline$_1$ trials], the antecedent of the uncued rule [critical trials], or neither of them [baseline$_2$ trials]).

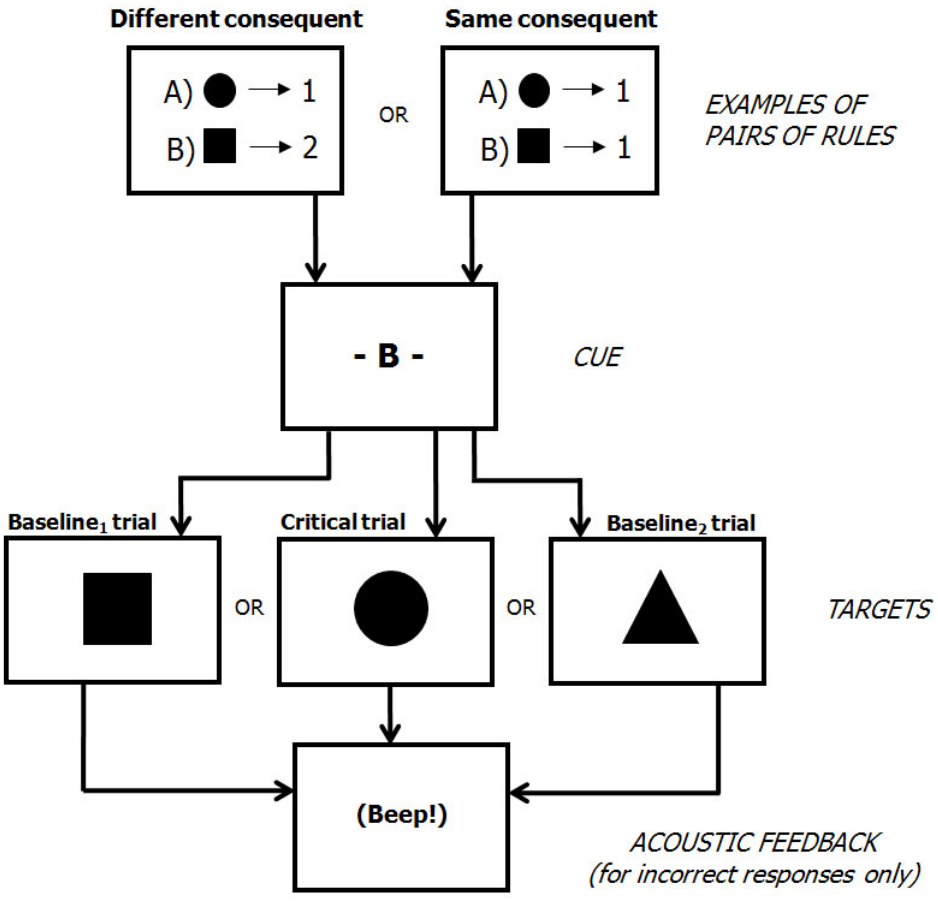


Figure 4.1 The experimental design, with examples of each type of trial. The two rules could present the same (pair of rules on the right) or a different (pair of rules on the left) consequent. The cue indicated which rule applied on the current trial and thus which one was irrelevant and could be forgotten. During *baseline$_1$* trials, the target figure matched the antecedent of the cued rule, in *critical* trials the target figure matched the antecedent of the uncued rule, and in *baseline$_2$* trials the target figure did not match any of the figures in the rules. An acoustic feedback was provided when the response was incorrect.

In the trials, we presented all possible combinations of rule antecedents (triangle, square, or circle), consequents ("press key 1" or "press key 2"), cues (A or B), and triggers (triangle, square, or circle). Trials in which the two rules had the same antecedent were omitted; therefore, six combinations of rule antecedents were used instead of nine, yielding a total of 144 experimental trials (6 combinations of antecedents × 4 combinations of consequents × 2 cues × 3 triggers). Twenty-four trials, which were all different from one another, were distributed across each cell of the design.

Before the experiment began, the participants read self-paced instructions displayed on a computer, and two blocks of practice trials were administered. The first section of the instructions and the first practice block trained the participants to interpret each task rule biconditionally: "If you see [figure], then press [numbered key]; otherwise, press [alternative numbered key]". The practice block displayed only one rule at a time; therefore, no cue was presented. After the rule, a trigger figure appeared, and the participants had to press the appropriate key (with accuracy feedback). The practice block ended as soon as the participant provided five correct responses out of the last six trials. All participants easily learned how to use a rule in a few practice trials (mean = 5.9, indicating that most participants gave five correct responses to the first five or six practice trials).

The second section of instructions detailed the two-rule structure of the experimental trials and explained how to interpret a cue. Participants were explicitly told that only the cued rule should govern responses; the uncued rule was irrelevant and should be ignored. These instructions were followed by five practice trials, which were randomly selected from the 144 experimental trials.

During each experimental trial, two rules (A and B) appeared on the screen. When participants were sure that they would remember the rules and wished to go ahead, they nodded to the experimenter, who pressed a key to proceed with the trial. A cue was displayed (A or B, balanced across trials) indicating which rule should be applied to the trigger. Then, the trigger figure was shown and participants were asked to apply the cued rule and respond accordingly (see Figure 4.2).

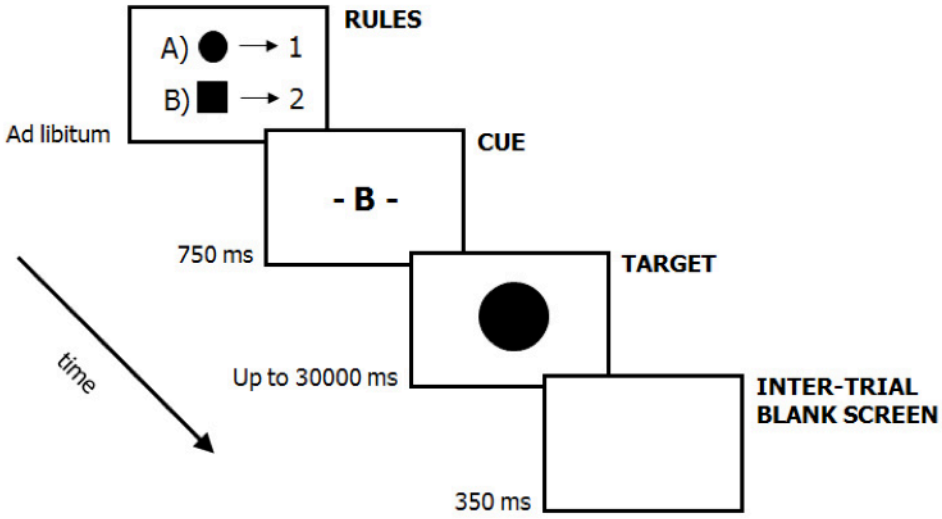


Figure 4.2 An example of experimental trial (objects in the rules, cue, and target are not to scale). At the beginning of each trial, two rules (A and B) were presented; participants had to memorise both rules and to signal to the experimenter when they were ready to proceed. A cue (A or B) was displayed for 750ms and then the target figure appeared. Participants had to respond as fast as possible according to the rule that was active in the current trial.

The experiment was programmed using the E-Prime™ software (Psychology Software Tools, Inc., Pittsburgh, PA, US) and was administered on a personal computer equipped with a 17" CRT colour monitor. Participants sat 60cm away from the computer screen in a dimly lit room and responded by pressing the left or right button (numbered 1 or 2, with the order counterbalanced across participants) of an analog computer mouse with their right index and middle fingers, respectively.

4.2.1. Critical contrasts

During *baseline$_1$ trials*, the trigger figure matched the antecedent of the cued rule (e.g., rule: "if circle, then press 1"; trigger: "circle"; conclusion: "press 1"); thus, determining the correct response required a Modus Ponens inference. During *baseline$_2$* and critical *trials*, the trigger figure did not match the antecedent of the cued rule (e.g., rule: "if circle, then press 1"; trigger: "square"; conclusion: "press 2") and required a Denial of the Antecedent inference that, for biconditional syllogisms, is slower and more difficult than the Modus Ponens inference (Grosset & Barrouillet, 2003). Therefore, baseline$_1$ trials should have elicited responses that were easier and faster than those elicited by baseline$_2$ trials. The logical structure of the responses to critical trials was the same as the structure of responses to baseline$_2$ trials; however, in *critical trials*, the

trigger matched the antecedent of the uncued rule. If the uncued rule is not properly excluded by selective attention, and if Modus Ponens inferences are automatically triggered whenever a set of premises matches the corresponding logical schema as some scholars claim (Braine & O'Brien, 1998), the uncued rule should affect responses in a predictable manner. When the two rules have the same consequents, the uncued rule suggests the *inappropriate* response to the critical trigger (e.g., cued rule: "if square, then press 2"; uncued rule "if circle, then press 2"; trigger figure: "circle"; correct conclusion: "press 1" [Denial of the Antecedent]; conclusion suggested by the irrelevant rule: "press 2" [Modus Ponens]), possibly causing *interference* with the correct response, in which case the correct Denial of the Antecedent response would be harder and slower than in the baseline_2 condition. When the two rules have different consequents, the uncued rule suggests the *appropriate* response (e.g., cued rule: "if square, then press 2"; uncued rule "if circle, then press 1"; trigger figure: "circle"; correct conclusion: "press 1" [Denial of the Antecedent]; conclusion suggested by the irrelevant rule: "press 1" [Modus Ponens]), possibly causing *facilitation* of the correct response, in which case the correct Denial of the Antecedent response would be easier and faster than in the baseline_2 condition.

We normalised the participants' mean latencies and accuracies with respect to the corresponding baseline_1 trials and used **normalised differences** between the conditions of interest (critical trial or baseline_2) and baseline_1 for the analyses. For baseline_2 trials, the normalised values captured the proportional increased cost of Denial of the Antecedent inferences plus random variations. For critical trials, the normalised values capture the increased cost of Denial of the Antecedent inferences, random variations, and interference caused by the uncued rule. Hence, the difference between normalised critical and baseline_2 trials is a pure measure of the effects of the irrelevant rule on the responses, as follows:

Facilitation for correct response latencies

$$\Delta\text{RT}_{\text{different}} = \frac{\text{MeanRT}_{\text{critical different trials}}}{\text{MeanRT}_{\text{baseline1 different trials}}} - \frac{\text{MeanRT}_{\text{baseline2 different trials}}}{\text{MeanRT}_{\text{baseline1 different trials}}}$$

Interference for correct response latencies

$$\Delta\text{RT}_{\text{same}} = \frac{\text{MeanRT}_{\text{critical same trials}}}{\text{MeanRT}_{\text{baseline1 same trials}}} - \frac{\text{MeanRT}_{\text{baseline2 same trials}}}{\text{MeanRT}_{\text{baseline1 same trials}}}$$

Facilitation for accuracy rates

$$\Delta \mathrm{ACC}_{\mathrm{different}} = \frac{\mathrm{ACC}_{\mathrm{critical\ different\ trials}}}{\mathrm{ACC}_{\mathrm{baseline1\ different\ trials}}} - \frac{\mathrm{ACC}_{\mathrm{baseline2\ different\ trials}}}{\mathrm{ACC}_{\mathrm{baseline1\ different\ trials}}}$$

Interference for accuracy rates

$$\Delta \mathrm{ACC}_{\mathrm{same}} = \frac{\mathrm{ACC}_{\mathrm{critical\ same\ trials}}}{\mathrm{ACC}_{\mathrm{baseline1\ same\ trials}}} - \frac{\mathrm{ACC}_{\mathrm{baseline2\ same\ trials}}}{\mathrm{ACC}_{\mathrm{baseline1\ same\ trials}}}$$

Accuracy was scored as the percentage of correct responses in each experimental cell. Before normalisation, RTs were pruned by eliminating participant's correct responses when they exceeded the mean RT of that participant in that condition by more than two SDs. Fewer than 2% of the trials were eliminated.

The normalisation procedure effectively negated the possible effects of differences in the WM load imposed by the rules in the two types of trials. Normalisation also removed the effects of differences in individual participants' absolute speed and accuracy, the rate of random errors, and the rate of errors caused by failures in perceiving, encoding and maintaining the rules, cues, or triggers. Normalisation allowed us to obtain a conservative measure of the effects of irrelevant rules on performance.

According to the critical contrast computations, if participants do not perfectly inhibit the uncued rule, facilitation effects should occur during *different consequent* trials (because the cued and the uncued rules suggest the same response), and interference effects should occur during *same consequent* trials (because the uncued rule suggests a different response than the one required according to the cued rule), with negative $\Delta \mathrm{RT}_{\mathrm{different}}$ and $\Delta \mathrm{ACC}_{\mathrm{same}}$ values and positive $\Delta \mathrm{RT}_{\mathrm{same}}$ and $\Delta \mathrm{ACC}_{\mathrm{different}}$ values.

4.3. Analyses and results

The mean normalised RTs and accuracies for the four conditions involved in the critical comparisons and the respective deltas are reported in Table 4.1.

Data were analysed using the nonparametric Wilcoxon exact test and all tests were one-tailed because our predictions specified the expected directions of the effects. The effects were in the expected direction for both latency and accuracy. However, the differences in latency and accuracy were significant only for the *same consequent* conditions (ΔRT_{same}: Z = 2.73, p = .003, r = .43; ΔACC_{same}: Z = 3.07, p = .001, r = .49;

$\Delta RT_{different}$: Z = 1.01, p = .157, r = .16; $\Delta ACC_{different}$: Z = 0.13, p = .45, r = .02), indicating the presence of an **interference**, but not of a facilitation, effect.

Table 4.1 RTs and accuracy scores normalised with respect to the baseline$_1$ values for the four conditions involved in the critical comparisons and their Δ values. For the sake of deriving absolute values: different actions condition baseline$_1$: Mean RT = 783ms; Mean Accuracy = .948; same action condition baseline$_1$: Mean RT = 745ms; Mean Accuracy = .948.

		Critical trials	Baseline$_2$ trials	Δ
Normalised RT	Different	.432	.519	-.087
	Same	.444	.325	.119
Normalised Accuracy	Different	-.053	-.061	.008
	Same	-.073	.023	-.096

4.4. Discussion

The results show that an irrelevant rule temporarily encoded in WM causes significant interference. When the irrelevant rules suggested incorrect responses, correct responses were 9.6% lower and 11.9% slower relative to the corresponding baseline condition. Because of the normalisation procedure, this difference can be viewed as a pure measure of the uncued rule's interference. Contrary to other known interference effects, the one observed here cannot originate from well-learned associations stored in long-term memory. The two rules were presented and memorised at the beginning of each trial, and the relevant rule was signalled immediately before the trigger stimulus. Furthermore, each combination of possible rules and cues was presented only once in each experimental cell, making learning and the transfer of experience from trial to trial impossible in this task. In this study, interference between the rules occurred in WM during the reasoning processes required to decide the response to each trial and it was not caused by the retrieval of responses to previous trials from long-term memory.

The lack of facilitation effects can be explained in terms of the following:

1. A ceiling effect. The task was quite simple; under baseline conditions, accuracy rates were high and correct responses were fast. If baseline performance is at the ceiling level, then decline can be observed, but improvement cannot be observed.

2. Interference and facilitation effects are not symmetrical, and one does not imply the other; therefore, it is possible that these effects are related to two different cognitive mechanisms. There are some theoretical proposals to this effect (Lindsay & Jacoby, 1994) and, in previous studies, facilitation is not reported as often as interference (MacLeod, 1991).

Additional studies should disambiguate the two interpretations. If (1) is correct, then facilitation should emerge along with interference when the task is made more difficult (for example, by adding a concurrent memory load or requiring more complex logical inferences). If (2) is correct, making the task more complex should result in increased interference but not necessarily in increased facilitation effects.

Our findings are **important** for many theoretical fields of cognitive psychology. First, these results improve our understanding of the role of *attention* in reasoning and WM. Not all contents of WM are equally active at any one time (Barrouillet, Bernardin, & Camos, 2004). However, if selective attention is focused on some content of WM, other content still has a residual effect on performance, causing errors and delays.

Second, our findings corroborate the hypothesis that some rule-based inferential steps are *automatic* (Braine & O'Brien, 1998). In our study, delays due to interference could have only derived from participants involuntarily drawing Modus Ponens inferences in response to the irrelevant rules. There have been previous proposals and empirical studies concerning whether Modus Ponens inferences are automatic (Reverberi, Pischedda, et al., 2012); our findings support the hypothesis that Modus Ponens inferences are automatic.

Third, rule-based mistakes or capture errors (Rasmussen, 1982; Reason, 1990) are an important class of *human errors*. These errors occur when performance that should be driven by certain rules is actually driven by other, stronger rules. Usually, the interfering rules are the result of previously well-learned and practised habits. Here, we show that capture errors can occur even when the interfering rule is transient and unpractised.

Finally, the task we devised is short, easy, and suitable for elderly people and for some clinical populations (as shown by pilot testing that we performed with elderly subjects and with Parkinson's disease patients). Accordingly, this task could be useful

for *testing* the ability to control and inhibit contents of WM in populations that are at risk of developing inhibitory deficits.

4.5. Conclusion

The present study showed that not just well-learned and practised task rules or associations temporarily maintained for successive use, but also unpractised rules transiently encoded in WM and promptly marked as irrelevant can have a detrimental effect on performance when they are involuntarily triggered by target stimuli that match their antecedents. These results improve our understanding of the role of attention in reasoning and corroborate the hypothesis that some rule-based inferential steps are automatic.

CHAPTER 5

5. Study 3: Hierarchies of control in Prefrontal Cortex

5.1. Theoretical background

In daily life, humans rely on different types of **rules** to define complex plans and orchestrate their actions in order to achieve specific goals (Bunge & Wallis, 2008). In the simplest situation, it is sufficient to perform an action (e.g., cross the road) in response to a specific stimulus (e.g., a green traffic light) that is linked to that action (e.g., by the rule "if the traffic light is green, you can cross the road"). However, in more challenging conditions, it is necessary to apply rules organised in hierarchies, where high-level rules influence the application of lower-level rules (e.g., if the traffic policeman directs the traffic, the rules that link a specific gesture of the traffic policeman to an action will take priority over the usual associations between the colour of the traffic lights and the appropriate response).

Previous studies, either on monkeys or humans, have demonstrated that a wide frontoparietal network is involved in rule representation and application (Asaad et al., 2000; Bengtsson et al., 2009; Bode & Haynes, 2009; Bunge et al., 2003; Cavina-Pratesi et al., 2006; Genovesio et al., 2005; Hoshi et al., 1998; Muhammad et al., 2006; Petrides & Baddeley, 1996; Reverberi, Görgen, et al., 2012b, 2012a; Sakai & Passingham, 2003, 2006; Stoet & Snyder, 2004; Wallis et al., 2001; Wallis & Miller, 2003; White & Wise, 1999; Woolgar, Hampshire, et al., 2011; for more information see Paragraphs 1.2 and 1.3). Recently, growing interest has been devoted to the investigation of how PFC represents rule sets of increasing complexity. Several theories have been proposed suggesting that different areas in PFC are organised along an anterior-to-posterior **gradient** (e.g., Badre & D'Esposito, 2009; Badre, 2008; Christoff et al., 2009; Fuster, 1997; Koechlin et al., 2003; Koechlin & Summerfield, 2007; Petrides, 2005) depending on the level of abstraction of the rules to be applied or on the type of control signal involved. A large body of evidence supports the existence of such a gradient (Badre & D'Esposito, 2007; Charron & Koechlin, 2010; Koechlin, Ody, & Kouneiher, 2003;

Kouneiher, Charron, & Koechlin, 2009; Nee & Brown, 2012; but see, e.g., Duncan, 2001), but the exact feature that defines the hierarchy is debated. For example, in their cascade model of cognitive control, Koechlin and colleagues (Koechlin et al., 2003; Koechlin & Summerfield, 2007; Kouneiher et al., 2009) propose that PFC is organised according to a hierarchy of representations processing different *types of signals* involved in the control of action selection. In particular, more caudal regions, such as premotor (PM) cortex and BA 9/44/45, would control the selection of motor responses or of representations of imminent actions, respectively; instead, more rostral areas, such as BA 46 and FPC (BA 10), would be involved in selecting the appropriate task set and in cognitive branching (the process that "enables a task or a behavioral episode to be interrupted and temporarily maintained in a pending state while another is being performed, and/or to revert back to a pending task or episode following completion of the ongoing one", Koechlin & Summerfield, 2007, p. 233), respectively. More specifically, in *sensory* control, motor actions are selected to respond to perceptual stimuli; in *context* control, appropriate S-R associations are chosen based on context signals; in *episodic* control, the relevant task set is selected according to previous events or to the current internal goals; and in *branching*, a specific task is chosen based on reward expectations. In the experiments devised by Koechlin's group, sensory and contextual control are operationalised as the number of responses required and of tasks to be performed within an experimental block, respectively (with low control when one response or one task is possible and high control when several responses or many tasks have to be performed). Episodic control is defined according to the number of possible S-R associations for each stimulus across blocks (with low control when a stimulus is always associated with the same response and high control when a stimulus is associated with a response on one block and with another response on a different block). Finally, branching is defined as the need to suspend the execution of a primary task until a second task has been performed (responses to the primary task are postponed to the end of second task execution, Charron & Koechlin, 2010).

In contrast, Badre and colleagues (Badre & D'Esposito, 2007, 2009; Badre, 2008, 2008) suggest that the rostro-caudal axis within PFC is hierarchically organised based on the level of *abstraction* of the representations that have to be selected for action. The definition of abstraction provided by the Authors states that "a representation may be

defined as more abstract to the extent that it generalises over specific instances" implying that "a more abstract or superordinate representation comprises a category or class of subordinate representations." (Badre & D'Esposito, 2007, p. 2084) In four experiments, the authors manipulated the level of abstraction of cues necessary to define the appropriate response; in other words, they varied the hierarchical nesting of the features governing action selection. The results showed that caudal regions in PFC, such as PM cortex (BA 6), control concrete representations (e.g., motor responses), while progressively more anterior areas, such as DLPFC (BA 9/46), control more abstract representations.

To sum up, these two theoretical frameworks explain the gradient in PFC in terms either of the temporal proximity of the control signals required or of the level of abstraction of the representations to be selected. Further ideas on the functional organisation of PFC derive from recent studies suggesting that the recruitment of more anterior areas in PFC depends rather on task difficulty (Crittenden & Duncan, 2012) or on task-dependent maintenance demands (Reynolds et al., 2012). Crittenden and Duncan (2012) used a discrimination task (selecting the shortest in a set of lines) and varied the difficulty of the task by either increasing the set size (four *vs.* eight lines), reducing the difference in length between the target and the other lines, or changing the S-R mapping compared with the normal mapping. According to the theories previously described, all the experimental conditions should activate only caudal regions in PFC, since only one task requiring low-level control information has to be performed. However, the results showed that more anterior areas in PFC were activated when the *task difficulty* was higher compared with the baseline condition. Reynolds and colleagues (2012) proposed yet another hypothesis (*adaptive context maintenance* hypothesis) to account for the recruitment of different areas in PFC. They suggested that both posterior and anterior regions in PFC modulate their activity depending on the specific maintenance demand of the task (i.e., neural activity is transient when contextual information is relevant for a single trial and sustained when the same information has to be maintained across multiple trials). In an elegant experiment, they varied the level of abstraction of the representations (defined as the number of cues to be considered in order to respond) and the time interval in which a cue was active (for a single trial *vs.* for a block of trials) independently and directly tested the predictions of

the three theories. The results supported Reynolds' hypothesis, since both anterior and posterior regions were active in all conditions considered and the neural activity of each area was transient when contextual information was relevant for a single trial and sustained when it applied to a block of trials.

All the experiments described so far analysed brain activity during rule application. Other studies, instead, investigated the pure *representational* phase, when the rules have to be maintained and before they are applied. For example, Nee and Brown investigated rule representation when information relative to high- or low-level context was updated (2012a) or when either high-level context information was represented in isolation or it was integrated with low-level context information (2012b). Updating of high-level context information was associated with increased activation in DLPFC (BA 46) and basal ganglia (striatum and pallidum), while PM area (BA 6), ACC (BA 32), PPC (BA 40) and precuneus were active during low-level context information updating. Information about high-level context was present in OFC (BA 10/11) and ventromedial PFC (VMPFC, BA 11) when represented in isolation and in sensorimotor cortex (BA 3 and BA 4), PM area (BA 6), IFJ (BA 8), mid-DLPFC (BA 9/46), FPC (BA 10), mid-cingulate cortex, supplementary motor area (SMA, BA 6), precuneus, IPS (BA 40), and temporoparietal junction (TPJ, BA 39) when integrated with low-level context information.

Taken together, growing interest is currently devoted to the issue of a possible functional specialisation of different areas within PFC involved in representations of rule sets of increasing complexity. Still, it is currently unclear whether those regions systematically encode qualitatively different types of information. In the **present study**, we investigated whether the different features defining a complex rule set are represented in different brain areas depending on the level of control they enforce, focusing on the encoding and maintenance of rules in WM rather than on their application. In addition, we considered both high- and low-level rules when represented either in isolation or together with the alternative type of rule.

5.2. Methods

5.2.1. Participants

Twenty subjects (11 females) participated in this experiment in exchange of monetary payment. All were right-handed (score > 40 at the Edinburgh Handedness Inventory – EHI, Oldfield, 1971), native German speakers, had normal or corrected-to-normal vision, no neurological or psychiatric history, and no anatomical brain abnormalities. Four participants were discarded: two because of excessive head movement during scanning, one because of poor performance in the task, and one because of technical problems during data acquisition. We report results from the 16 remaining subjects (mean age 24.9 years; range 19-30). All the experimental materials were provided in German. The local ethics committee approved the study and all the participants gave written informed consent.

5.2.2. Stimuli and experimental procedure

Participants were required to retrieve, maintain, and apply multiple conditional rules to different target stimuli (Figure 5.1). We used two different types of **rules**, hierarchically organised: *low-level* rules (basic rules, BRs) and *high-level* rules (hierarchical rules, HRs). More precisely, we define rules at high level of control as those that influence the selection and application of other lower-level rules; therefore, high-level rules determined when low-level rules had to be applied. It is important to point out that also different definitions of hierarchies in action control exist (for alternative definitions than the one we used here see Paragraph 5.1 or, e.g., Badre & D'Esposito, 2009).

To avoid possible confounds, we used conditional rules with the same logical form. Each rule was composed of two single rules (Figure 5.1 A). For the BRs, each single rule (e.g., "if there is a banana, then left") associated the presence of an object (e.g., "banana") with a motor response (press either left or right). For the HRs, each single rule (e.g., "if there is a square, then blue") linked the presence of a figure (e.g., "square") with an attentive response (focus on images with a specific background colour, e.g., "blue"). We used eight different single rules (R1-R8) and combined them in pairs to create four compound rules: BR1, BR2, HR1, and HR2 (Figure 5.1 A). The same procedure was

applied in previous studies of our group (Reverberi, Görgen, et al., 2012a, 2012b). We used two different pairs of objects, figures, and colours in all possible combinations to define the rule sets for each participant. For the first subgroup of subjects (N = 10), the BRs were composed of the following single rules: R1, "if there is a banana, then left"; R2, "if there is a guitar, then right"; R3, "if there is a banana, then right"; R4, "if there is a guitar, then left". For the second subgroup (N = 6), banana/guitar were replaced by car/chair, respectively. The HRs were composed using various combinations of figures and colours for different subjects, for example: R5, "if there is a hexagon, then yellow"; R6, "if there is a square, then blue"; R7, "if there is a hexagon, then blue"; R8, "if there is a square, then yellow". The figures hexagon/square and the colours yellow/blue were replaced by the alternative pairs circle/star and red/green, respectively. Overall, for nine subjects the HRs were created using hexagon/square and for the remaining seven subjects using circle/star; similarly, for nine subjects the relevant colours were red/green and for the other seven were yellow/blue. Each single rule was associated with two visually unrelated cues (Figure 5.1 B). Participants had learnt the associations between the cues and the rules in a separate training session (see below). The associations were randomised across subjects so that each participant learnt different cue-rule associations.

The images in the target were coloured 3D pictures of four different objects: banana, guitar, car, and chair (courtesy of Michael J. Tarr, Canergie Mellon University, http://wiki.cnbc.cmu.edu/TarrLab). For each subject, only two of the objects were used in the rules. The target screen was composed of three images embedded in a shape defined by a white border (Figure 5.1 C). Each picture was presented on a coloured background. During the target phase, participants had to apply the rules in a specific order: the HR active in the current trial had to be applied first. When the relevant figure was present in the target screen, subjects had to consider only the objects with the same background colour as the one stated in the HR. Then, if the object in the matching background colour was relevant, subjects had to press the key specified by the BR active for that trial. The experiment was devised and administered via MATLAB® (The Mathworks, Inc., Natick, MA, US) using the Cogent 2000 toolbox (Functional Imaging Laboratory and Institute of Cognitive Neuroscience, University College London, London, UK).

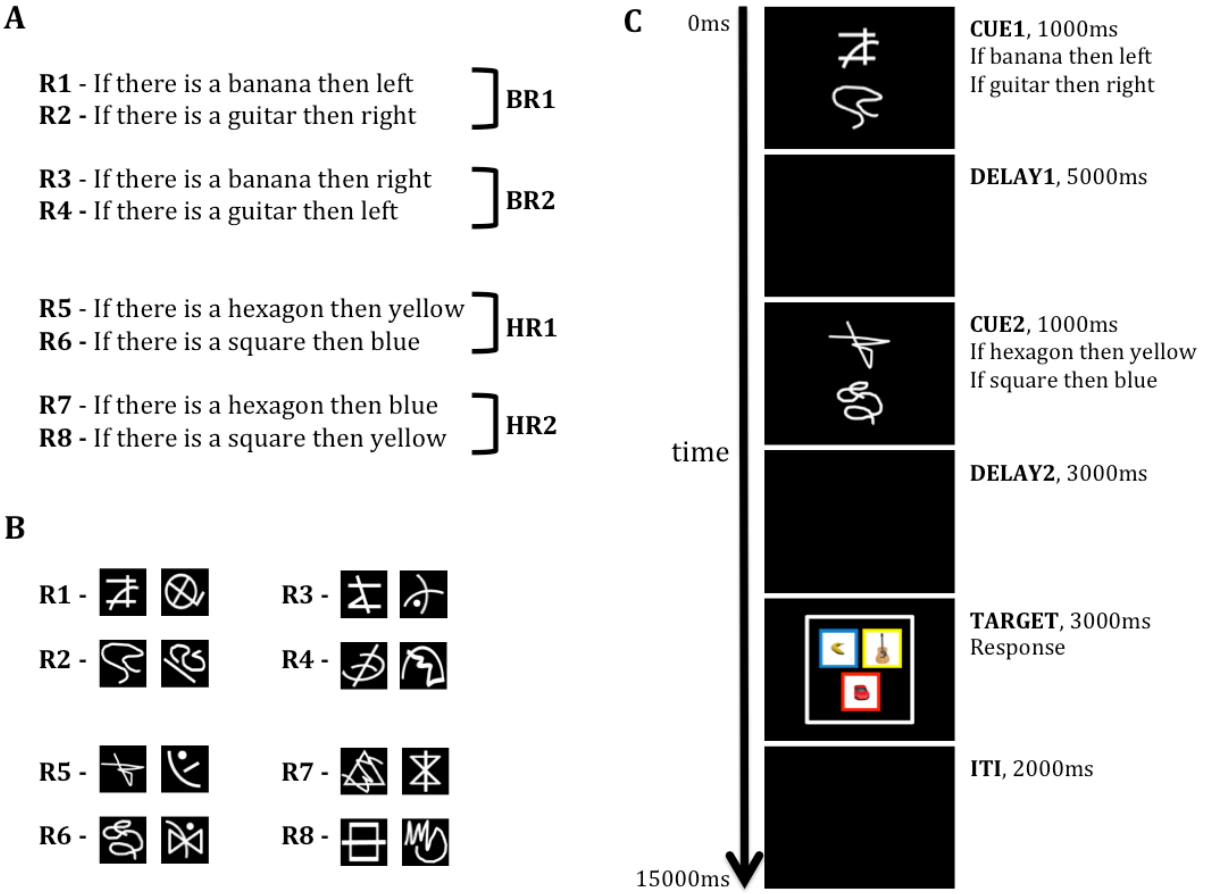


Figure 5.1 Schema of the experimental paradigm. From eight different conditional rules (R1–R8) four complex rules were produced, either low (BR1 and BR2) or high (HR1 and HR2) level rules (A). Each of the eight single rules used in the experiment was associated with two visual cues (B). Timeline of the experiment (C). At the beginning of each trial, two cues were presented, indicating which BR (or HR) had to be applied in the current trial. After a delay of 5s another pair of cues, specifying which HR (or BR) was active for the current trial, was shown. A second delay of 3s followed the second cue presentation; then the target image was presented. Participants had to apply the active rules to the target stimuli and derive as fast as possible the appropriate response(s).

At the beginning of each **trial** (for an example see Figure 5.1 C), two cues were presented for 1s, one above the other aligned vertically to the centre of the screen (cue1). After the cue screen, a delay period of 5s followed (delay1) and then another similar pair of cues was presented for 1s (cue2). A second delay period of 3s (delay2) followed cue2 screen and then the target was shown for 3s. The cues informed subjects about the rule set active on the current trial. During delay1 participants retrieved and maintained one of the compound rules (e.g., BR1), while during delay2 they retrieved a second compound rule belonging to the alternative type (e.g., HR1) and maintained both rules. For half of the trials cue1 comprised cues coding for one of the BRs and cue2 showed cues associated to one of the HRs; in the other half of trials the opposite was true. Subjects had to evaluate first the active compound HR and to infer which colour

was appropriate; then they evaluated the active compound BR and pressed the key(s) associated with the object(s) with a matching background colour. Multiple responses were possible (if more than one relevant object was displayed on the matching colour or if the same object was associated with two different keys, as in some catch trials – see below); in this case, participants could press the keys in any order. Given the active rules and the target shown in a particular trial, four *different outcomes* were possible: subjects had to apply 1) both rules (Both, when a relevant figure was present and the background of a proper object matched the colour in the rule), 2) only the BR (BR, when the figure was not relevant but a pertinent object was present), 3) only the HR (HR, when the figure was relevant but either the background colour of none of the image was appropriate or an irrelevant object had the matching colour), or 4) none of the rules (None, when both the figure and the objects were irrelevant). When it was not possible to apply the BR (HR or None conditions), participants pressed a different key meaning "no response". Subjects had to respond as quickly as possible pressing the inner button of either the left or the right response box using the index finger of the respective hand. The "no response" button was the second button on the right response box and participants pressed it using the middle finger of the right hand.

In addition to experimental trials (Exp), some **catch trials** were interspersed throughout the standard trials. We used three different types of catch trials: trials with shorter delays (Short), with an *uncommon* combination of single rules in cue1 (Catch1), or with an uncommon combination of single rules in cue2 (Catch2). In Short trials, both delay1 and delay2 lasted only 2s, in order to ensure that participants immediately retrieved the rules and represented the rule set. In Catch1 trials, the single rules composing the compound BR instructed by cue1 shared either the object (e.g., "if there is a banana, then left" and "if there is a banana, then right") or the motor response (e.g., "if there is a banana, then left" and "if there is a guitar, then left"); when cue1 coded for a compound HR, the single rules composing the rule shared either the figure (e.g., "if there is a square, then blue" and "if there is a square, then yellow") or the colour (e.g., "if there is a square, then blue" and "if there is a hexagon, then blue"). In Catch2 trials the same uncommon combinations described for Catch1 trials appeared in cue2. The uncommon combinations of rules were used to prevent participants to establish a strong association between single rules (e.g., "R1 is always together with R2") and to use strategies to

reduce the cognitive load (e.g., remembering only "If banana, then left" and performing the alternative motor response for the guitar).

During scanning, participants performed 300 trials, divided into 6 runs. In each run, 50 trials (40 experimental, 4 Short, 2 Catch1, and 4 Catch2 trials) were administered in random order. The intertrial interval (ITI) was 2s, and the whole experiment lasted about 73min. Before scanning, participants performed a "refresher session" for about 10min to ensure that they remembered the cue-rule associations. Once in the scanner, participants performed 5 experimental trials to get used to the new setting. These trials were administered for practice purpose only and thus were not further analysed. After scanning, a *questionnaire* was administered to investigate if and what types of strategies were used to perform the task (see Appendix B1).

Before the fMRI scanning, subjects underwent two **training** sessions scheduled on separate days at most three days before scanning. On the first day of training, participants learnt the cue-rule associations; on the second day, they practised the experimental task and received feedback on their accuracy on each trial. Three levels of difficulty were implemented for the training procedure on the second day, by gradually reducing the duration of the trial delays (4s for both delay1 and delay2 in the first level, 3s for both delays in the second level and 2s for both delays in the last level). Overall, the training procedure lasted about 2.5h (mean duration day1 = 70min; mean duration day2 = 84min). Only participants who reached a high accuracy in the final level of the training (at least 12 correct responses in the last 15 trials) were allowed to the fMRI session.

Such a **paradigm** is particularly fruitful since it allows to: 1) independently assess the encoding of high- and low-level rules, 2) evaluate the difference between the encoding of the two types of rules (comparing high- and low-level rule representations during delay1, when only one type of rule is maintained), and 3) investigate rule integration (by comparing independent representations during delay1 and integrated representation in delay2, in which rules at a different level have to be integrated in order to respond).

5.2.3. Image acquisition

fMRI data were collected using a 3-T Siemens Trio scanner (Erlangen, Germany), equipped with a 12-channel head coil. In each of the six scanning sessions, we acquired

376 T_2*-weighted volumes in descending order, using GRE EPI sequences. The images were composed of 33 slices (3mm thick), separated by a gap of 0.75mm. Imaging parameters were as follows: TR 2000ms, TE 30ms, FA 78°, matrix size 64 × 64, and field of view (FOV; i.e., the spatial extent of an image) of 192mm × 192mm, thus yielding an in-plane voxel resolution of $3mm^2$, resulting in a voxel size of 3mm × 3mm × 3.75mm. A T_1-weighted anatomical dataset and magnetic field mapping images were also acquired. Imaging parameters for the anatomical scan were: TR 1900ms, TE 2.52ms, FA 9°, matrix size 256 × 256 × 192, FOV 256mm × 256mm × 192mm, 192 slices (1mm thick), and resolution 1mm × 1mm × 1mm. For the field maps the parameters were the following: TR 400ms, TE 5.19ms and 7.65ms, FA 60°, matrix size 64 × 64, FOV of 192mm × 192mm, 33 slices (3mm thick), and resolution 3mm × 3mm.

5.3. Data analyses

5.3.1. Pre-processing

fMRI data were pre-processed and analysed using SPM8 (Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, London, UK). During pre-processing, the images were realigned and slice-time corrected. Low-frequency noise was removed using a high-pass filter with a cutoff period of 128s (Worsley & Friston, 1995) and an autoregressive (AR) model was fit to the residuals to allow for temporal autocorrelations (Friston et al., 2002). The volumes were neither spatially smoothed nor normalised in order to preserve fine-grained patterns of activation.

5.3.1. Multivariate Pattern Analysis

We applied MVPA to identify which brain areas encode information about rules at different hierarchical levels. To investigate the representation of a specific rule, we considered only fMRI data from *delay1*, since during this time window only one type of rule was encoded; instead, in delay2, the two active rules were represented at the same time. To look for areas coding for the identity of the two types of rules, we defined two distinct **models**. Each model comprised two regressors corresponding to the two pairs of rules: BR1 and BR2 for model1 and HR1 and HR2 for model2. The two GLMs were estimated using an *FIR* model and applying it to the realigned and slice-time corrected

images (Henson, 2004). Each condition was modelled using eight time bins of 2s each. Using this procedure, it is possible to identify where and when information about the rules is first available and if and how representations modify over time (Bode & Haynes, 2009). The time vectors of all regressors were defined using cue1 onset times. Distinct regressors were computed for each of the six runs. The beta parameters estimated by the FIR model were used to perform a whole-brain MVPA using a *searchlight* approach (see Figure 5.2 A). For this analysis, we implemented a 6-fold *leave-one-run-out* CV procedure to obtain a better estimate of the classifier accuracy. A *linear SVM* with a fixed regularisation parameter C = 1 was trained to distinguish between the two rules (BR1 *vs.* BR2 in model1 and HR1 *vs.* HR2 in model2, Figure 5.2 B) using data from five of the six runs available; the classifier performance was then tested on the data from the left-out run (see Figure 5.2 C). The procedure was repeated six times (using each of the six runs as test set once) and the results were averaged. Only data from the FIR time bins 3, 4, and 5 were used, which corresponded to the time window from 4 to 10s after cue onset. Based on the delay of the HRF, this interval reflected only activity related to cue1 and delay1 (see Figure 5.3). For each searchlight sphere (radius = 4 voxels), a measure of DA (with respect to chance level) was obtained. The values were then combined to produce accuracy maps. For each participant, the two accuracy maps (one for each model) resulting from the analyses described above were normalised to MNI space and then submitted to a one-way ANOVA (factor levels = number of bins) to test where in the brain DAs differed from chance level across all participants.

To investigate whether rule representation changes when information about a different rule has to be integrated, we repeated the same analyses on data from *delay2* where both types of rules were encoded at the same time. The procedure was the same as the one described above, except that the time vectors for the regressors were defined using cue2 onset times. We considered only time bins 3 and 4, corresponding to the interval between 4 and 8s after cue2 onset; in this way only the activity corresponding to cue2 and delay2 was analysed (see Figure 5.3). The aim of this analysis was to evaluate if and how the representation of a specific type of rule changes when it is encoded together with a second rule relative to a different level of action control.

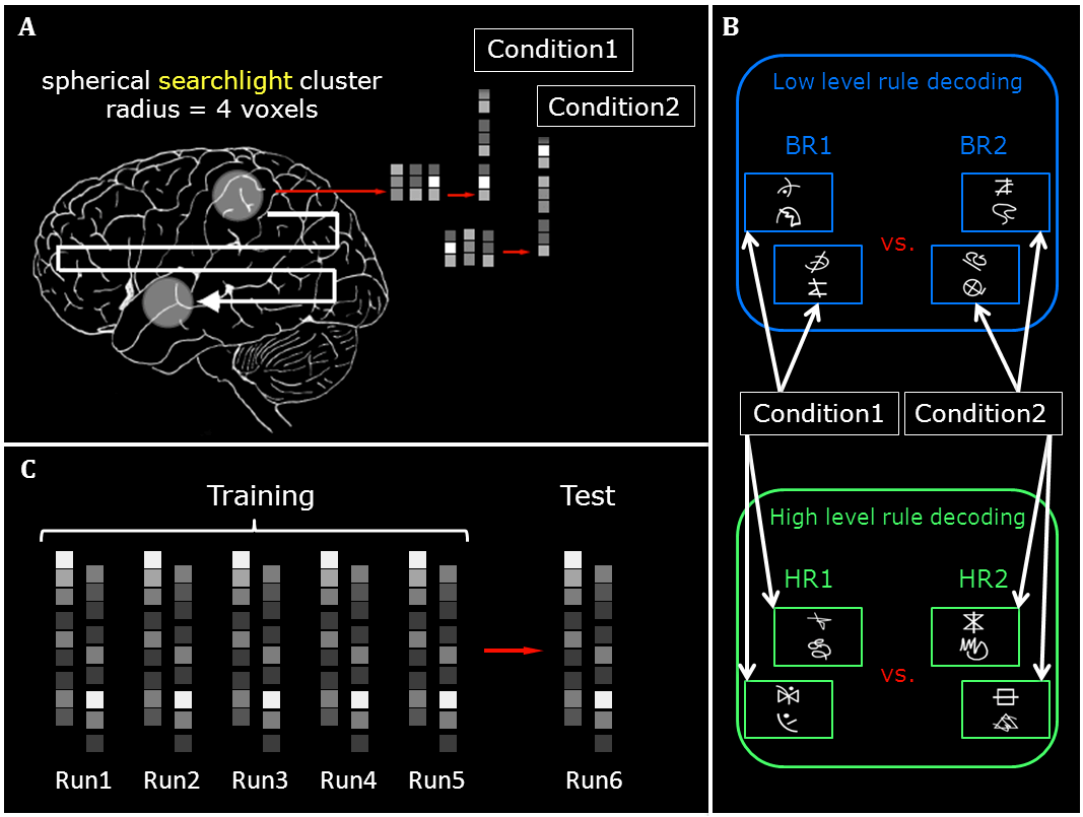


Figure 5.2 Multivariate pattern analysis: extraction of voxel intensity vectors for the two conditions (A). Classification is performed on spherical clusters (radius = 4 voxels) centred on each voxel of the brain. Models used for the two analyses (B): model1 for the BRs (on the top) and model2 for the HRs (on the bottom). Classification procedure (C): pattern vectors of all but one run are assigned to the training data set to train the classifier to distinguish between the conditions; vectors from the left-out run are used to evaluate the accuracy of the classifier.

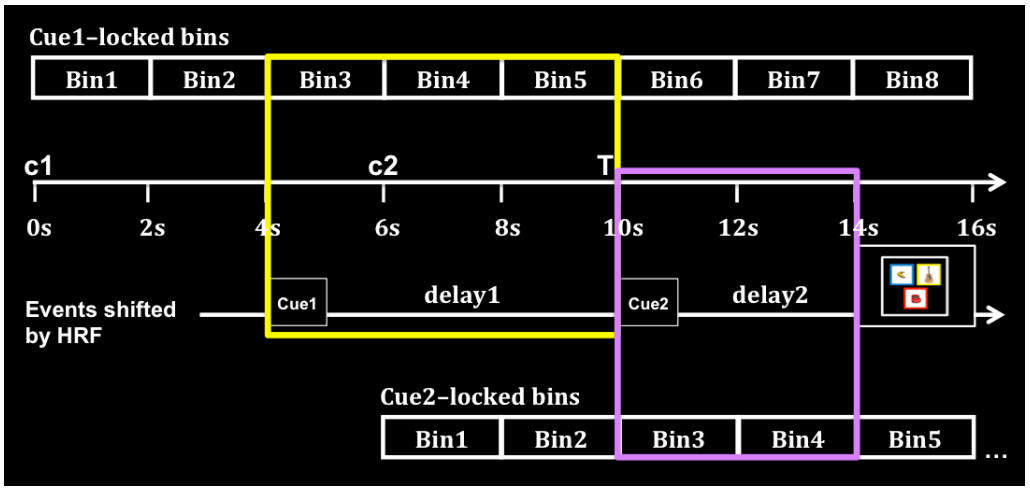


Figure 5.3 FIR models for the analyses performed on fMRI data from delay1 (cue1-locked bins) and delay2 (cue2-locked bins). The FIR models consisted of eight time bins lasting 2s each to model the whole trial duration. The first cue (c1) was presented at the beginning of each trial (onset 0s), followed by a delay of 5s (onset 1s), a second cue (c2, onset 6s), a second delay (onset 7s) and the target (T, onset 10s). To account for the temporal delay of the BOLD signal, we considered only time the bins from the third on, since time bin 3 was the earliest that could reflect cue-related activity, as shown by the timeline representing the events shifted by the first two volumes. In delay1 model, cue1 presentation corresponded to the first second of time bin 3, while delay1 coincided with time bins 3, 4, and 5. In delay2 model, cue2 presentation corresponded to the first second of time bin 3 and delay2 coincided with time bin 3 and 4.

5.4. Results

5.4.1. Behavioural results

During scanning, subjects applied the rules with high **accuracy**. The experimental trials received an average of 94.6% correct responses (SD = 23%, for catch trials see Figure 5.4 A). They were also fast in generating the responses; the first button press was given in about half of the total time allowed for responding (3s). Mean **RT** for the experimental trials was 1642.5ms (SD = 482.8ms, for catch trials see Figure 5.4 B). To assess potential differences in difficulty between the rules, we extracted all trials in which subjects were required to apply a specific experimental compound rule and compared the relative mean accuracies and RTs. No significant difference between any of the four compound rules emerged either for accuracy or RT (accuracy, $F(3,4796) = .75$, $p = .52$; RT, $F(3,4445) = .34$, $p = .79$).

Three types of **catch trials** were also used in the experiment. In Short trials, participants had a mean accuracy of 91.1% (SD = 28%) and an average RT of 1652.8ms (SD = 465.7ms); in catch trials with uncommon combinations of rules (Catch1 and Catch2) the average accuracy was 81.2% (SD = 39%) and 80.5% (SD = 40%) and the mean RT was 1587.1ms (SD = 505.3ms) and 1611.9ms (SD = 530.9ms), respectively. RTs in catch trials were not significantly different from those of experimental trials as well as accuracy in Short and experimental trials. However, accuracy in both Catch1 and Catch2 trials was significantly lower than experimental trials (Catch1, $t(197.4) = 4.70$, $p < .001$; Catch2, $t(408) = 6.88$, $p \ll .001$). This can be due to the performance of the subjects who reported to have applied the strategy to remember only one of the rules for at least one pair; if these participants (N = 7) are excluded, the significance of the difference in accuracy between experimental and catch trials is reduced (Catch1, $t(110.6) = 3.12$, $p = .002$; Catch2, $t(228.6) = 4.77$, $p < .001$). The absence of differences in both accuracy and RTs between experimental and Short trials confirms that participants promptly represented the relevant rule set.

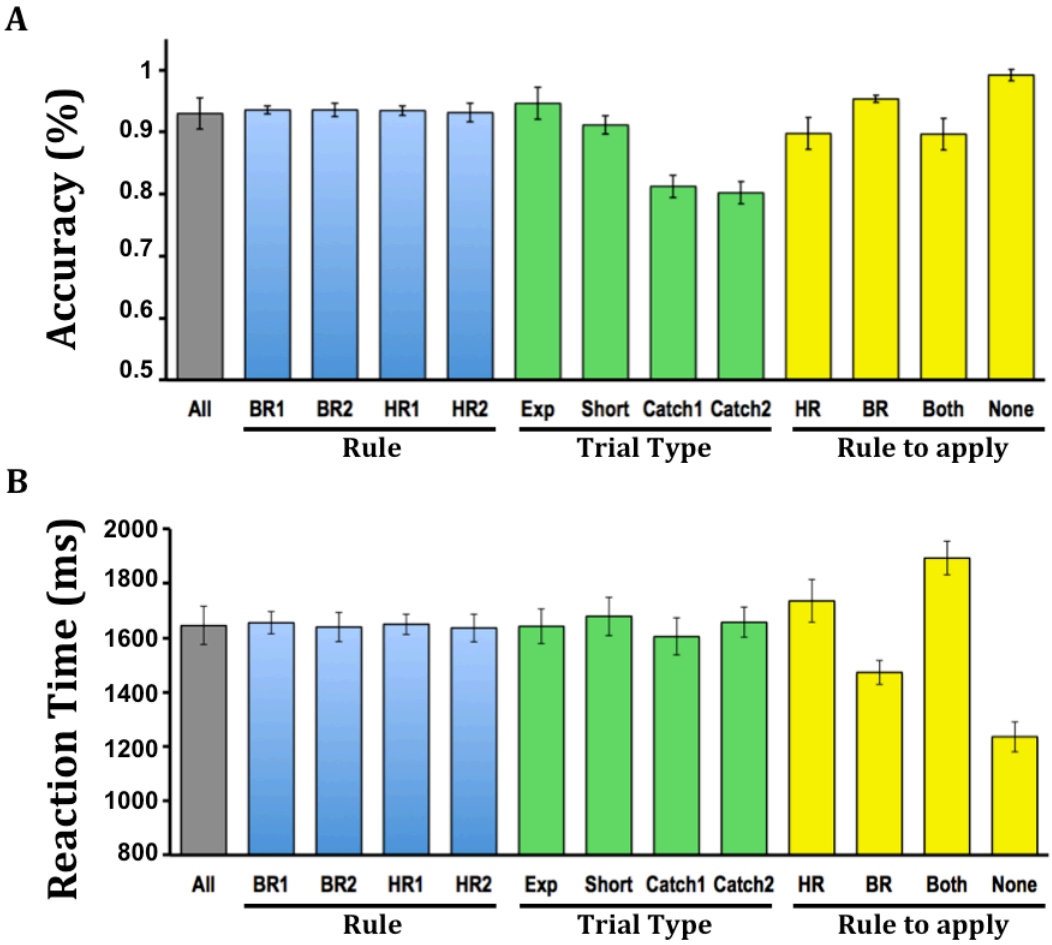


Figure 5.4 Plot of the mean response accuracies (A) and RTs (B) for different trials. All: average across all trial types. BR1 and BR2: basic compound rules; HR1 and HR2: hierarchical compound rules. Exp: experimental trials; Short: catch trials with delay1 and delay2 lasting 2s; Catch1: uncommon combination of rules on cue1; Catch2: uncommon combination of rules on cue2. HR: only the HR applies; BR: only the BR applies; Both: both rule apply; None: neither the BR nor the HR applies. Error bars display the SEM.

On each trial, four **outcomes** were possible, depending on the number and the type of rule that had to be applied. In general, the less the rules to apply, the faster and the more accurate the responses were (see Table 5.1). The effect of the possible outcome on both RTs and accuracy was significant ($F(3,4445) = 287.5$, $p < .001$ and $F(3,4796) = 21.5$, $p < .001$, respectively). *Post hoc* tests for RTs showed a significant difference between conditions for all the pairwise comparisons (all *p*s < .001); accuracies were significantly different between the condition in which either only the HR or both the BR and the HR applied and either only the BR or neither of the rules applied (all *p*s < .001).

Table 5.1 Mean RTs and accuracies for each possible outcome. SDs are shown in parentheses.

Rule	Both	Only HR	Only BR	None
RT (ms)	1862.9 (11.6)	1703.6 (15.9)	1460.6 (9.2)	1244.3 (38.6)
Accuracy (%)	89.7 (0.7)	89.8 (0.9)	95.4 (0.5)	99.2 (2.3)

Differences in RTs and accuracies when the HR had to be applied may be due to the higher cognitive load necessary to process rules at a higher level of control or to the need to inhibit a response when a proper object background didn't match the relevant colour.

5.4.2. Decoding representations of high- and low-level rules

The aim of the first set of decoding analyses was to identify brain areas coding for rules at *different levels* in the hierarchy of action control. The results of the analyses (all $p < .001$ uncorrected at voxel-level and $p < .05$ FWE-corrected at cluster level) are shown in Figure 5.5. For the **BRs**, during time bin 3 (time window from 4 to 6s after cue1 presentation), information was encoded in occipital (middle occipital gyrus, MOG, BA 17/18; left hemisphere peak: -18, -88, -5, accuracy 67%; right hemisphere peak: 36, -76, 1, accuracy 73%), right parietal (IPL, BA 40; peak: 48, -28, 37, accuracy 71%), and left temporal (inferior temporal gyrus, ITG, BA 20; peak: -51, -73, -5, accuracy 67%) areas. During time bin 4, it was possible to decode the rules in the right parietal cortex (IPL, BA 40; peak: 54, -31, 31, accuracy 65%; superior parietal lobule, SPL, BA 7; peak: 45, -52, 58, accuracy 67%). Finally, in time bin 5, information was represented only in right prefrontal cortex (middle frontal gyrus, MFG, mainly BA 47; peak: 48, 47, -8, accuracy 67%). During the whole time window in which the rule was represented (cue1 and delay1), it was possible to decode the BRs from occipital (MOG, BA 17/18; left hemisphere peak: -30, -97, -14, accuracy 82%; right hemisphere peak: 36, -91, 13, accuracy 80%), right temporal (ITG, BA 20; peak: 39, -58, -5, accuracy 85%), right prefrontal (MFG, BA 47; peak: 39, 50, -5, accuracy 84%; ACC, BA 32; peak: 6, 44, 19, accuracy 83%), and particularly parietal (right IPL, mainly BA 40; peak: 48, -25, 40, accuracy 90%; left SPL, BA 7; peak: -30, -61, 58, accuracy 84%; right IPS, BA 7; peak: -30, -43, 37, accuracy 83%) areas.

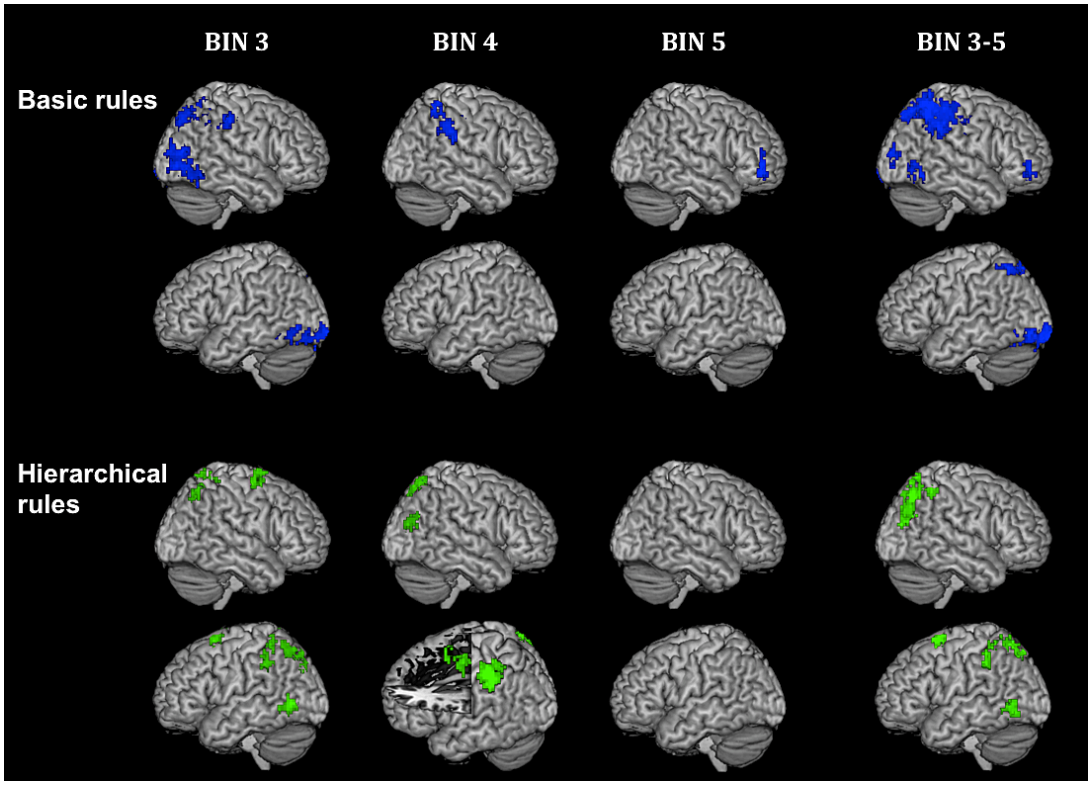


Figure 5.5 Results of the analyses on data from cue1 and delay1 phases for both BRs (on the top) and HRs (on the bottom), for each time bin considered separately and for all time bins taken together (in the last column). Brain areas in which it was possible to decode BRs are shown in blue; they were BA 17/18, 20, and 40 during time bin 3, BA 7 and 40 for time bin 4, BA 47 for time bin 5, and BA 7, 17/18, 20, 32, 40, and 47 for the whole time window considered. Regions encoding information about HRs are shown in green and comprise BA 6, 7, 21, 40, and 44 in time bin 3, BA 6, 7, 17, and 40 for time bin 4, and BA 6, 7, 21, and 40 for the whole time window considered.

Information about the **HRs** was available in parietal (SPL, BA 7; left hemisphere peak: -30, -52, 52, accuracy 69%; right hemisphere peak: 30, -67, 49, accuracy 67%; left IPL, BA 40; peak: -42, -40, 37, accuracy 69%), left temporal (middle temporal gyrus, MTG, BA 21; peak: -51, -67, 1, accuracy 72%), and frontal (SMA, BA 6; left hemisphere peak: -21, 11, 64, accuracy 69%; right hemisphere peak: 9, 5, 64, accuracy 70%; left inferior frontal gyrus, IFG, BA 44; peak: -36, 5, 28, accuracy 70%) areas during time bin 3; in right MOG (BA 17; peak: 42, -79, -19, accuracy 65%), left parietal (IPL, BA 40; peak: -54, -37, 55, accuracy 70%; SPL, BA 7; peak: 21, -58, 55, accuracy 68%), and prefrontal (PM area/SMA, BA 6; peak: -3, -7, 37, accuracy 68%) regions during time bin 4. It was not possible to decode the HRs from any area during time bin 5. Overall, considering all time bins, information about the HRs was available in SPL (BA 7; left hemisphere peak: -24, -76, 55, accuracy 82%; right hemisphere peak: 33, -61, 49, accuracy 89%:), left IPL (BA 40; peak: -48, -40, 40, accuracy 79%), left PM area (BA 6; peak: -24, 8, 40, accuracy 90%), and left MTG (BA 21; peak: -51, -70, -8, accuracy 83%).

5.4.3. Representations during rule integration

We analysed brain activity during *delay2* to decode rule representations of the BRs and the HRs when they were encoded at the same time. The results of the analyses are shown in Figure 5.6. We couldn't decode the **BRs** from any brain area during the single bins of delay2. Nevertheless, across the whole time window considered (bins 3 and 4, corresponding to the interval between 4 and 8s following cue2 presentation, Figure 5.3) information about the BRs was available bilaterally in MFG (FPC, BA 10; left hemisphere peak: -33, 56, 4, accuracy 74%; right hemisphere peak: 24, 50, 19, accuracy 73%). In contrast, it was possible to decode the **HRs** from right MOG/MTG (BA 19/21; peak: 48, -70, 7, accuracy 79%) and left MOG/fusiform gyrus (BA 19/37; peak: -30, -88, 1, accuracy 68%) during time bin 3 and from right entorhinal cortex/cerebellum (EC/hemispheric lobules [Hem] I-IV; peak: 21, -31, -32, accuracy 70%) and amygdala (peak: -15, -10, -20, accuracy 66%) during time bin 4. Across the whole time interval, information about the HRs was present in left inferior occipital gyrus/ITG (IOG/ITG, BA 19/20; peak -45, -70, -14, accuracy 82%), left amygdala/cerebellum (Hem I-IV; peak: -21, -31, -32, accuracy 80%), right MOG/MTG (BA 19/21; peak 45, -88, -2, accuracy 68%), and right amygdala/EC (peak: -15, -13, -5, accuracy 74%).

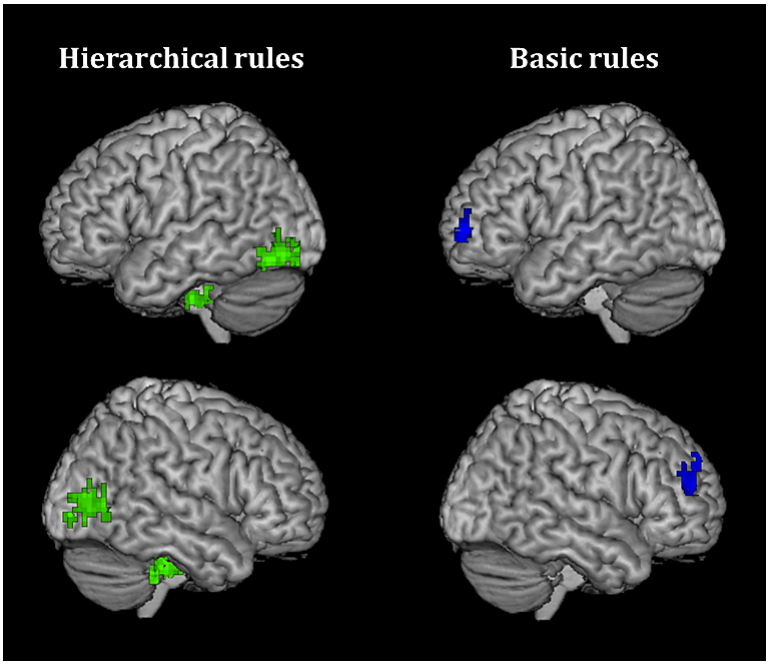


Figure 5.6 Results of the analyses on data from cue2 and delay2 phases for both HRs (on the left) and BRs (on the right) for time bins 3-4. The only brain area in which it was possible to decode BRs (in blue) was BA 10 on both hemispheres. Instead, regions encoding information about HRs (in green) were BA 19/20 and 19/21, amygdala, and cerebellum.

5.4.4. Comparison with results from previous studies

To evaluate whether the regions found in the present study corresponded to the results of previous studies on rule representation, we made direct **comparisons** with previous findings from our group (Figure 5.7). Brain areas encoding the BRs were compared with the regions found to hold information about rule identity in Reverberi, Görgen et al. 2012a (in Figure 5.7, the areas are shown in blue and yellow, respectively). In that experiment, the Authors used rules similar to our BRs: compound conditional rules linking a category with a motor response (e.g., "if face then left"). The clusters in the right *VLPFC* (BA 47) found in the two studies, as well as those in the left *parietal lobe* (IPL, BA 40) were highly overlapping (overlaps are shown in white in Figure 5.7). In our study, additional parietal and temporal areas emerged in the right hemisphere together with an occipital cluster in the left hemisphere.

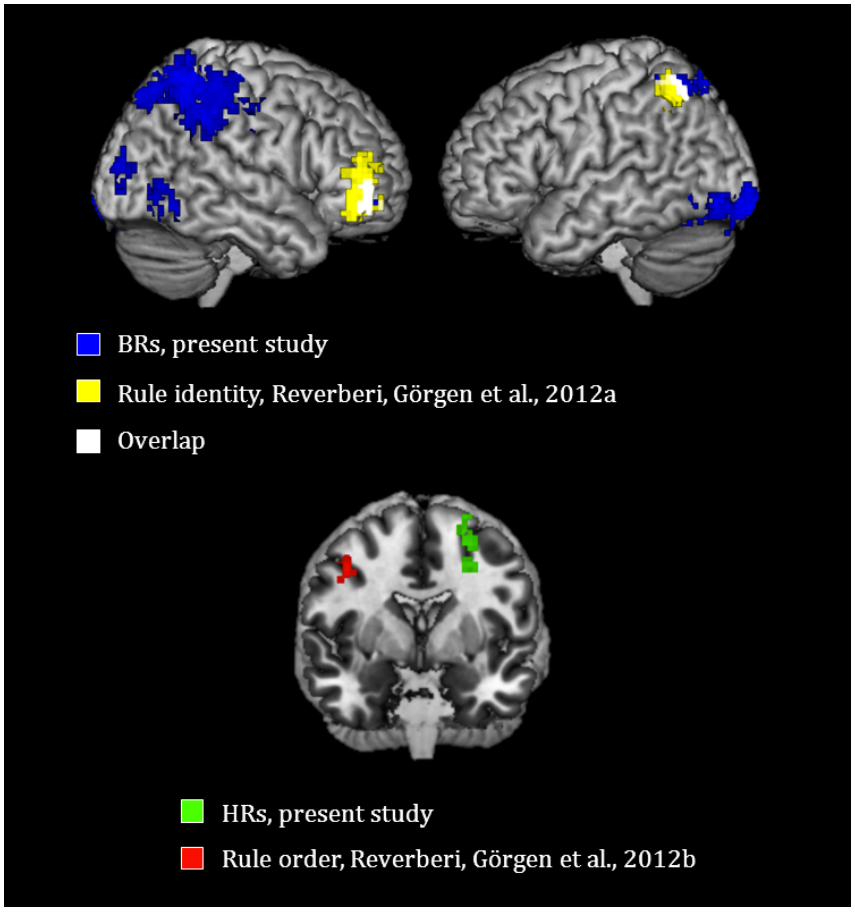


Figure 5.7 Comparisons with results from previous studies on rule representation. The clusters encoding rule identity in right VLPFC and parietal cortex from the present study and from Reverberi, Görgen et al. 2012a overlap. The cluster in PM area related to HRs in the present study corresponds to the area found in Reverberi, Görgen et al. 2012b for rule order, but it is located on the opposite hemisphere.

In addition, we compared areas containing information about the HRs in the present study with those encoding rule order in Reverberi, Görgen et al. 2012b (in green and red, respectively in Figure 5.7). In Reverberi's study, each single rule (e.g., "if food then A") associated a category (e.g., "food") with a symbol (e.g., "A"). Information about order was conveyed by the position of the cues (rules associated with the cue displayed on the top of the screen had to be applied first) and then influenced the application of the single rules. Hence, order constitutes a different feature of the rule set compared with the response mappings established by the compound rules. Similarly, our HRs represented an additional aspect of the rule set in that they established a hierarchy in action selection, since they influenced the application of the BRs. The result of the comparison showed that clusters in the *PM area* were active in both studies, although on different hemispheres (left in our study and right in Reverberi, Görgen et al. 2012b).

5.5. Discussion

This study investigated the neural basis of the representation of rules at different levels (low and high) in the hierarchy of action control. We considered both the condition in which each type of rule is encoded separately and the situation in which both types of rules are represented jointly. Different from previous studies addressing the same issue, we *aimed* to identify brain areas containing information about rule identity independently for each rule type and condition. We found *partially different* networks in the brain encoding distinct types of rules; furthermore we could decode each type of rule from more restricted and partially distinct areas when they were encoded together with another rule at a different level of action control. In particular, when rules were represented in isolation, information on low-level rules was present in occipital, temporal, parietal, and prefrontal areas, with the highest DA in parietal cortex (IPL, BA 40). Information related to high-level rules, instead, was found in parietal, temporal and frontal regions, with the highest DA in PM area (BA 6). In contrast, when rules of both types were represented jointly, information about the identity of the BRs could be decoded in rostral regions of PFC (FPC, BA 10) bilaterally, while information about the HRs was present bilaterally in occipital/temporal cortex and in amygdala/cerebellum, with the highest DA in left IOG/ITG (BA 19/20).

The finding that partially different brain networks underlie the representations of rules with the same formal structure (i.e., conditional rules) but at a different level in the hierarchy of action control, supports the idea that the brain represents complex rules by breaking them in their defining features (Reverberi, Görgen, et al., 2012b) and thus that different aspects of the rule set are encoded in distinct control layers (Badre & Frank, 2012; Botvinick, 2007; Frank & Badre, 2012; Koechlin & Summerfield, 2007). However, our results do *not* support the hypothesis of the existence of a gradient where more abstract rule representations are controlled by progressively more anterior regions in PFC. In fact, information about the low-level BRs was present in more anterior areas compared with the high-level HRs, both when they were maintained in isolation and when they were encoded together with the other rule type. If any, these findings would support a gradient proceeding in the opposite direction compared to the one usually described in literature (Badre & D'Esposito, 2007; Badre, 2008; Christoff et al., 2009; Koechlin et al., 2003; Koechlin & Summerfield, 2007; Nee & Brown, 2012b). To account for this discrepancy, it is important to point out the *peculiarity* of our study with respect to previous experiments investigating the functional specialisation of sub-regions in PFC. First, most of the evidence provided by those studies referred to brain activation during task execution; therefore, differences in results may simply reflect the distinct cognitive processes underlying rule processing and rule representation. Second, previous findings were often obtained using univariate analysis techniques; since univariate and multivariate methods are used to address different types of questions and adopt different perspectives ("activation-based" *vs.* "representational" view, Reverberi, Görgen, et al., 2012a), the use of different analysis techniques could in part account for differences in the results they produce (e.g., Bode & Haynes, 2009; Reverberi, Görgen, et al., 2012a). In fact, when comparing our findings with results obtained in experiments investigating the pure representation of task information and using the same analysis method, a greater overlap is observable. As shown in Paragraph 5.4.4, our results are in line with previous findings from our group using similar experimental paradigms. In particular, VLPFC (BA 47) and IPL (BA 40) held information about rules associating a category with a motor response. Right VLPFC was also found to represent information on rule identity in Reverberi, Görgen et al. 2012b. However, in that study rules couldn't be decoded from parietal cortex but instead from temporal areas. This finding was

explained by the differential involvement of temporal and parietal cortices in rule representation depending on the type of association entailed by the rule currently active (Reverberi, Görgen, et al., 2012b). In fact, temporal cortex seems to be recruited when combinations of objects are encoded (Baron & Osherson, 2011; Bunge, 2004) and thus when a rule links at least two visual objects, as in Reverberi, Görgen et al. 2012b, where a category was associated with a letter. Instead, parietal cortex is active when an object is directly linked with an action, as for the BRs in the present study and in Reverberi, Görgen et al. 2012a (Chao & Martin, 2000; Kellenbach, Brett, & Patterson, 2003). In line with this hypothesis, we found information about the HRs (where a shape was associated with a colour rather than with a motor response), but not related to the BRs, in the left MTG (BA 21). However, the temporal region identified in Reverberi, Görgen et al. 2012b was located in the right hemisphere; this difference in lateralisation was observed also in the corresponding PM clusters identified in the direct comparison with the results of that study.

Since the recruitment of a given area seems to depend on the specific *relation* established by the rule to be applied, it could be argued that differences in the brain areas involved depend on the specific rules considered rather than on other features defining the rule set. In this study, we used rules with the same formal structure (i.e., conditional rules) but at a different level in the hierarchy of action control. However, the BRs established a link between an object and a motor response, while the HRs associated an object with a feature of another object, thus requiring an attentional response. If the representation of rules requiring different response types involves distinct regions in the brain, as the aforementioned hypothesis and the results reported above suggest, then differences in the nature of the relations established by the rules rather than other features defining the rule set may account for the segregation of information in separate brain areas. However, the results discussed above, referred to different studies and a direct comparison between rules requiring different response types that would allow for testing this hypothesis was still lacking. Hence, we carried out an experiment involving rules identical in every respect but the type of response required; we are currently performing the analyses to assess this hypothesis.

Another study also investigated the representation of task information using *MVPA* (Nee & Brown, 2012b, see Paragraph 5.1). As in the present study, Nee and Brown

analysed brain activity during a first delay in which only one type of task information was encoded and during a second delay in which different types of information were represented. However, during the first delay the encoded information was always high-level, therefore low-level rules were only represented together with the alternative rule type. Moreover, fMRI data from the second delay were analysed to assess in which brain area it was possible to discriminate between different combinations of high- and low-level information (a total of four combinations considering the two high- and the two low-level rules); therefore the results of these analyses are qualitatively different and cannot be compared with our findings. Information about high-level context during the first delay was present in OFC (BA 10/11) and VMPFC (BA 11); in our study, we couldn't decode HRs from neither of these areas. However, *FPC* (BA 10) showed higher DA when high- and low-level information was integrated compared with the condition in which only high-level information was represented. This suggests a role of FPC in rule integration and may account for our finding that information about the BRs was encoded in that area when both types of rules were represented. Moreover, delay2 is proximal to action execution, compared with delay1 that is much farther in time from response. Since the BRs defined the response mappings and representations of the appropriate action have to be selected in order to respond, this information is more relevant in proximity to response onset. Therefore, the activation of FPC may reflect processes underlying *preparation* to select an appropriate response or to withhold an equally proper one when required by high-level information (e.g., in our study, when motor responses associated with a relevant object had to be withheld because its background colour didn't match the proper one). An alternative explanation could be that FPC controls the *retrieval* of associations relevant to react to an incoming stimulus and thus was involved in delay2 but not in delay1 since only in the former response mappings were retrieved and immediately applied. This hypothesis would explain also the findings from Nee and Brown (2012b), since low-level information determining the relevant S-R associations was always provided after high-level information, thus right before the target appeared and a response was required. Indeed, the involvement of FPC in rule retrieval has been demonstrated in preceding studies (Braver & Bongiolatti, 2002; Buckner & Koutstaal, 1998; Kuhl, Dudukovic, Kahn, & Wagner, 2007; Rugg & Wilding, 2000; Tulving, Kapur, Craik, Moscovitch, & Houle, 1994). However, while the

role of FPC in rule processing has been widely investigated and several cognitive processes were shown to recruit this region (e.g., managing the execution of multiple tasks, sub-goaling, evaluation of reward-related information, see Paragraph 1.3), the contribution of FPC in rule representation is less clear. In this study, we suggest different hypotheses that need to be further tested to shed light on the exact role played by FPC in rule representation.

Another unexpected finding is that information about the same type of rule was present in partially *different* brain areas when they were represented in isolation compared to the condition in which they were encoded together with another rule at a different level of action control. In a previous study, Reverberi, Görgen, and Haynes (2012a) found that the brain represents conditional rules using a compositional code (i.e., the representation of the rule is the same when it is represented alone and when it is part of a more complex rule set). Thus, we would have expected to decode information about a specific type of rule in similar areas in the two conditions considered in our study. Compound rules in Reverberi's study were composed of single rules linking a category with a response (e.g., "if there is a house press right; if there is a face press left"), with no difference in terms of relevance for selecting the appropriate response. In contrast, in the present study a hierarchy was established between the different types of rules that were part of the rule set. Thus, this finding could be explained assuming that a different *neural code* is used to represent rule sets when more complex features have to be considered. Further research is needed to identify the exact nature of the code that the brain uses to represent rule sets involving hierarchies of rules.

5.6. Conclusion

This study provides interesting results for research on the neural basis of rules representation. In particular, we found that rules with the same formal structure but at different levels in the hierarchy of action control are encoded in partially different brain networks, in line with recent findings from research on the functional *specialisation* of sub-regions in PFC. In addition, our results converge with recent findings from our group investigating simpler rule sets or different features defining such sets. However, contrary to our expectation, the results of the present study do *not* support the existence

of a functional gradient in PFC in which increasingly abstract information is represented in progressively more anterior areas. This finding could be ascribed to differences in the specific cognitive processes investigated or in the particular analysis techniques used respect to previous studies. Future studies should also take into account the possibility that the segregation of information in separate brain areas in PFC may be due to *qualitative* differences in the relations established by the rules rather than other relevant aspects of the rule sets.

We also showed that rule information was present in more restricted and partially different areas when high- and low-level rules were encoded *jointly* compared with the condition in which only one type of information was represented. In particular, information on the BRs was present in FPC (BA 10) during rule integration. This result is at odds with existing hypotheses on the role of this area in cognitive control. Hence, we propose that FPC may govern *preparation* processes underlying the selection of appropriate responses or the withholding of proper actions in light of high-level information; in alternative, FPC could control the *retrieval* of associations relevant to react promptly to an incoming stimulus. Further research testing these hypotheses may pinpoint the precise role of FPC in rule representation.

CHAPTER 6

6. Study 4: Neural basis of propositional connectives

6.1. Theoretical background

A logical **connective** is a word (or a symbol) used to modify the truth-value of a linguistic term or expression or to link two pieces of information, such as two sentences or two rules. The most basic and common logical connectives are: *not*, *and*, *or*, *if-then*, and *if and only if*. Formally, they are also known, respectively, as negation, conjunction, disjunction, conditional (or implication), and biconditional (or double implication) and are regarded as the key syntactic operators of propositional logic. These connectives are highly frequent in natural language and are often crucial to understand its meaning. In fact, removing logical connectives from language would make it harder to convey ideas and express concepts clearly.

If different cognitive domains exploit the same basic logic, as some scholars claim (Stenning & van Lambalgen, 2008), then logical connectives may be crucial not just for discourse processing but also for other cognitive processes, such as reasoning or *executive functions*. Indeed, executive processes such as planning usually involve the selection, manipulation, and integration of numerous pieces of information in order to reach a certain goal or produce a specific output; logical connectives often define how such chunks of information have to be combined in order to achieve such goals.

In addition, logical connectives can be represented in abstract, symbolic systems, such as formal **propositional logic** or mathematics. In his *Logisch-Philosophische Abhandlung*, Wittgenstein put forth a complete formal *truth-functional* propositional logic (Wittgenstein, 1921). Within this system, the logical connectives (or operators) *not*, *and*, *or*, *if-then,* and *if and only if* are represented by symbols: $\sim$, $\wedge$, $\vee$, $\rightarrow$, and $\leftrightarrow$ respectively. Instead, elementary propositions (the elements linked by connectives) are represented by letters: *p* and *q*. The first proposition (p) is referred to as the *antecedent* and the second proposition (q) is referred to as the *consequent*. This allows natural language sentences to be symbolically represented and the truth-value of the sentence

in question to be evaluated depending on the truth of each proposition. For example, the sentence "Sarah is reading a book and Tom is cooking" can be represented as p ∧ q. The truth of the entire statement can then be evaluated based on the truth of each component: if p is true (1) but q is false (0), the entire utterance is false. Likewise, if q = 1 but p = 0, the whole utterance is false. The only case in which p ∧ q = 1 is when both p = 1 and q = 1 (i.e., both Sarah is reading a book and Tom is cooking). This truth-functionality can be represented, for each logical connective and for every possible combination of truth or falsehood of p and q, in a truth table (Figure 6.2 B).

The strength of this formalisation is that it is both sound and complete. Soundness refers to the fact that everything that can be derived syntactically is true semantically. Completeness refers to the fact that everything that is true semantically can be derived syntactically. These properties make propositional logic an essential tool for mathematical reasoning and provide mathematics with some of the basic operators that constitute its language. The logical consequence of the combination of true or false mathematical statements is by definition either true or false; that is the reason why a counterexample to a theorem makes it false (Stenning & van Lambalgen, 2008).

The propositional logic system described above is a classical logic, in which each logical connective has a definite *truth-functional* **interpretation**. For example, as stated above, for an *and*-statement to be true, both propositions must be true and all other possible combinations are false (Figure 6.2 B). The classical interpretation of *or* is *inclusive*, meaning that either one proposition or both must be true for the statement to be true (Int1 for OR in Figure 6.2 B). The conditional *if-then* is viewed as material implication (Stenning & van Lambalgen, 2008) (Int1 for IF-THEN in Figure 6.2 B), meaning that the only false condition is when the first proposition is true and the second one is false. Finally, the biconditional *if and only if* is true when both the propositions have the same truth-value (they are both true of both false). The negation *not*, instead, simply modifies the truth-value of a single proposition (from true to false and vice versa). Interestingly, in everyday life these logical connectives are often interpreted differently with respect to the classical logic described here. In natural language, *or* is usually used with the meaning "one or the other, but not both" (e.g., Chevallier et al., 2008). In other words, its meaning is most commonly *exclusive* (Int2 for OR in Figure 6.2 B). *If-then* can also be interpreted either as "and", as in its *conjunctive* or conjunctive-

defective interpretation (Int2 for IF-THEN in Figure 6.2 B), or as "if and only if", as in its *biconditional* interpretation (Int3 for IF-THEN in Figure 6.2 B). The alternative interpretations of the conditional have been a matter of deep debate (see, e.g., Johnson-Laird & Byrne, 2002) and are probably due to difficulty in reasoning in two instances: when the antecedent is false but the consequent is true ("If p then q"; ~p), or when both the antecedent and the consequent are false ("If p then q"; ~p and ~q) (Byrne & Johnson-Laird, 2009).

The picture outlined so far depicts logical connectives as a critical component of human cognition across many domains, from natural language to executive functions, as well as in logic and mathematics. Specifically, it seems that logical connectives play a role in the way humans *organise* thoughts across many domains. Hence, the differences in logical interpretation provide an interesting opportunity to study the link between the formal structure of the connectives and their pragmatic or cognitive use. Yet, despite their important role as building blocks of human cognition, little is known about how they are represented in the brain.

Several studies have started to shed light on the **neuroanatomy** of reasoning with propositional logic (for a meta-analysis see Prado et al., 2011), using either fMRI (e.g., Knauff, Mulack, Kassubek, Salih, & Greenlee, 2002), EEG (e.g., Prado, Kaliuzhna, Cheylus, & Noveck, 2008), or positron emission tomography (PET) (Parsons & Osherson, 2001). The majority of them used *deduction tasks* (see Paragraph 3.1). Deductive inferences drawn from premises containing logical connectives are important for discourse and thinking in everyday life; this has motivated the use of deductive tasks in the laboratory to further understand propositional logic.

In previous studies, multiple premises were presented and subjects were required to integrate them to *generate* an inference (e.g., Reverberi et al., 2007), to *choose* the correct conclusion for a specific problem (e.g., Prado et al., 2010), or to *evaluate* the validity of the provided conclusion (e.g., Reverberi et al., 2009). Houdé and colleagues (2000), instead, used single conditional statements comprising negations and required participants to explicitly *falsify* them (see also Canessa et al., 2005, but with the introduction of social content). Only two works focused on the *encoding* phase of conditional sentences, either with null results (Reverberi et al., 2010) or with an explicit focus on word processing (Bonnefond & Van Der Henst, 2013). Moreover, most of the

studies investigated conditional problems (e.g., Noveck et al., 2004) and only few introduced also the disjunction (e.g., Monti, Parsons, & Osherson, 2009). In either case, potential differences between groups with alternative *interpretations* of the connective (IF or OR) were not investigated. Some research took into account also the *difficulty* of the conditional problems, in terms of deductive complexity (e.g., Monti et al., 2007), and the *linguistic* component of the reasoning task (Parsons & Osherson, 2001), since almost all of the studies used verbal stimuli (but see Houdé et al., 2000). This factor should be especially considered since, in the majority of the experiments, a left-lateralised network emerged and areas related to language were often activated. For example, the left **IFG** (BA 44/45/47), regarded as part of the "Broca's Complex" (Hagoort, 2009), was related to reasoning processes in many studies (Houdé et al., 2000; Noveck et al., 2004; Prado et al., 2008; Reverberi et al., 2010). Since Broca's Complex is important for syntactic operations in natural language, one possibility is that its function in reasoning is to perform formal rule-based manipulations on the premises in order to draw a conclusion (Reverberi et al., 2007). Although in some studies controlling for linguistic aspects of the task Broca's Area (BA 44/45) did not come out (e.g., Monti et al., 2009; Parsons & Osherson, 2001; for a discussion see Monti & Osherson, 2012), in others it was observed (Reverberi et al., 2012). However, in the first case the task implied some form of reasoning on both logic and linguistic arguments, while in the last study a deductive task was contrasted with a memory task on the same stimuli. In Prado's meta-analysis (2011), the only areas that were activated consistently across all the eleven studies investigating propositional reasoning were the left precentral gyrus (PG, BA 6), medial frontal gyrus (MeFG, BA 8), and left PPC (BA 39). The latter, in particular, was associated more specifically with this specific type of argument.

The neuroimaging studies reviewed so far provide insights on the neural correlates of the mechanisms underlying how propositional connectives are processed. Information about where and how these connectives are represented, instead, is still too little. Most of our knowledge comes from research on cognitive control that investigated rule *representation*. Some of these studies used MVPA to decode information about the specific rule active on a given trial based on patterns of neural activity discriminative for that rule (see Paragraph 2.4). Typically, the tasks devised use conditional rules associating a stimulus or condition with a response (i.e., if-then statements). For

example, Reverberi, Görgen, and Haynes (2012a, 2012b) investigated the encoding of conditional rules linking a category with a response (e.g., "if there is a house press right"). On each trial, rules were instructed by graphic *symbols* that subjects had learnt to associate to the different rules in a preliminary training session. Thus, linguistic stimuli were not involved. They could decode information about rule identity from right prefrontal (BA 47 and BA 8/9), temporal (BA 20/21), and left parietal (BA 7/40) regions. Interestingly, they also used single together with *compound* rules (i.e., a conjunction of two conditional rules, similar to the statements composed using a combination of different connectives in difficult deductive tasks). They found that the brain represents conditional rules using a compositional code (i.e., the representation of the rule is the same when it is represented alone and when it is part of a more complex rule set). However, the focus of those works was not the logical form of the rules and the areas identified to decode rule identity could have likely ignored any information related to the formal structure since it was invariant across all rules.

Taken together, the studies discussed in this introduction are far from fully clarifying the cognitive processes underlying the encoding and evaluation of rules composed using logical connectives. Specifically, despite some knowledge on the neural basis of the reasoning processes involving some of these connectives is available, more specific information about how the brain encodes different types of connectives and operates on them to evaluate the truth, consistency, or validity of rules with a different formal structure is still *lacking*.

To overcome the limitations of the previous investigations, in the **present study** we: 1) used simple rules composed using *three* different logical connectives (conjunction, disjunction, and conditional) to allow direct comparisons between them; 2) chose graphic *symbols* to code for rules to minimise the influence of linguistic processes in rule encoding and application; 3) analysed brain activity during *both* the encoding (to gather information about the pure representation of the rules) and the assessment of the compatibility of target scenes with the active rule (to assess potential differences in the processing of distinct connectives); 4) introduced a *control* condition to better assess the specificity of the neural signal associated with the variables of interest; 5) directly compared alternative *interpretations* of the same connective; 6) used classical univariate analysis techniques in *combination* with more advanced MVPA methods to get hints

about both the regions activated by the conditions of interest and where and how information on the active rule is represented in the brain.

A clear understanding of these elementary building blocks of logical forms is an essential step forward in understanding the neural basis of deductive reasoning, at both a basic and a more highly complex/abstract level. Elucidating the neural basis of logical connectives can provide an intriguing insight into complex areas of cognition, many of which are unique to humans (e.g., Kandel, 2000).

6.2. Methods

6.2.1. Participants

Thirty-four participants (18 females) took part in the study in exchange of a monetary payment. All were right-handed (score > 40 at the EHI), native German speakers, screened to ensure no neurological or psychiatric disorder and had normal or corrected-to-normal vision. The local ethics committee approved the study and all participants provided written informed consent. Four participants were discarded because of poor performance (consistency < .80 in at least one of the experimental conditions). Thus, the results reported were obtained from the remaining 30 participants (mean age 25.4 years; range 19-35). All the experimental materials were provided in German. Participants were recruited through a *pre-screening* procedure; they filled in a paper and pencil logic questionnaire (see Appendix C1) and were included based on two requirements: 1) all responses had to be correct according to any of the possible logical interpretations of the connectives, and 2) the interpretation of the conditional had to be classical or conjunctive. For the conditional rule, one subject had a biconditional interpretation, 16 had a classical interpretation (IF_{clas}) and 17 a conjunctive interpretation (IF_{conj}); for the disjunction, 23 had an inclusive interpretation (OR_{inc}) of the connective and 11 an exclusive interpretation (OR_{exc}).

6.2.2. Stimuli and experimental procedure

Participants performed a **task** requiring to retrieve and maintain three logical rules (conjunction, disjunction, and conditional – R1, R2, and R3 in Figure 6.1, respectively) in order to evaluate whether they were compatible or incompatible with different target

scenes. Each logical rule had the same antecedent (i.e., yellow square) and consequent (i.e., green circle) and therefore differed from the others *only* in the logical connective used to compose that rule. We introduced a further *control* condition (baseline, B in Figure 6.1) identical to the experimental conditions except that it didn't require to evaluate the scene but simply to respond to a stimulus (a letter changing colour). Each rule was associated with two visually unrelated cues (Figure 6.1 A); participants had learnt the associations between the cues and the rules in a separate training session (see below). The associations were randomised across subjects so that each one learnt different cue-rule associations.

At the beginning of each **trial** a cue was shown (1s) at the centre of the screen and the participant had to retrieve the corresponding rule. After cue presentation, a delay period (6s) followed and then a target scene appeared (3s). The scene could be of *four* different types, comprising all the possible combinations of antecedent and consequent (Figure 6.2 A): 1) antecedent and consequent (AC), 2) antecedent only (A), 3) consequent only (C), 4) and neither antecedent nor consequent (None). Participants had to evaluate whether the target scene was *compatible* (C) or *incompatible* (I) with the rule active in the current trial (i.e., whether the scene make the rule true or false, respectively). Please note that, while the displayed shapes were always the same (a square and a circle), shape colours were changed to define presence or absence of the rule antecedent/consequent. Additionally, the location of the shapes (top or bottom of the screen) was randomised. Specifically, the shape presented at the top of the screen matched the antecedent of the rule (square) in about half of the trials, and vice versa. The centre of the target screen displayed two letters: "C" for compatible and "I" for incompatible. The location in which each letter appeared (right or left side of the screen) was randomised throughout trials. This randomisation was used so that participants couldn't plan a motor response in advance during the delay period; thus, they were forced to wait until the target appeared to figure out on which side each letter was. Participants had to respond as quickly as possible pressing the inner button of either the left or the right button box, according to the side of the letter they wanted to choose. They used the index finger of the respective hand to respond. For the *baseline* condition, either letter C or I would turn orange at approximately 1505ms (average RT for experimental trials) and the participant had to immediately press the corresponding

response button. In addition to experimental trials, some *catch* trials, identical to the experimental trials but with a shorter delay period (1s), were interspersed throughout the standard trials, in order to ensure that participants immediately retrieved and represented the rule.

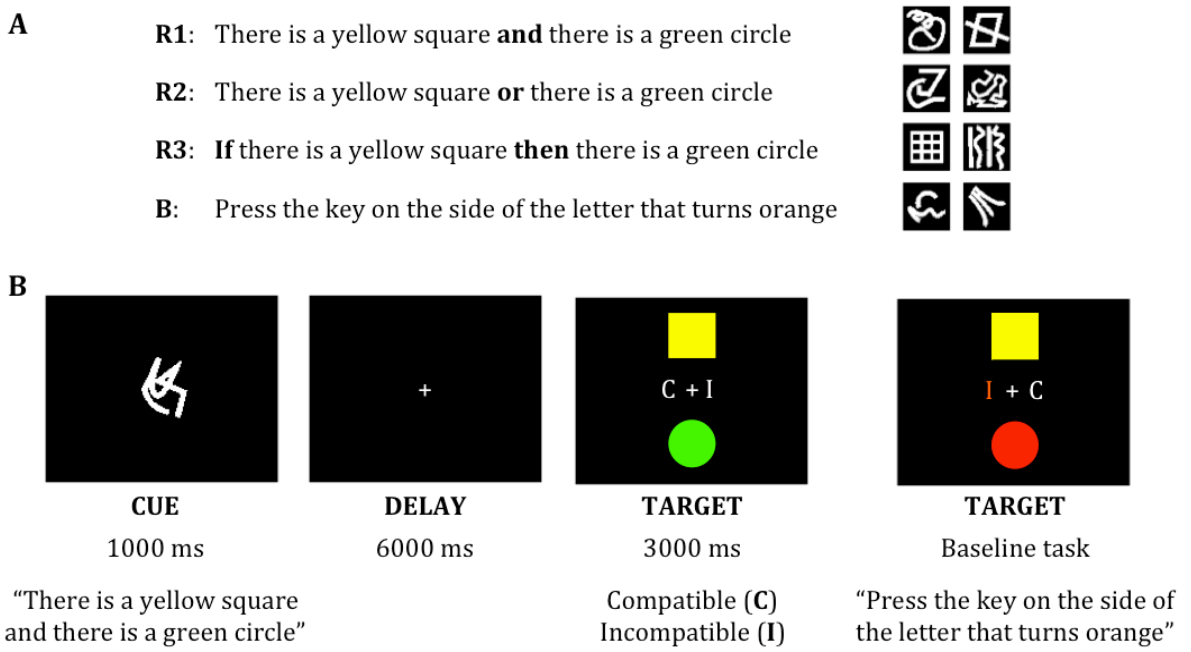
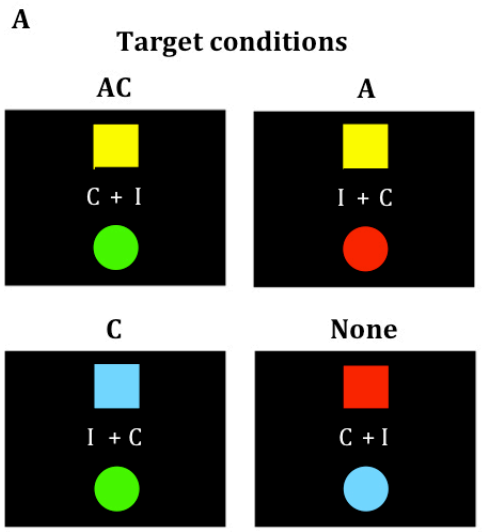


Figure 6.1 Rules employed (A). We used three rules composed using different logical connectives (R1-*and*, R2-*or*, and R3-*If-then*) and a low level baseline (B, "Press the key on the side of the letter that turns orange"). Each of the three rules and the low-level baseline were associated with two visual cues (abstract symbols, on the right). Timeline of the experiment (B). At the beginning of each trial, a cue was presented, indicating which rule was active in the current trial. After a delay of 6s, the target scene appeared. Participants had to evaluate if the scene was compatible or incompatible with the active rule and press as fast as possible the appropriate button.



B

True table

	A Yellow square	C Green circle	Int1	Int2	Int3
AND	1	1	1		
	1	0	0		
	0	1	0		
	0	1	0		
OR	1	1	1	0	
	1	0	1	1	
	0	1	1	1	
	0	1	0	0	
IF-THEN	1	1	1	1	1
	1	0	0	0	0
	0	1	1	0	0
	0	1	1	0	1

Figure 6.2 Possible targets (A). AC = antecedent and consequent, A = antecedent only, C = consequent only, and None = neither antecedent nor consequent. True table for the three connectives (B). Truth-value: 1 = true, 0 = false. Interpretation: OR, Int1 = inclusive, Int2 = exclusive; IF-THEN: Int1 = classical, Int2 = conjunctive, Int3 = biconditional.

During scanning, participants performed 360 trials, divided into 6 runs. In each run, 60 trials (36 experimental, 12 baseline, and 12 catch trials) were administered in random order. The ITI was 2050ms, and the whole experiment lasted about 61min. Before scanning, participants performed a "refresher session" for 10min to ensure that they remembered the cue-rule associations. Once in the scanner, participants performed 15 experimental trials to get used to the new setting. These trials were administered for practice purpose only and thus were not further analysed. After scanning, a *questionnaire* was administered to investigate if and what types of strategies were used to perform the task (see Appendix C2). The experiment was devised and administered via MATLAB® (The Mathworks, Inc., Natick, MA, US), using the Cogent 2000 toolbox (Functional Imaging Laboratory and Institute of Cognitive Neuroscience, University College London, London, UK).

At most three days prior the scanning session, participants underwent a **training** session. On the first part of the training (mean duration = 45min), they learnt the cue-rule associations; on the second part (mean duration = 60min), they practised the experimental task and received feedback on their accuracy. Two participants performed the sessions on two different consecutive days. Only participants who applied the rules with high consistency (>= 80%) were allowed to the fMRI session.

6.2.3. Image acquisition

Functional imaging data were collected using a 3-T Siemens Trio scanner (Erlangen, Germany), equipped with a 12-channel head coil. In each of the six scanning sessions, we acquired 340 T_2*-weighted images in descending order, using GRE EPI sequences. The volumes were composed of 33 slices (3mm thick), separated by a gap of 0.75mm. Imaging parameters were as follows: TR 2000ms, TE 30ms, FA 78°, matrix size 64 × 64, and FOV of 192mm × 192mm, thus yielding an in-plane voxel resolution of $3mm^2$, resulting in a voxel size of 3mm × 3mm × 3.75mm. At the beginning of the scanning session, a T_1-weighted structural dataset was collected, with the following parameters: TR 1900ms, TE 2.52ms, FA 9°, matrix size 256 × 256 × 192, FOV of 256mm × 256mm × 192mm, 192 slices (1mm thick), and resolution 1mm × 1mm × 1mm. After the experiment, a magnetic field mapping sequence was also run. Imaging parameters were

as follows: TR 400ms, TE 5.19ms and 7.65ms, FA 60°, matrix size 64 × 64, FOV of 192mm × 192mm, 33 slices (3mm thick), and resolution 3mm × 3mm.

6.3. Data analyses

6.3.1. Pre-processing

Functional data were pre-processed and analysed using SPM12 (Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, London, UK). During pre-processing, the volumes were realigned, corrected for geometric distortions using unwrapped field maps, and slice-time corrected. Low-frequency noise was removed using a high-pass filter with a cutoff period of 128s (Worsley & Friston, 1995) and an AR model was fit to the residuals to allow for temporal autocorrelations (Friston et al., 2002). For the univariate analysis, the data were also normalised and spatially smoothed (kernel size = 6 voxels Full Width at Half Maximum – FWHM).

6.3.2. Univariate analyses

We used the univariate method to analyse target scene **processing**. The introduction of a low-level *baseline* allowed us to filter out neural activity related to ancillary processes and thus not reflecting the evaluation of the rules (Monti & Osherson, 2012). In fact, the baseline involved the same processes as the experimental rules (e.g., cue representation, rule encoding, target evaluation, response execution) but didn't imply the assessment of the compatibility of the target scene with the rule. A fist-level GLM model was estimated using one regressor for each rule, yielding a total of four *regressors* for each subject: AND (R1), OR (R2), IF (R3), and low-level baseline (B). The regressors of the GLM corresponded to the target-onset time vectors of each condition convolved with the HRF; duration was set to 0. The effects of interest were estimated using all six runs. *Contrasts* were used to determine the main effect of each condition of interest (AND, OR, IF, and baseline), resulting in the production of four contrast images per subject. These images were normalised to MNI space and subjected to a one-way ANOVA. Contrasts were devised to assess differences in target processing between all

pairs of different rule types at the group level. We performed a conjunction analysis[2] (Friston, Penny, & Glaser, 2005; Nichols, Brett, Andersson, Wager, & Poline, 2005) to identify brain areas significantly activated by a given rule compared to both the other rules (e.g., a conjunction analysis on the effects IF > OR and IF > AND will identify only brain regions significantly active in both contrasts). Statistical inferences at the group level relied on a RFX model. Alternative interpretations of the same connective were analysed as between-group factors.

6.3.3. Multivariate Pattern Analysis

We applied MVPA to identify which areas in the brain **encode** information about rules with logical connectives. An FIR model was applied to the realigned and slice-time corrected images (Henson, 2004). The volumes were neither spatially smoothed nor normalised in order to preserve fine-grained patterns of activation. We modelled four *regressors* corresponding to the four rules: AND, OR, IF, and baseline. Each condition was modelled using 16 time bins of 2s each. The time vectors of all regressors were defined using cue onset times. Distinct regressors were computed for each of the six runs. The beta parameters estimated by the FIR model were used to perform a whole-brain MVPA using a *searchlight* approach (see Figure 5.2 A). For this analysis we implemented a 6-fold *leave-one-run-out* CV procedure to obtain a better estimate of the classifier accuracy. A *linear SVM* with a fixed regularisation parameter C = 1 was trained to distinguish between two of the experimental conditions using data taken from five of the six runs available; the classifier performance was then tested on the data from the remaining run (see Figure 5.2 C). The procedure was repeated six times and the results were averaged. Three different classifications were performed using all possible pairs of experimental conditions: AND *vs.* OR, AND *vs.* IF, and OR *vs.* IF. Only data from the FIR time bins 3, 4, and 5 were used, which corresponded to the time window from 4 to 10s after cue onset. Based on the delay of the HRF, this interval reflects only activity related to the cue and delay phases. For each searchlight sphere (radius = 4 voxels) a measure of DA (with respect to chance level) was obtained. The values were then combined to

[2] Conjunction analysis is used to test whether the activations in two or more conditions (or tasks) are significant when considered jointly. This method looks for similarity in effects between different conditions. Therefore, a conjunction analysis will produce significant results only when the effect is present in both the conditions at the same time.

produce *accuracy maps*. For each participant, the three accuracy maps resulting from the analyses described above were normalised to MNI space and averaged before performing group-level analyses. The resulting mean accuracy maps were submitted to a one-way ANOVA (factor levels = number of bins) to test where in the brain DAs differed from chance level across all participants.

6.3.4. ROI analyses

ROI analyses were also performed to further investigate potential differences between rules composed by different connectives, *both* in rule encoding and in scene evaluation. Since no previous imaging study specifically investigated the evaluation of the truth of simple rules involving logical connectives, we referred to studies on deductive reasoning. In particular, for this set of analyses we used the six **ROIs** identified in Reverberi et al. 2012 (Figure 6.3 A). In that study, the Authors explored inference generation in quantified syllogisms and found that different cognitive processes involved in inference generation activated separate brain regions. In particular, they proposed that BA 44/45 and BA 6/8 are involved in extracting the formal structure of the deductive problem, while BA 47 is active when selecting and applying the proper inferential rules. Similarly, our study required the extraction of the logical structure of the rule and the evaluation of its validity with respect to information conveyed by a target scene. The six ROIs were processed using the MarsBar toolbox (Brett, Anton, Valabregue, & Poline, 2002). In the first set of ROI analyses, we extracted an average effect across all voxels within each of the six ROIs. The effects were computed as *differences* in activation between either OR or IF and AND. The average activations were submitted to one-sample t-tests to look for significant effects in any of the six ROIs. The second ROI analysis tested potential differences in activation during target processing between groups of subjects using alternative *interpretations* of the same connective (OR or IF). We calculated a mean effect (as difference in activation between the two groups: $IF_{clas} > IF_{conj}$ for the interpretation of IF and $OR_{inc} > OR_{exc}$ for the interpretation of OR) across all voxels from each of the six ROIs. The resulting mean activations were subjected to one-sample t-tests to identify significant effects in any ROI.

Finally, ROI analyses were performed also on the *DAs* estimated by the MVPA. For the first analyses, we used data from the FIR time bins 3 (to account for the delay of the

BOLD signal) to 8, thus covering the time window from 4 to 16s after cue onset. Since the latency of the peak in the HRF depends on both the specific task being measured and the brain region involved (Heeger & Ress, 2002), using an extended interval might allow for detecting effects also in regions and for rules with a longer latency. For each participant, we computed the mean accuracy reached by the classifier across all voxels in each ROI. The resulting values were submitted to a one-sample t-test to identify significant effects in any of the six ROIs.

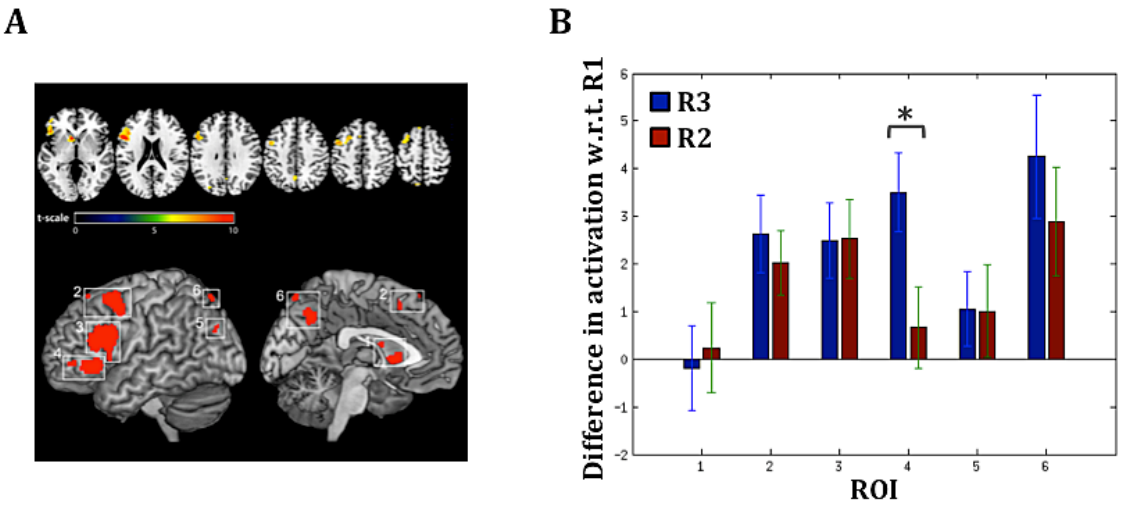


Figure 6.3 ROI analyses. ROIs from Reverberi et al. 2012 (A): 1 = basal ganglia (caudate nucleus), 2 = PG, SMA and frontal middle gyrus (BA 6 and FMG, BA 8), 3 = posterior IFG (BA 44/45), 4 = anterior IFG (BA 45/47), 5 = occipital middle gyrus (OMG, BA 19), and 6 = medial parietal cortex (mPC, precuneus, BA 7). Adapted from Reverberi et al. 2012, Neuroimage, 59, 1752-1764. Differences in activation for either OR or IF with respect to AND (B). The difference between OR and IF is significant in anterior IFG (ROI 4).

A further ROI analysis was performed to test for possible differences between groups with alternative *interpretations* of IF. Only data from time bins 3, 4, and 5 were analysed. We calculated a mean effect (as difference in the DAs between the two groups IF_{clas} and IF_{conj}) across all voxels from each of the six ROIs. The average accuracies were tested with a one-sample t-test.

6.4. Results

6.4.1. Behavioural results

During scanning, subjects evaluated the rules with high **consistency** (Figure 6.4); the average index of consistency was 89% (SD = 4.9%). The mean RTs for each

experimental rule and possible target scene are reported in Table 6.1. Subjects responded with high accuracy (mean = 93%, SD = 3%, Figure 6.5 A) and short RTs (mean = 1493.3ms, SD = 177.7ms, Figure 6.5 B) also to the *baseline* trials.

Table 6.1 Mean RTs (in ms, with SDs in parentheses) for each experimental rule and possible target scenes. Possible target scenes are: A (only the antecedent is present), AC (both the antecedent and the consequent are present), C (only the consequent is present), and None (neither the antecedent nor the consequent is present).

Rule	A	AC	C	None
AND	1351.4 (242.81)	1287.4 (204.86)	1328.1 (226.98)	1375.3 (198.95)
OR	1640.1 (209.31)	1461.5 (206.26)	1464.8 (189.78)	1461.4 (201.95)
IF	1449.9 (283.75)	1439 (243.53)	1423.3 (230.69)	1390.4 (202.55)

Generally, subjects pressed the button in less than half of the total time at their disposal to respond (3s). Such short RTs, together with the high level of consistency, attest that subjects represented the active rule during the delay before the scene appeared.

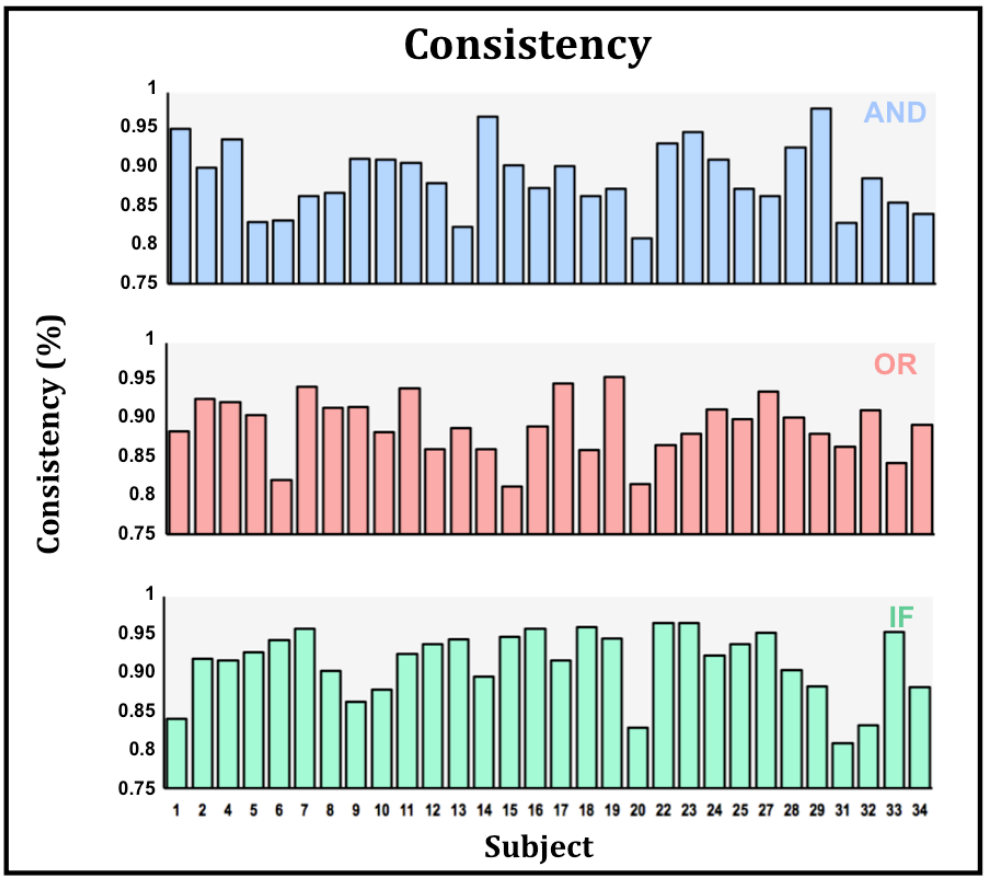


Figure 6.4 Consistency in the evaluation of the scenes for each experimental rule and for each subject: AND rule (top histogram, blue), OR rule (middle histogram, red), and IF rule (bottom histogram, green).

Consistency and RTs were similar in groups with different **interpretations** of the connectives. Subjects with an inclusive interpretation of OR responded to all the experimental rules equally fast and consistently as the group with an exclusive interpretation of that connective. Subjects with a classical interpretation of IF responded to the AND and OR rules with similar accuracies and RTs compared with the group with a conjunctive interpretation of the conditional. However, they were significantly more consistent (mean 94% *vs.* 90%, $t(21) = 2.52$, $p = .02$, $d = 0.87$) and slower (mean 1540.4ms *vs.* 1326ms, $t(22) = 3.22$, $p = .004$, $d = 1.21$) in evaluating scenes when required to assess the IF rule. Differences in RTs between the two groups may be due to the higher difficulty in processing the conditional compared to the conjunction (e.g., Johnson-Laird, Byrne, & Schaeken, 1992), since for the group with a conjunctive-defective interpretation of IF this rule reduced to the AND rule.

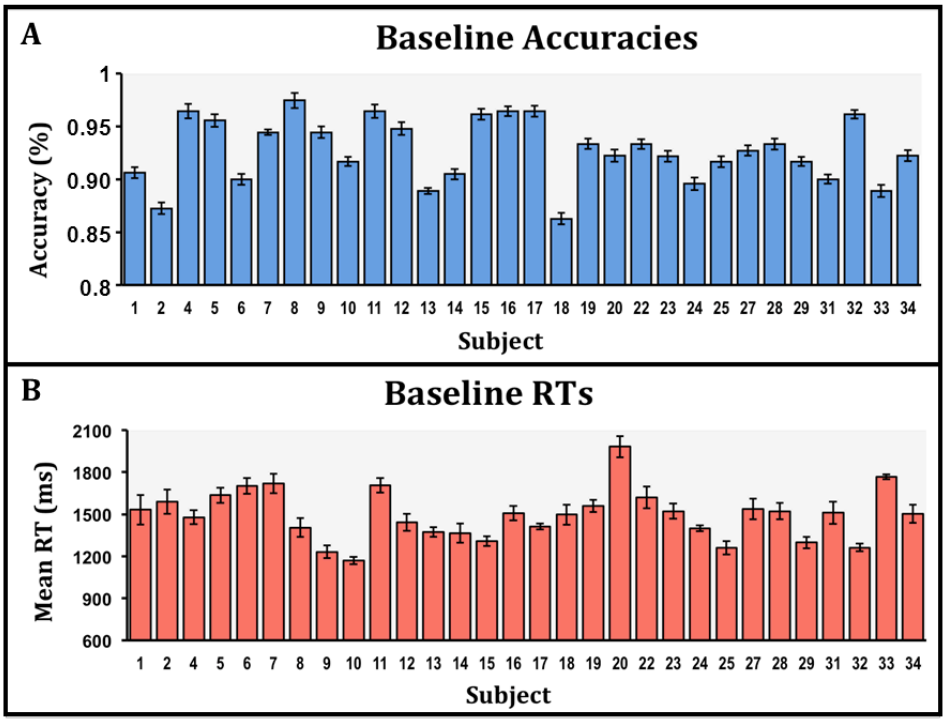


Figure 6.5 Accuracy (A) and mean RT (B) in the baseline rule for each subject.

6.4.2. Decoding representations of rules with logical connectives

The **aim** of the decoding analysis was to identify brain areas coding for rules with logical connectives. The results of the analysis (all *p*s < .001 uncorrected at voxel-level and $p < .05$ FWE-corrected at cluster level) are shown in Figure 6.6. A flow of

information (Bode & Haynes, 2009) across time is observable. During time bin 3 (time window from 4 to 6s after cue presentation), information was predominantly represented in occipital areas (MOG, BA 17; left hemisphere peak: -48, -79, 5, accuracy 61%; right hemisphere peak: 45, -79, 5, accuracy 62%), with only a small cluster in the frontal lobe (left IFG, BA 44; peak: -51, 2, 35, accuracy 58%). During time bins 4 and 5, information shifted progressively more anteriorly: in bin 4 (from 6 to 8s) it was represented partly in occipital (MOG, BA 17; left hemisphere peak: -45, -73, 11, accuracy 57%; right hemisphere peak: 33, -82, 8, accuracy 60%; left superior occipital gyrus, SOG, BA 17; peak: -12, -91, 11, accuracy 57%; right lingual gyrus; peak: 24, -46, -10, accuracy 58%), left temporal (MTG, BA 21; peak: -48, -34, -7, accuracy 59%), and left parietal (IPL, BA 40; peak: -33, -73, 47, accuracy 57%) areas and mostly in prefrontal cortex (left SMA, BA 6; peak: -9, 11, 56, accuracy 58%; left IFG, BA 44; peak: -54, 8, 20, accuracy 60%; right MFG, BA 46; peak: 21, 50, -1, accuracy 56%); in bin 5 (from 8 to 10s), information about rules was present only in right FPC (BA 10; peak: 0, 62, 14, accuracy 58%) and left IFG (BA 44; peak: -57, 11, 17, accuracy 58%). To test if any of the areas that were informative in each time bin was able to decode rules throughout the whole interval, we performed a conjunction analysis across the three time bins. The *only* brain area coding for rules with logical connectives throughout all seven seconds of the cue and delay phases was left **IFG** (BA 44; peak: -54, 11, 20, accuracy 75%; Figure 6.7 A).

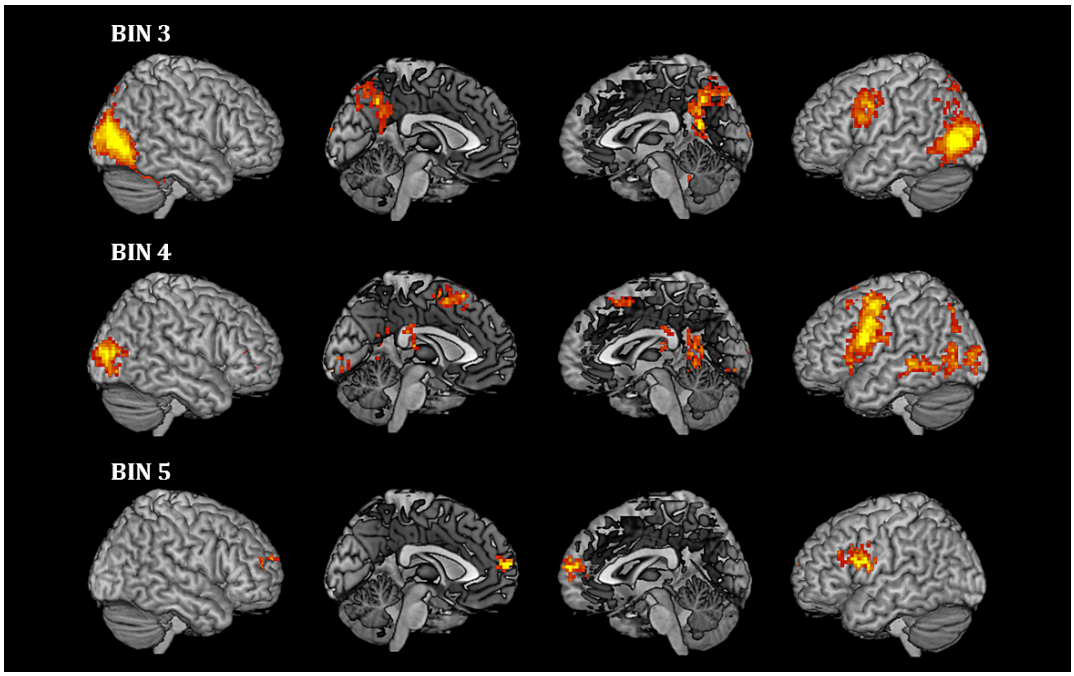


Figure 6.6 Results of the MVPA on data from cue and delay phases, for each time bin. Brain areas where it was possible to decode the rules during time bin 3 (BA 17 and 44), 4 (BA 6, 17, 21, 40, 44, and 46), and 5 (BA 10 and 44).

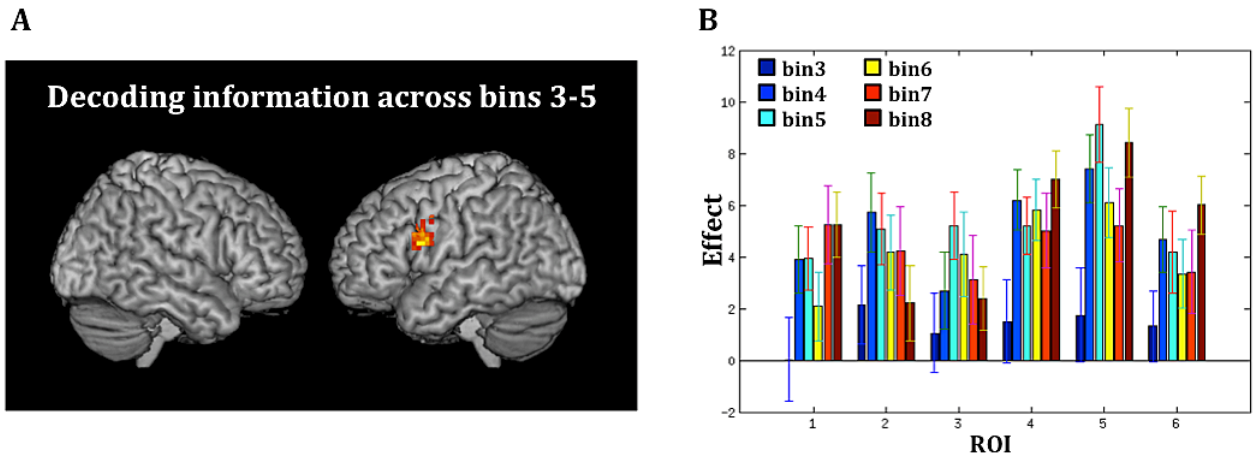


Figure 6.7 Conjunction analysis across all time bins (3, 4, and 5) considered (A). The only area that contained information about the rules throughout the entire cue and delay period was the left IFG. Results of the ROI analysis (B). All ROIs but basal ganglia (ROI 1) held information about the currently active rule.

The ROI analyses performed to assess whether the six areas encoded information about the currently active rule showed that all ROIs but basal ganglia held information about the rule (all *ps* < .05, Figure 6.7 B). In contrast, in the ROI analysis testing for differences in DA between groups with alterative interpretations of IF was not significant in any of the six ROIs considered.

6.4.3. Rule processing

A set of univariate analyses on data from the **target phase** (defined as the time interval from the onset of the target scene to the offset at subject response) examined brain activity while subjects evaluated the *validity* of distinct logical rules. We compared activations related to rules composed using different connectives to identify brain areas selectively involved in the evaluation of a specific connective. Figure 6.8 A shows the differences in activation when evaluating OR as compared to AND. We tested the effect of interest OR > AND and used the contrast OR > baseline (p < .05 uncorrected) as an inclusive mask to ensure that only voxels specifically active during the evaluation of the OR rule were considered. The contrast map was thresholded at p < .001 uncorrected at voxel-level and FWE-corrected at cluster level at p < .05. Reasoning with disjunction recruited different areas in the left hemisphere, in particular the posterior IFG (BA 44; peak: -45, 8, 32; BA 44/45; peak: -42, 23, 29), insula (peak: -30, 23, -4), a small posterior portion of the SMA (BA 6; peak: -6, 20, 44), and part of the IPL (BA 40; peak: -45, -40, 44).

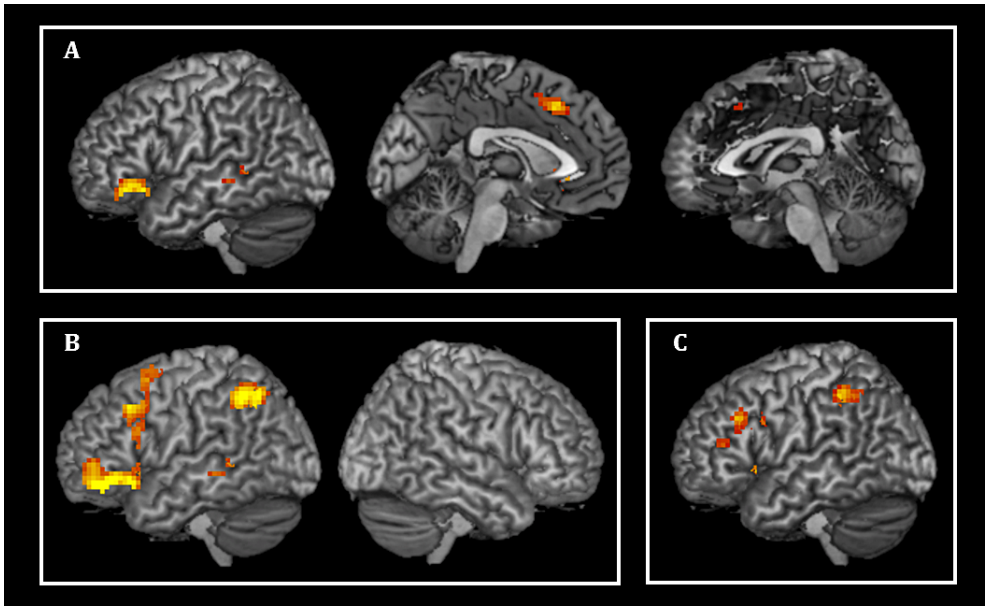


Figure 6.8 Brain regions active when evaluating either the OR (A) or the IF (B) rule compared to the AND rule. Disjunctions activated a left-lateralised network comprising: BA 6, 40, 44, and 45 and insula. Conditionals activated BA 21, 40, 44, 45, and 47. Brain areas active when evaluating the IF rule compared to both the AND and the OR rule (C). Two brain regions emerged: IFG and MTG.

We repeated a similar analysis for the IF rule (IF > AND masked by IF > baseline at p < .05 uncorrected). The contrast map was thresholded at p < .001 uncorrected at voxel-level and FWE-corrected at cluster level at p < .05. The evaluation of conditional rules activated a left-lateralised network comprising the IFG (BA 47; peak: -45, 38, -10; BA 44/45; peak: -45, 23, 41), the IPL (BA 40; peak: -45, -55, 50) and a portion of the left lateral temporal cortex (MTG, BA 21; peak: -48, -43, 5) (Figure 6.8 B).

We performed a conjunction analysis (see Paragraph 6.3.2) to assess whether any brain region was active both when precessing the IF compared to the AND rule and when evaluating the IF compared to the OR rule. Hence, we carried out the conjunction analysis considering the contrasts IF > AND and IF > OR (masked by IF > baseline at p < .05 uncorrected). The contrast map was thresholded at p < .001 uncorrected at voxel-level and FWE-corrected at cluster level at p < .05. Areas crucial for reasoning with conditionals proved to be the left anterior IFG (BA 47; peak: -48, 23, -10) and the left MTG (BA 21; peak: -51, -40, 2) (Figure 6.8 C).

A second set of analyses assessed whether alternative **interpretations** of the same connective result in differences in brain activity. First, we evaluated differences in activation between IF and AND (IF > AND, masked by IF > baseline, p < .05 uncorrected) in both the group with a classical interpretation of the conditional and in the group with

a conjunctive interpretation separately. The resulting map was thresholded at $p < .001$ uncorrected at voxel-level and FWE-corrected at $p < .05$ at cluster level. The first group showed a significant activation in the left IFG (BA 47; peak: -45, 38, -10) and in the IPL (BA 40; peak: -39, -58, 35) (Figure 6.9 A). In contrast, no brain areas emerged in the group with a conjunctive interpretation of the conditional, thus showing that the processing of the IF and AND rules was not different in this subgroup. Results from a similar analysis evaluating differences in activation between OR and AND (OR > AND, masked by OR > baseline, $p < .05$ uncorrected) in both the group with an inclusive interpretation of the disjunction and in the group with an exclusive interpretation were not significant.

The first set of **ROI analyses** evaluated the existence of differences in activation between either OR or IF and AND for each of the six ROIs. The results are shown in Figure 6.3 B (significant threshold $p < .05$). The difference in activation between IF and AND was significant in PG/SMA/FMG (BA 6/8), posterior IFG (BA 44/45), anterior IFG (BA 45/47), and mPC (BA 7). The difference between OR and AND showed a significant effect in PG/SMA/FMG (BA 6/8), posterior IFG (BA 44/45), and mPC (BA 7). In BA 45/47 the contrast IF > OR was also significant ($p < .05$).

A further ROI analysis evaluated potential differences in activation between groups with alternative *interpretations* of the same connective. The contrast $IF_{clas} > IF_{conj}$ was significant in BA 45/47 and mPC (all $p < .05$, Figure 6.9 B). In contrast, the difference in activation between OR_{inc} and OR_{exc} was not significant in any of the ROI considered.

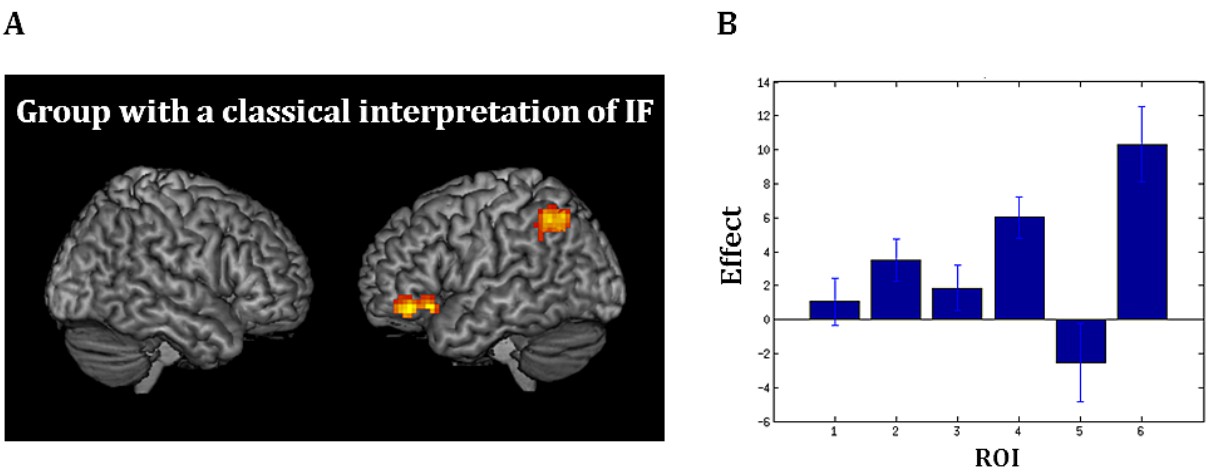


Figure 6.9 Brain areas showing higher levels of neural activity when evaluating the IF rule compared to the AND rule in the group with a classical interpretation of the conditional (A). Two brain regions emerged: IFG and IPL. Differences in activation between the group with a classical and the group with a conjunctive interpretation of IF for each ROI (B). The effect is significant in anterior IFG (ROI 4) and mPC (ROI 6).

6.5. Discussion

In this study we investigated the neural basis of the representation and evaluation of three different elementary logical connectives: *and*, *or*, and *If-then*. A crucial area for rule **representation** was the left *IFG* (BA 44), from which it was possible to decode rules composed using the three connectives throughout the whole delay period. This area is part of the Broca's complex (see Paragraph 6.1), a network central in natural language for processing the linguistic information contained in an utterance. In the literature on deductive reasoning, it has been argued that the function of BA 44 is either to perform *syntactic* operations on the premises in order to draw a valid conclusion (Prado et al., 2010; Reverberi et al., 2007) or simply to initially encode *verbal* stimuli (Monti & Osherson, 2012), since this region is implicated in retrieval and maintenance of rules only when they are subvocally rehearsed (Smith & Jonides, 1999; Wagner, Bunge, & Badre, 2004). However, two main aspects of our study rule out the latter hypothesis. First, we used visual cues coding for rules instead of verbal stimuli, hence no reading process was involved. Second, even if rules had been learnt as sentence strings (e.g., "There is a yellow square and there is a green circle"), when performing the experimental task subjects used a variety of strategies (including visual ones) to maintain the rules during the delay, as reported in the post-experimental questionnaires (see Paragraph 6.2.2 and Appendix C2). This makes it unlikely that subvocal rehearsal can account by itself for the role of BA 44 in rule representation. The first hypothesis, instead, argues that inferences (especially based on sentential connectives) rely on linguistic manipulations, implying that deduction is a language-based process. Moreover, the role of linguistic information is restricted to the "integration stage", at least for conditional reasoning (Reverberi et al., 2010). Such theoretical account was formulated based on the finding that BA 44/45 was active during inference generation but not during premise encoding in propositional reasoning. This result is apparently in contrast with the crucial role of BA 44 in rule representation emerged in the present study. However, it should be noted that in Reverberi and colleagues (2010) representation was assessed by comparing brain activity related to the presentation of the same premise when encoded either for a reasoning or for a memory task, thus emphasising either the exact structure or the meaning of the sentence. Instead, our task required to judge the validity of logical rules differing only in the specific connective

used; therefore, in the encoding phase the focus was on the formal structure of the rule that actually distinguished between the different logical rules. In a later work, the same Authors advanced the hypothesis that deduction is a *multi-componential* cognitive function, where different cognitive processes rely on separate brain regions. In particular, they proposed that BA 44/45 is involved in *extracting* the *formal* structure of a deductive problem while BA 47 is implicated in selecting and applying the appropriate inferential rules (Reverberi et al., 2012).

A further hypothesis has been proposed in the field of language processing based on an analogy with the role of other areas in PFC, that is to select and integrate pieces of information from perception or memory over time (Fuster, 2002; Koechlin & Summerfield, 2007). This hypothesis suggests that left IFG "is crucial for the unification operations required for binding single word information into larger structures" (Hagoort, 2005, p. 419) and, more broadly, to integrate lexical and non-linguistic information. The analogy with PFC is reinforced by the proposal of an anterior-to-posterior gradient in Broca's area functions, where BA 47/45, BA 44/45, and BA 44 and BA 6 contribute to semantic, syntactic, and phonological processing, respectively (Hagoort, 2005). Hence, the role of BA 44 in representing rules using logical connectives may be to integrate pieces of information (in our case the two propositions linked by the connective) into a comprehensive structure (the abstract form of the cued rule), regardless of whether this representation is linguistic or non-linguistic in nature. This interpretation is in line with the abovementioned explanation proposed by Reverberi and colleagues that BA 44/45 is involved in the "extraction and representation of the formal structure of premises during processing of deductive problems" (Reverberi et al., 2012, p. 1762). Moreover, it is supported by our finding that there is no difference in the representation of rules with the same formal structure but a different meaning (i.e., the IF rule in the group with a classical *vs.* a conjunctive interpretation of the conditional).

Other brain regions related to rule representation were found in occipital, temporal and parietal cortices. A previous study using the same paradigm for cueing rules with abstract symbols (Reverberi, Görgen, et al., 2012a) found that it was possible to decode symbol identity (i.e., activity specifically related to the visual features of the cues) in *occipital* areas and the left superior parietal lobe (BA 7). These results suggest that the early information in occipital regions found in the present study may reflect cognitive

processes related to the *encoding* of the *cue*. Parietal areas have been implicated in the representation of task sets (Bode & Haynes, 2009; Sakai & Passingham, 2003), the activation of the appropriate S-R associations (Cavina-Pratesi et al., 2006; Woolgar, Thompson, et al., 2011), rule maintenance (Bunge et al., 2003; Rowe et al., 2007), and translating the visual cue into meaning (i.e., the associated rule) (Reverberi, Görgen, et al., 2012a). Temporal regions have been associated with cue encoding (Reverberi, Görgen, et al., 2012b; Woolgar, Thompson, et al., 2011), rule retrieval (Bunge et al., 2003; Donohue et al., 2005), object recognition (Eichenbaum, Yonelinas, & Ranganath, 2007; Squire, Wixted, & Clark, 2007), concept representation (Martin, 2007; Mion et al., 2010; Visser, Jefferies, & Lambon Ralph, 2010), and language processing (Binder et al., 1997). In particular, the involvement of temporal areas in rule representation is considered to depend on the type of cue (visual *vs*. verbal) (Bunge et al., 2003) or task (spatial *vs*. verbal) (Sakai & Passingham, 2003), to the complexity of the rules (Bunge et al., 2003), or to the specific S-R association involved (direct association with a motor response *vs.* with a symbol) (Muhammad et al., 2006; Reverberi, Görgen, et al., 2012b). Therefore, we propose that the parietal and temporal areas identified in the present study are involved in the retrieval and representation of the meaning of the rules. In particular, *IPL* might support the *retrieval* of the relevant rule (i.e., the activation of the association between the symbol and the corresponding rule), while *MTG* might play a role in *storing* the meaning of the rule (i.e., the relation established between the two figures by the logical connective present in the rule).

Finally, an additional area in PFC where we could decode information about rules using connectives at a later stage was *FPC* (BA 10). Studies on deductive reasoning implicated this region in performing the logical operations necessary to assess the structure of a problem and derive valid conclusions from a set of premises (Monti & Osherson, 2012; Reverberi et al., 2009). FPC has been also associated with rule representation (Bunge et al., 2005, 2003; Sakai & Passingham, 2003), in particular when integration of information is needed, for example between opposing S-R mappings in hierarchies of rules (Bunge & Zelazo, 2006), multiple relations (Christoff et al., 2001; Kroger et al., 2002), or subtask outcomes and information maintained in WM (De Pisapia & Braver, 2008). It is also sensitive to an increasing in the amount of relational processing needed beyond relational integration (Bunge, Helskog, & Wendelken, 2009).

Therefore, if BA 44 may be involved in integrating pieces of information to identify the logical form of the rules, the role of BA 10 may be to *reconstruct* the relation established by the connective between that pieces and extract the exact *meaning* of the rule.

For the **evaluation** phase, we found that assessing the validity of the OR compared to the AND rule engaged a left-lateralised frontoparietal network comprising IFG (BA 44 and BA 44/45), insula, SMA (BA 6), and IPL (BA 40). Beyond BA 44/45 and the IPL (BA 40), evaluating the validity of the IF rule compared to the AND rule additionally recruited the left MTG (BA 21) and the anterior IFG (BA 47). These areas were identified to be crucial for reasoning with conditionals in the conjunction analysis, in line also with the results of the ROI analyses where the same ROIs showed a significant activation when evaluating either the conditional or the disjunction compared to the conjunction. Interestingly, processing the IF rule recruited *BA 47* significantly more than evaluating the OR rule. Previous studies also found that BA 47 is involved in deductive reasoning with conditional problems (Monti et al., 2007; Noveck et al., 2004; Prado et al., 2010; Reverberi et al., 2012) and in representing the identity of simple as well as compound conditional rules (Reverberi, Görgen, et al., 2012a, 2012b). Moreover, the recruitment of this area is specific for conditional arguments (Prado et al., 2010) and the level of activation is higher when assessing the truth of linguistic expressions implying higher cognitive load (Hagoort, Hald, Bastiaansen, & Petersson, 2004) or when evaluating more complex logical problems, for example Modus Tollens compared to Modus Ponens (Noveck et al., 2004). Put together, these results suggest that BA 47 is crucial in processing conditional rules compared to the easier disjunctive and conjunctive rules.

As stated above referring to Hagoort's theory on language processing, BA 47 is also thought to play a role in semantic unification, that is in integrating the meaning of incoming words in the representation of the preceding context (Hagoort, 2005). Similarly, the function of this region in evaluating the validity of rules using logical connectives may be to *integrate* incoming information (the target scene) into the current context (established by the rule) in order to reconstruct the overall meaning (for the validity judgement). The special involvement of BA 47 in dealing with the semantic aspects of the rule is supported by the fact that we found no difference in activation between rules with the same meaning but a different formal structure (i.e., the AND and the IF rules in the group with a conjunctive interpretation of the conditional).

A *peculiar* aspect of our experimental design is that we recruited two groups of individuals with different logical **interpretations** of the conditional: either a classical or a conjunctive interpretation. This allowed us to examine potential differences in the activation profiles underlying different ways of representing and processing the same stimulus. As already mentioned, no difference emerged between the two groups in the representation phase. In the evaluation phase, instead, in the group with a classical interpretation of the *conditional*, the left IFG (BA 47) and the IPL (BA 40) were more active when evaluating the IF compared to the AND rule. In contrast, no difference in activation emerged between the two experimental conditions in the group with a conjunctive interpretation of the conditional, suggesting that rules with identical meaning (i.e., sharing the same truth table) are processed in the same way, irrespective of their logical structure. In an ROI analysis, we directly compared the activation profiles of the two groups; IFG (BA 45/47) and mPC (BA 7) were significantly more active in individuals with a classical interpretation of the conditional rule. This result is particularly interesting: showing a difference in how rules with the same formal structure but a different meaning are interpreted, it suggests that the relevant factor that discriminate between different types of rules in rule processing is the *meaning* rather than the formal structure of the rule.

We also investigated potential differences in the activation profiles of subjects with different interpretations of the *disjunction*. No difference emerged between the group with an inclusive and the group with an exclusive interpretation of the OR rule, neither in the univariate nor in the ROI analyses. However, since these two groups are very different in size (the interpretation of the disjunction was not a selection criterion for participant recruitment) and the group comprising subjects with an exclusive interpretation of the disjunction is very small (N = 9), this null result might be due to lack of power of the present study and should not be considered conclusive in this respect.

Taken together, our results are in line with the findings from Reverberi et al. 2012 (Reverberi et al., 2012) and support the hypothesis of a multi-componential **model** where different brain areas are involved in distinct stages of the cognitive process. In particular, in an early *representational* stage, BA 44/45 extracts the formal structure of the logical rule; then, in the subsequent *reasoning* stage, BA 47 is critical for the

reconstruction of the meaning and the application of the inferential rule. In fact, we found BA 44 to hold information about the rules during the entire encoding phase and BA 47 to be active during the evaluation but not during the representation phase. Moreover, the presence of a difference in the activation profiles between groups with alternative interpretations of the conditional in the evaluation but not in the representation phase, suggests the possibility that different factors are relevant in distinct phases of rule processing: when rules have to be encoded, the *logical structure* is crucial (supported by the absence of differences between the representations of rules with the same logical structure but different meaning – i.e., the conditional in the groups with alternative interpretations of the IF rule), while during rule processing the *meaning* of the rule is more important (supported by the fact that there was no difference in activation between rules sharing the same meaning but with a different formal structure – i.e., the conjunction and the conditional in the group with a conjunctive interpretation of the IF rule – while rules with the same formal structure but different meaning showed differences in activation). Moreover, investigating the encoding phase (where a simple rule was represented and no additional information had to be integrated to derive an inference) separately from the evaluation phase, different from previous studies, in the present one we could actually isolate the two processes involved and assess which areas control the pure representation of logical rules.

However, it is important to point out also the **limitations** of the current study. One potential shortcoming is that we included as stimuli just *three* elementary logical *connectives* presented in isolation. While this choice guarantees the benefit of clarity and simplicity, at the same time it limits the generalisability of our findings to the combined and more complex use of logical connectives typical of both abstract logic and natural language statements. For example, we didn't consider negation, an equally important operator in propositional logic. The inclusion of negation would have allowed us to make further interesting comparisons between rules with the same meaning (identical truth-functionality) but a different logical structure. Another limitation is that, for pragmatic reasons, we had to limit our investigation to only *two* of the possible *interpretations* of the conditional, excluding a third common interpretation (i.e., the biconditional one). In addition, for the same reasons, we couldn't include enough

participants with an exclusive interpretation of the disjunction to ensure enough *power* for the analyses.

A final shortcoming is that we used only *simple* logical rules. Since the activation of some brain areas is modulated by the difficulty of the task to be performed or by the complexity of the information to be processed (Hagoort, Hald, Bastiaansen, & Petersson, 2004; Monti et al., 2007; Noveck et al., 2004), it is likely that more complex task sets will involve additional cortical areas. A study with progressively more complex logical forms and higher cognitive requirements, within the same paradigm and using the same stimuli, would be necessary to explore this possibility.

6.6. Conclusion

This study provides unique insight into the neural basis of representing and reasoning with elementary *propositional logic*. Specifically, we found different brain areas within the left IFG involved in the representation and evaluation of rules using logical connectives, with BA 44 crucial for the representation and BA 47 for the evaluation of the logical rules. Similarly to the hypothesis of a multi-componential model of deduction advanced by Reverberi et al. 2012, we propose that separate brain regions are involved in different stages of logical rule processing. In particular, the function of BA 44 would be to generate a *representation* of the *formal* structure of the rule, while BA 47 would be involved in the subsequent reconstruction of the *meaning* of that rule. In addition, we were able to investigate the activation profiles underlying different *interpretations* (meanings) of the same stimuli (logical rules).

Although we found a left-lateralised network for rule representation and processing including areas within the IFG commonly associated with linguistic functions, we caution against asserting that reasoning is a language-based process. Language comprehension and production are multifaceted cognitive functions and reconstructing the meaning involves the integration of many non-linguistic aspects of information (e.g. world knowledge based on episodic and semantic memory, Hagoort, 2005). In support of this position, Broca's area, that was shown to be crucial for the representation and processing of propositional rules, has been found to be activated also by non-linguistic tasks (Binkofski et al., 1999; Fink et al., 2006; Hamzei et al., 2003; Thoenissen, Zilles, &

Toni, 2002). Moreover, we didn't use verbal stimuli but rather visual cues coding for rules, thus greatly reducing the linguistic component of the task. Therefore, the recruitment of these linguistic areas in our study may account for more abstract computational properties of these regions, such as information integration. These processes may be necessary, but not limited to, both reasoning and discourse comprehension. At the same time, we do not make claims for a reasoning system completely separate from language or other cognitive domains. It has been pointed out that the same brain area is likely to participate in multiple functions or play a role in more than one network (Carpenter, Just, & Reichle, 2000; Elliott, 2003; Goldman-Rakic, Cools, & Srivastava, 1996; Mesulam, 1998; Sheth et al., 2012). Therefore, presupposing dedicated systems, such as a "language" or a "reasoning" system, may be a misleading theoretical framework for interpreting results. Instead, we suggest that using a *bottom-up* approach to investigate the basic components and computations shared by these complex domains provides a more valid and accurate method to understand their neural basis.

CONCLUSION

"Insanity is doing the same thing over and over again and expecting different results."

(Rita Mae Brown (1983), Sudden Death, p. 68)

This conclusive section attempts to provide an organic description of the results emerged from the experiments presented in this thesis. The **general idea** that inspired the single studies was to improve our understanding of how we select and apply rules to determine our actions and to react flexibly to a continuously changing environment.

The first study investigated *which* rules entail associations that can be retrieved *automatically* whenever the proper stimulus is presented. We assessed whether subliminally presented information of which participants were not aware could trigger inferences when matching one of the terms in a previously presented rule. Different inferential schemata were used, such as valid or invalid logical inferences (e.g., Modus Ponens, Disjunctive Syllogism, Modus Tollens), spatial relations, and quantified sentences. The only inference that could be triggered automatically was Modus Ponens, suggesting that the rule involved (i.e., conditional) is processed differently compared to the other types of associations we considered. In light of this result, we speculated that, if conditional associations were processed automatically, they could cause interference effects even when simply encoded in WM and irrelevant for the task. In the second study, we tested whether condition-action rules instructed on a trial-by-trial basis (thus not retrieved from long-term memory) and promptly marked as irrelevant for the task to be performed may cause *interference* effects. The results showed a detrimental effect of the irrelevant rule when *involuntarily* triggered by target stimuli matching the condition in the rule and requiring a response conflicting with the correct one. This finding corroborates the hypothesis that some rule-based inferences are automatic. The rule sets used in the first two studies were very simple in that they required the evaluation or application of a single rule at a time. However, in everyday life it is often necessary to integrate multiple rules to choose an appropriate course of action. Therefore, in the third study we considered more *complex* task sets in which rules at different levels in

the hierarchy of action control had to be integrated in order to respond. fMRI and multivariate pattern classification (Bode & Haynes, 2009; Haynes & Rees, 2006; Norman et al., 2006; Pereira et al., 2009) were used to identify brain regions that encoded information about different types of rules, either when they were represented alone or when they were encoded jointly. We added a further layer of higher-level conditional rules to a set of associations similar to those used in the first two studies. High-level rules influenced the application of the simple low-level rules. The results showed that rules with the same formal structure but at different levels in the hierarchy of action control are encoded in partially different brain areas, in keeping with recent findings on the functional organisation of PFC. However, our results did not support the existence of a functional gradient in PFC as conceived by the main theories on the role of this area in cognitive control. Additionally, we showed that different regions encoded rule information when the rules were represented in isolation or were encoded together with a different type of rule. An interesting result was that FPC (a region typically associated with high-level control functions) encoded low-level information during rule integration. We proposed two alternative hypotheses on the role of FPC in rule representation that should be tested to clarify this aspect. In this study, the two types of rules were both conditionals, thus they shared the same logical form. However, different cognitive processes may be involved in the evaluation of different logical rules (Knauff et al., 2002; Monti et al., 2009; Reverberi et al., 2007, 2012), as the first study also demonstrated (Reverberi, Pischedda, et al., 2012). The fourth study tested whether this difference in processing between various logical rules was present also at the neural level. We used rules with a *different logical form* (i.e., conjunctive, disjunctive, and conditional rules) to assess whether brain areas involved in rule representation and evaluation varied depending on the specific rule used. The results showed that different regions in PFC are involved during rule representation (BA 44) and processing (BA 47) and that reasoning with conditional rules activates areas involved in rule evaluation to a higher degree.

Taken together, our results suggest that **conditional rules** hold a "special" status in the human cognitive system. First, they are processed in a different way compared to other types of associations, as the first study confirmed. Second, functional differences are reflected by neural dissimilarities in representation and evaluation between

different logical rules, as the last study showed. This finding represents a valuable contribution to our understanding of the mechanisms underlying rule-guided behaviour.

However, this work presents also some **shortcomings**. A first limitation resides in the possible *aspects* taken into account. As already pointed out in the introduction, we considered only some of the processes involved in rule usage (for example, we didn't investigate rule learning and storage) and only a small set of all possible associations that might guide action. Moreover, we focused on the last stages of rule implementation, neglecting processes related, for example, to the perception and encoding of the stimulus. Nevertheless, in the discussion section of each study we already proposed some interesting research questions or working hypotheses that could broaden the range of aspects considered by the present work. Another shortcoming related to the last two studies derives from the fact that fMRI as a research *method* has various limitations, mainly in terms of the questions that can be possibly addressed using this technique, as highlighted in the method section (Paragraph 2.2). Finally, we used a small set of abstract rules entailing arbitrary associations between stimuli and actions. In everyday life, we are often confronted with larger and more complex rule sets in which meaningful associations link stimuli and actions. Therefore, experiments that better resemble the actual situations in which we act according to rules would provide a more *ecologically* valid perspective on the issue here addressed. However, the level of complexity of real-life situations makes this goal hard to accomplish. Hence, combining results obtained using simplified experimental tasks allowing for investigating only a single aspect involved in the process or using progressively more complex tasks, might be the only viable method to pinpoint the building blocks of rule-guided behaviour.

To conclude, rules are a useful tool for representing effectively meaningful variables and relations, allowing us to react properly to a dynamic environment and thus to behave meaningfully. The **importance** of rules for our behaviour becomes much more evident when insanity or brain injury prevents us to flexibly apply them to adapt to a situation that has changed, producing maladaptive behaviours such as perseveration. Hence, researching to improve our understanding of the cognitive mechanisms underlying the selection and implementation of rules for action is indeed an important, useful, and valuable goal.

ACKNOWLEDGEMENTS

During the last few years, many people have been telling me that you need to be mean, selfish, and better than everyone else to be successful in science. I'm still not convinced though… I like to think about science as climbing a mountain in a roped party, where you are glad to share the credit for reaching the summit with every single person in your group. Here I want to thank all those people who climbed the mountain with me.

First, I'd like to thank my advisers Prof. Paolo Cherubini and Dr. Carlo Reverberi for letting me collaborate in many of the interesting projects they were working on, for introducing me to neuroscience, for entrusting me with responsibilities, for believing in my capabilities, for allowing me to spend a significant part of my PhD abroad, and for the long-distance conversations when I was lost. They actually built that mountain for me.

Prof. John-Dylan Haynes deserves my gratitude for giving me the opportunity to work in his awesome lab at the Bernstein Centre for Computational Neuroscience for one and a half year and for providing generous resources to carry out my research. I'm especially thankful to him for challenging me and trying to get the best out of me; he taught me a lot, in many ways.

I'm very grateful to all people in Haynes' lab for being always so nice, friendly, helpful, supportive, and patient with me, for giving me part of their precious time, for making the most complex concepts easily understandable for me, for never minding when I bothered them, for all the inspiring and stimulating conversations we had, and for making me feel part of the lab, even an important one. I had the privilege to work in such a great group and it would be extremely hard to find another equally good one. In particular, I'd like to thank Kai Görgen who more closely collaborated with me in all the projects we carried out when I was in Berlin. He denoted an incredible patience but also a lot of enthusiasm in teaching me all the technical skills I needed and in helping me to understand the "nerdish" language; his help throughout the projects, his smart suggestions during brainstorming, his encouragement and support, and his precise and insightful comments on my thesis represent an invaluable contribution to this work.

A special thanks goes to my colleague and friend Carsten Bogler, who pushed me to start writing this thesis organising writing sessions at the library, gave me valuable advice and tips for working at sustained levels and for finding motivation, helped me to figure out relevant steps of the analyses, and gave honest, precise, and insightful comments on the method chapter, allowing me to learn a huge amount of information and to make the concepts in my mind clearer, greatly improving the quality of my writing. I'm extremely grateful to him for being always frank, for stating his opinions firmly, for giving me an objective perspective on what I was facing, for making me aware of my errors, and for "shaking" me when I was going crazy, but also for encouraging me and for calming me down when I was in panic. His contribution was essential for me to achieve the final goal.

I owe many thanks also to Anna Blumenthal who worked very hard for the last study described in this thesis and put a lot of passion in it. I'm thankful to her for keeping me alert reporting promptly all the problems as they emerged and especially for all the fruitful conversations, for making my days in the lab and at the scanner much funnier, and for discovering and boosting my "social" side.

I'm very grateful to Raffaele Cabras for reading part of the thesis and pointing out the obscure paragraphs even though he was not familiar with the topic, for helping me to make my pictures even more attractive, for chasing me when I moved, and especially for being by my side all the time, for showing enthusiasm and admiration whenever I achieved an important professional goal, and for always encouraging my work aspirations wherever they would have led me. Indeed, these are all priceless qualities that I greatly appreciated and could profit from.

I'd like to thank my precious friend Emanuela Maggioni with whom I could share my troubles, who always supported my plans, believed I could make it, kept me focused on my goal, tried to take me away from home when I was working too much, reprimanded me when I messed up all my natural rhythms, and has always been there for me in the most difficult times.

I'm extremely thankful to my amazing family that still hasn't figured out which my job is but always shared my joy and was proud of me whenever something good in my career happened, supporting me no matter how far I was moving or how little time I could spend with them. In particular, I thank my sister for being my first fan, for

reminding me how good I am, and for making me smile whenever she asks me whether I'll ever start to look for a "real" job. Without their support and confidence in my abilities it would have been very hard to keep my motivation going.

Finally, I'm thankful to all people (friends, colleagues, outstanding professors, and committed and bright researchers) who inspired, supported, and helped me during my PhD and I cannot name here one by one; I got something valuable from any of them that made me a more capable scientist and a better and happier person.

REFERENCES

Aguirre, G. K. (2011). Experimental Design and Data Analysis for fMRI. In S. H. Faro, F. B. Mohamed, M. Law, & J. T. Ulmer (Eds.), *Functional Neuroradiology* (pp. 321–330). Boston, MA, US: Springer US.

Allport, A., Styles, E. A., & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In C. Umiltà & M. Moscovitch (Eds.), *Attention and performance XV: Conscious and nonconscious information processing* (pp. 421–452). Cambridge, MA, US: The MIT Press.

Allwein, E. L., Schapire, R. E., & Singer, Y. (2001). Reducing multiclass to binary: a unifying approach for margin classifiers. *The Journal of Machine Learning Research*, *1*, 113–141. doi:10.1162/15324430152733133

Amiez, C., Kostopoulos, P., Champod, A.-S., & Petrides, M. (2006). Local Morphology Predicts Functional Organization of the Dorsal Premotor Region in the Human Brain. *The Journal of Neuroscience*, *26*(10), 2724–2731. doi:10.1523/JNEUROSCI.4739-05.2006

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2004). Inhibition and the right inferior frontal cortex. *Trends in Cognitive Sciences*, *8*(4), 170–177. doi:10.1016/j.tics.2004.02.010

Asaad, W. F., Rainer, G., & Miller, E. K. (2000). Task-specific neural activity in the primate prefrontal cortex. *Journal of Neurophysiology*, *84*(1), 451–459.

Badre, D. (2008). Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes. *Trends in Cognitive Sciences*, *12*(5), 193–200. doi:10.1016/j.tics.2008.02.004

Badre, D., & D'Esposito, M. (2007). Functional Magnetic Resonance Imaging Evidence for a Hierarchical Organization of the Prefrontal Cortex. *Journal of Cognitive Neuroscience*, *19*(12), 2082–2099. doi:10.1162/jocn.2007.19.12.2082

Badre, D., & D'Esposito, M. (2009). Is the rostro-caudal axis of the frontal lobe hierarchical? *Nature Reviews Neuroscience*, *10*(9), 659–669. doi:10.1038/nrn2667

Badre, D., Doll, B. B., Long, N. M., & Frank, M. J. (2012). Rostrolateral Prefrontal Cortex and Individual Differences in Uncertainty-Driven Exploration. *Neuron*, *73*(3), 595–607. doi:10.1016/j.neuron.2011.12.025

Badre, D., & Frank, M. J. (2012). Mechanisms of Hierarchical Reinforcement Learning in Cortico-Striatal Circuits 2: Evidence from fMRI. *Cerebral Cortex*, *22*(3), 527–536. doi:10.1093/cercor/bhr117

Bandettini, P. A. (2009). Functional MRI Limitations and Aspirations. In E. Kraft, B. Gulyás, & E. Pöppel (Eds.), *Neural correlates of thinking*. Berlin, Heidelberg, DE: Springer-Verlag.

Baron, S. G., & Osherson, D. N. (2011). Evidence for conceptual combination in the left anterior temporal lobe. *NeuroImage*, *55*(4), 1847–1852. doi:10.1016/j.neuroimage.2011.01.066

Bengtsson, S. L., Haynes, J.-D., Sakai, K., Buckley, M. J., & Passingham, R. E. (2009). The representation of abstract task rules in the human prefrontal cortex. *Cerebral Cortex*, *19*(8), 1929–1936. doi:10.1093/cercor/bhn222

Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M., & Prieto, T. (1997). Human Brain Language Areas Identified by Functional Magnetic Resonance Imaging. *The Journal of Neuroscience*, *17*(1), 353–362.

Binkofski, F., Buccino, G., Stephan, K. M., Rizzolatti, G., Seitz, R. J., & Freund, H.-J. (1999). A parieto-premotor network for object manipulation: evidence from neuroimaging. *Experimental Brain Research*, *128*(1-2), 210–213. doi:10.1007/s002210050838

Bode, S., & Haynes, J.-D. (2009). Decoding sequential stages of task preparation in the human brain. *NeuroImage*, *45*(2), 606–613. doi:10.1016/j.neuroimage.2008.11.031

Bogler, C., Bode, S., & Haynes, J.-D. (2011). Decoding Successive Computational Stages of Saliency Processing. *Current Biology*, *21*(19), 1667–1671. doi:10.1016/j.cub.2011.08.039

Bonnefond, M., & Van Der Henst, J.-B. (2013). Deduction electrified: ERPs elicited by the processing of words in conditional arguments. *Brain and Language*, *124*(3), 244–256. doi:10.1016/j.bandl.2012.12.011

Boorman, E. D., Behrens, T. E. J., Woolrich, M. W., & Rushworth, M. F. S. (2009). How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action. *Neuron*, *62*(5), 733–743. doi:10.1016/j.neuron.2009.05.014

Botvinick, M. M. (2007). Conflict monitoring and decision making: Reconciling two perspectives on anterior cingulate function. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(4), 356–366. doi:10.3758/CABN.7.4.356

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. *Trends in Cognitive Sciences*, *8*(12), 539–546. doi:10.1016/j.tics.2004.10.003

Braine, M. D. S., & O'Brien, D. P. (1998). *Mental logic*. Mahwah, NY, US: Lawrence Erlbaum Associates.

Brass, M., Derrfuss, J., & von Cramon, D. Y. (2008). The Role of the Posterior Frontolateral Cortex in Task-Related Control. In S. A. Bunge & J. D. Wallis (Eds.), *Neuroscience of rule-guided behavior* (pp. 177–196). Oxford, NY, US: Oxford University Press.

Brass, M., Ruge, H., Meiran, N., Rubin, O., Koch, I., Zysset, S., … von Cramon, D. Y. (2003). When the same response has different meanings: *NeuroImage*, *20*(2), 1026–1031. doi:10.1016/S1053-8119(03)00357-4

Brass, M., Ullsperger, M., Knoesche, T. R., von Cramon, D. Y., & Phillips, N. A. (2005). Who Comes First? The Role of the Prefrontal and Parietal Cortex in Cognitive Control. *Journal of Cognitive Neuroscience*, *17*(9), 1367–1375. doi:10.1162/0898929054985400

Brass, M., & von Cramon, D. Y. (2002). The Role of the Frontal Cortex in Task Preparation. *Cerebral Cortex*, *12*(9), 908–914. doi:10.1093/cercor/12.9.908

Brass, M., & von Cramon, D. Y. (2004). Decomposing Components of Task Preparation with Functional Magnetic Resonance Imaging. *Journal of Cognitive Neuroscience*, *16*(4), 609–620. doi:10.1162/089892904323057335

Braver, T. S., & Bongiolatti, S. R. (2002). The Role of Frontopolar Cortex in Subgoal Processing during Working Memory. *NeuroImage*, *15*(3), 523–536. doi:10.1006/nimg.2001.1019

Brett, M., Anton, J. L., Valabregue, R., & Poline, J.-B. (2002). Region of interest analysis using the MarsBar toolbox for SPM 99. *NeuroImage*, *16*, S497.

Brouwer, G. J., & Heeger, D. J. (2009). Decoding and Reconstructing Color from Responses in Human Visual Cortex. *Journal of Neuroscience*, *29*(44), 13992–14003.

Brown, R. M. (1983). *Sudden Death*. New York, NY, US: Bantam Books.

Buckley, M. J., Mansouri, F. A., Hoda, H., Mahboubi, M., Browning, P. G. F., Kwok, S. C., … Tanaka, K. (2009). Dissociable Components of Rule-Guided Behavior Depend on Distinct Medial and Prefrontal Regions. *Science*, *325*(5936), 52–58. doi:10.1126/science.1172377

Buckner, R. L., & Koutstaal, W. (1998). Functional neuroimaging studies of encoding, priming, and explicit memory retrieval. *Proceedings of the National Academy of Sciences*, *95*(3), 891–898. doi:10.1073/pnas.95.3.891

Buhmann, M. D. (2009). *Radial basis functions*. Cambridge, UK: Cambridge University Press.

Bullmore, E., Brammer, M., Williams, S. C. R., Rabe-Hesketh, S., Janot, N., David, A., … Sham, P. (1996). Statistical methods of estimation and inference for functional MR image analysis. *Magnetic Resonance in Medicine*, *35*(2), 261–277. doi:10.1002/mrm.1910350219

Bunge, S. A. (2004). How we use rules to select actions: A review of evidence from cognitive neuroscience. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(4), 564–579. doi:10.3758/CABN.4.4.564

Bunge, S. A., Helskog, E. H., & Wendelken, C. (2009). Left, but not right, rostrolateral prefrontal cortex meets a stringent test of the relational integration hypothesis. *NeuroImage*, *46*(1), 338–342. doi:10.1016/j.neuroimage.2009.01.064

Bunge, S. A., Kahn, I., Wallis, J. D., Miller, E. K., & Wagner, A. D. (2003). Neural Circuits Subserving the Retrieval and Maintenance of Abstract Rules. *Journal of Neurophysiology*, *90*(5), 3419–3428. doi:10.1152/jn.00910.2002

Bunge, S. A., & Wallis, J. D. (2008). *Neuroscience of rule-guided behavior*. Oxford, NY, US: Oxford University Press.

Bunge, S. A., Wallis, J. D., Parker, A., Brass, M., Crone, E. A., Hoshi, E., & Sakai, K. (2005). Neural Circuitry Underlying Rule Use in Humans and Nonhuman Primates. *Journal of Neuroscience*, *25*(45), 10347–10350. doi:10.1523/JNEUROSCI.2937-05.2005

Bunge, S. A., & Zelazo, P. D. (2006). A Brain-Based Account of the Development of Rule Use in Childhood. *Current Directions in Psychological Science*, *15*(3), 118–121. doi:10.1111/j.0963-7214.2006.00419.x

Burgess, G. C., & Braver, T. S. (2010). Neural Mechanisms of Interference Control in Working Memory: Effects of Interference Expectancy and Fluid Intelligence. *PLoS ONE*, *5*(9), e12861.

Bussey, T. J., Wise, S. P., & Murray, E. A. (2002). Interaction of ventral and orbital prefrontal cortex with inferotemporal cortex in conditional visuomotor learning. *Behavioral Neuroscience*, *116*(4), 703–715. doi:10.1037/0735-7044.116.4.703

Buxton, R. B. (2002). *Introduction to functional magnetic resonance imaging principles and techniques*. Cambridge, UK: Cambridge University Press.

Buxton, R. B., Wong, E. C., & Frank, L. R. (1998). Dynamics of blood flow and oxygenation changes during brain activation: The balloon model. *Magnetic Resonance in Medicine*, *39*(6), 855–864. doi:10.1002/mrm.1910390602

Byrne, R. M. J. (1989). Suppressing valid inferences with conditionals. *Cognition*, *31*(1), 61–83. doi:10.1016/0010-0277(89)90018-8

Byrne, R. M. J., & Johnson-Laird, P. N. (2009). "If" and the problems of conditional reasoning. *Trends in Cognitive Sciences*, *13*(7), 282–287. doi:10.1016/j.tics.2009.04.003

Canessa, N., Gorini, A., Cappa, S. F., Piattelli-Palmarini, M., Danna, M., Fazio, F., & Perani, D. (2005). The effect of social content on deductive reasoning: An fMRI study. *Human Brain Mapping*, *26*(1), 30–43. doi:10.1002/hbm.20114

Carpenter, P., Just, M. A., & Reichle, E. D. (2000). Working memory and executive function: evidence from neuroimaging. *Current Opinion in Neurobiology*, *10*(2), 195–199. doi:10.1016/S0959-4388(00)00074-X

Cavina-Pratesi, C., Valyear, K. F., Culham, J. C., Köhler, S., Obhi, S. S., Marzi, C. A., & Goodale, M. A. (2006). Dissociating Arbitrary Stimulus-Response Mapping from Movement Planning during Preparatory Period: Evidence from Event-Related Functional Magnetic Resonance Imaging. *The Journal of Neuroscience*, *26*(10), 2704–2713. doi:10.1523/JNEUROSCI.3176-05.2006

Chao, L. L., & Martin, A. (2000). Representation of Manipulable Man-Made Objects in the Dorsal Stream. *NeuroImage*, *12*(4), 478–484. doi:10.1006/nimg.2000.0635

Charron, S., & Koechlin, E. (2010). Divided Representation of Concurrent Goals in the Human Frontal Lobes. *Science*, *328*(5976), 360–363. doi:10.1126/science.1183614

Chevallier, C., Noveck, I. A., Nazir, T., Bott, L., Lanzetti, V., & Sperber, D. (2008). Making disjunctions exclusive. *The Quarterly Journal of Experimental Psychology*, *61*(11), 1741–1760. doi:10.1080/17470210701712960

Christoff, K., & Gabrieli, J. D. E. (2000). The frontopolar cortex and human cognition: Evidence for a rostrocaudal hierarchical organization within the human prefrontal cortex. *Psychobiology*, *28*(2), 168–186.

Christoff, K., & Keramatian, K. (2007). Abstraction of mental representations: theoretical considerations and neuroscientific evidence. In S. A. Bunge & J. D. Wallis (Eds.), *Perspectives on Rule-Guided Behavior* (pp. 107– 126). Oxford, UK: Oxford University Press.

Christoff, K., Keramatian, K., Gordon, A. M., Smith, R., & Mädler, B. (2009). Prefrontal organization of cognitive control according to levels of abstraction. *Brain Research*, *1286*, 94–105. doi:10.1016/j.brainres.2009.05.096

Christoff, K., Prabhakaran, V., Dorfman, J., Zhao, Z., Kroger, J. K., Holyoak, K. J., & Gabrieli, J. D. E. (2001). Rostrolateral Prefrontal Cortex Involvement in Relational Integration during Reasoning. *NeuroImage*, *14*(5), 1136–1149. doi:10.1006/nimg.2001.0922

Christoff, K., Ream, J. M., Geddes, L. P. T., & Gabrieli, J. D. E. (2003). Evaluating Self-Generated Information: Anterior Prefrontal Contributions to Human Cognition. *Behavioral Neuroscience*, *117*(6), 1161–1168. doi:10.1037/0735-7044.117.6.1161

Constantinidis, C., Franowicz, M. N., & Goldman-Rakic, P. S. (2001). The sensory nature of mnemonic representation in the primate prefrontal cortex. *Nature Neuroscience*, *4*, 311–316. doi:10.1038/85179

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), 201–215. doi:10.1038/nrn755

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, *19*(2), 261–270.

Crittenden, B. M., & Duncan, J. (2012). Task Difficulty Manipulation Reveals Multiple Demand Activity but no Frontal Lobe Hierarchy. *Cerebral Cortex*, 1–9. doi:10.1093/cercor/bhs333

Crone, E. A., Donohue, S. E., Honomichl, R., Wendelken, C., & Bunge, S. A. (2006). Brain Regions Mediating Flexible Rule Use during Development. *The Journal of Neuroscience*, *26*(43), 11239–11247. doi:10.1523/JNEUROSCI.2165-06.2006

Dalrymple-Alford, E. C. (1972). Associative facilitation and interference in the Stroop color-word task. *Perception & Psychophysics*, *11*(4), 274–276.

Dalrymple-Alford, E. C., & Budayr, B. (1966). Examination of some aspects of the Stroop color-word test. *Perceptual and Motor Skills*, *23*(3f), 1211–1214.

Damadian, R. (1971). Tumor Detection by Nuclear Magnetic Resonance. *Science*, *171*(3976), 1151–1153. doi:10.1126/science.171.3976.1151

Danziger, K. (1980). Wundt's Theory of Behavior and Volition. In R. W. Rieber (Ed.), *Wilhelm Wundt and the Making of a Scientific Psychology* (pp. 89–115). Boston, MA, US: Springer US.

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., & Formisano, E. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *NeuroImage*, *43*(1), 44–58. doi:10.1016/j.neuroimage.2008.06.037

De Neys, W., & Franssens, S. (2009). Belief inhibition during thinking: Not always winning but at least taking part. *Cognition*, *113*(1), 45–61. doi:10.1016/j.cognition.2009.07.009

De Pisapia, N., & Braver, T. S. (2008). Preparation for integration: the role of anterior prefrontal cortex in working memory: *NeuroReport*, *19*(1), 15–19. doi:10.1097/WNR.0b013e3282f31530

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences*, *10*(5), 204–211. doi:10.1016/j.tics.2006.03.007

Dell'Acqua, R., & Grainger, J. (1999). Unconscious semantic priming from pictures. *Cognition*, *73*(1), B1–B15. doi:10.1016/S0010-0277(99)00049-9

Diamond, A. (2006). The Early Development of Executive Functions. In E. Bialystok & I. M. C. Fergus (Eds.), *Lifespan cognition: Mechanisms of change* (pp. 70–95). New York, NY, US: Oxford University Press.

Dias, R., Robbins, T. W., & Roberts, A. C. (1996). Dissociation in prefrontal cortex of affective and attentional shifts. *Nature*, *380*(6569), 69–72. doi:10.1038/380069a0

Donohue, S. E., Wendelken, C., Crone, E. A., & Bunge, S. A. (2005). Retrieving rules for behavior from long-term memory. *NeuroImage*, *26*(4), 1140–1149. doi:10.1016/j.neuroimage.2005.03.019

Dosenbach, N. U. F., Visscher, K. M., Palmer, E. D., Miezin, F. M., Wenger, K. K., Kang, H. C., … Petersen, S. E. (2006). A Core System for the Implementation of Task Sets. *Neuron*, *50*(5), 799–812. doi:10.1016/j.neuron.2006.04.031

Dove, A., Pollmann, S., Schubert, T., Wiggins, C. J., & von Cramon, D. Y. (2000). Prefrontal cortex activation in task switching: an event-related fMRI study. *Cognitive Brain Research*, *9*(1), 103–109. doi:10.1016/S0926-6410(99)00029-4

Draine, S. C., & Greenwald, A. G. (1998). Replicable unconscious semantic priming. *Journal of Experimental Psychology: General*, *127*(3), 286–303.

Dreher, J.-C., & Berman, K. F. (2002). Fractionating the neural substrate of cognitive control processes. *Proceedings of the National Academy of Sciences*, *99*(22), 14595–14600. doi:10.1073/pnas.222193299

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). New York, NY, US: Wiley.

Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. *Nature Reviews Neuroscience*, *2*(11), 820–829. doi:10.1038/35097557

Efron, B. (1983). Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *Journal of the American Statistical Association*, *78*(382), 316–331. doi:10.1080/01621459.1983.10477973

Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The Medial Temporal Lobe and Recognition Memory. *Annual Review of Neuroscience*, *30*(1), 123–152. doi:10.1146/annurev.neuro.30.051606.094328

Elliott, R. (2003). Executive functions and their disorders. *British Medical Bulletin*, *65*(1), 49–59. doi:10.1093/bmb/65.1.49

Etkin, A., Klemenhagen, K. C., Dudman, J. T., Rogan, M. T., Hen, R., Kandel, E. R., & Hirsch, J. (2004). Individual Differences in Trait Anxiety Predict the Response of the Basolateral Amygdala to Unconsciously Processed Fearful Faces. *Neuron*, *44*(6), 1043–1055. doi:10.1016/j.neuron.2004.12.006

Evans, J. S. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: the psychology of deduction*. Hove, UK: Lawrence Erlbaum Associates.

Fink, G. R., Manjaly, Z. M., Stephan, K. E., Gurd, J. M., Zilles, K., & Amunts, K. (2006). A role from Broca's area beyond language processing: evidence from neuropsychology and fMRI. In Y. Grodzinsky & K. Amunts (Eds.), *Broca's region*. Oxford, UK: Oxford University Press.

Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A., & Noll, D. C. (1995). Improved Assessment of Significant Activation in Functional Magnetic Resonance Imaging (fMRI): Use of a Cluster-Size Threshold. *Magnetic Resonance in Medicine*, *33*(5), 636–647. doi:10.1002/mrm.1910330508

Formisano, E., De Martino, F., & Valente, G. (2008). Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning. *Magnetic Resonance Imaging*, *26*(7), 921–934.

Frank, M. J., & Badre, D. (2012). Mechanisms of Hierarchical Reinforcement Learning in Corticostriatal Circuits 1: Computational Analysis. *Cerebral Cortex*, *22*(3), 509–526. doi:10.1093/cercor/bhr114

Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between frontal cortex and basal ganglia in working memory: A computational model. *Cognitive, Affective, & Behavioral Neuroscience*, *1*(2), 137–160.

Friston, K. J., Ashburner, J., Frith, C. D., Poline, J.-B., Heather, J. D., & Frackowiak, R. S. J. (1995). Spatial registration and normalization of images. *Human Brain Mapping*, *3*(3), 165–189.

Friston, K. J., Ashburner, J., Kiebel, S., Nichols, T. E., & Penny, W. D. (2006). *Statistical parametric mapping: the analysis of funtional brain images* (1st ed.). London, UK: Elsevier.

Friston, K. J., Glaser, D. E., Henson, R. N. A., Kiebel, S., Phillips, C., & Ashburner, J. (2002). Classical and Bayesian Inference in Neuroimaging: Applications. *NeuroImage*, *16*(2), 484–512. doi:10.1006/nimg.2002.1091

Friston, K. J., Holmes, A. P., Poline, J.-B., Grasby, P. J., Williams, S. C. R., Frackowiak, R. S. J., & Turner, R. (1995). Analysis of fMRI Time-Series Revisited. *NeuroImage*, *2*(1), 45–53.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., & Frackowiak, R. S. J. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, *2*(4), 189–210.

Friston, K. J., Josephs, O., Zarahn, E., Holmes, A. P., Rouquette, S., & Poline, J.-B. (2000). To Smooth or Not to Smooth? *NeuroImage*, *12*(2), 196–208. doi:10.1006/nimg.2000.0609

Friston, K. J., Penny, W. D., & Glaser, D. E. (2005). Conjunction revisited. *NeuroImage*, *25*(3), 661–667. doi:10.1016/j.neuroimage.2005.01.013

Friston, K. J., Stephan, K. E., Lund, T. E., Morcom, A., & Kiebel, S. (2005). Mixed-effects and fMRI studies. *NeuroImage*, *24*(1), 244–252. doi:10.1016/j.neuroimage.2004.08.055

Funahashi, S., & Takeda, K. (2002). Information Processes in the Primate Prefrontal Cortex in Relation to Working Memory Processes. *Reviews in the Neurosciences*, *13*(4), 313–345. doi:10.1515/REVNEURO.2002.13.4.313

Fuster, J. M. (1997). Network memory. *Trends in Neurosciences*, *20*(10), 451–459. doi:10.1016/S0166-2236(97)01128-4

Fuster, J. M. (2000). Executive frontal functions. *Experimental Brain Research*, *133*(1), 66–70. doi:10.1007/s002210000401

Fuster, J. M. (2002). Physiology of Executive Functions: The Perception-Action Cycle. In D. T. Stuss & R. T. Knight (Eds.), *Principles of Frontal Lobe Function* (pp. 96–108). New York, NY, US: Oxford University Press.

Gade, M., & Koch, I. (2007). The influence of overlapping response sets on task inhibition. *Memory & Cognition*, *35*(4), 603–609. doi:10.3758/BF03193298

Garavan, H., Ross, T. J., Kaufman, J., & Stein, E. A. (2003). A midline dissociation between error-processing and response-conflict monitoring. *NeuroImage*, *20*(2), 1132–1139. doi:10.1016/S1053-8119(03)00334-3

Genovese, C. R., Lazar, N. A., & Nichols, T. E. (2002). Thresholding of Statistical Maps in Functional Neuroimaging Using the False Discovery Rate. *NeuroImage*, *15*(4), 870–878. doi:10.1006/nimg.2001.1037

Genovesio, A., Brasted, P. J., Mitz, A. R., & Wise, S. P. (2005). Prefrontal cortex activity related to abstract response strategies. *Neuron*, *47*(2), 307.

Goel, V. (2007). Anatomy of deductive reasoning. *Trends in Cognitive Sciences*, *11*(10), 435–441. doi:10.1016/j.tics.2007.09.003

Goldman-Rakic, P. S., Cools, A. R., & Srivastava, K. (1996). The Prefrontal Landscape: Implications of Functional Architecture for Understanding Human Mentation and the Central Executive [and Discussion]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *351*(1346), 1445–1453. doi:10.1098/rstb.1996.0129

Golland, P., & Fischl, B. (2003). Permutation Tests for Classification: Towards Statistical Significance in Image-Based Studies. In C. Taylor & J. A. Noble (Eds.), *Information Processing in Medical Imaging* (Vol. 2732, pp. 330–341). Berlin, Heidelberg, DE: Springer-Verlag.

Greenwald, A. G. (1992). New Look 3: Unconscious cognition reclaimed. *American Psychologist*, *47*(6), 766–779. doi:10.1037/0003-066X.47.6.766

Grosset, N., & Barrouillet, P. (2003). On the nature of mental models of conditional: The case of **If** , **If then** , and **Only if**. *Thinking & Reasoning*, *9*(4), 289–306. doi:10.1080/1354678034000240

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, *46*(1-3), 389–422.

Hadj-Bouziane, F., & Boussaoud, D. (2003). Neuronal activity in the monkey striatum during conditional visuomotor learning. *Experimental Brain Research*, *153*(2), 190–196. doi:10.1007/s00221-003-1592-4

Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends in Cognitive Sciences*, *9*(9), 416–423. doi:10.1016/j.tics.2005.07.004

Hagoort, P. (2009). Reflections on the Neurobiology of Syntax. In D. Bickerton & E. Szathmáry (Eds.), *Biological foundations and origin of syntax* (pp. 279–297). Cambridge, MA, US: The MIT Press.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of Word Meaning and World Knowledge in Language Comprehension. *Science*, *304*(5669), 438–441. doi:10.1126/science.1095455

Hamers, J. F., & Lambert, W. E. (1972). Bilingual interdependencies in auditory perception. *Journal of Verbal Learning and Verbal Behavior*, *11*(3), 303–310.

Hampshire, A., Thompson, R., Duncan, J., & Owen, A. M. (2010). Lateral Prefrontal Cortex Subregions Make Dissociable Contributions during Fluid Reasoning. *Cerebral Cortex*, *21*(1), 1–10. doi:10.1093/cercor/bhq085

Hamzei, F., Rijntjes, M., Dettmers, C., Glauche, V., Weiller, C., & Büchel, C. (2003). The human action recognition system and its relationship to Broca's area: an fMRI study. *NeuroImage*, *19*(3), 637–644. doi:10.1016/S1053-8119(03)00087-9

Handley, S., J., Newstead, S. E., & Trippas, D. (2011). Logic, beliefs, and instruction: A test of the default interventionist account of belief bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 28–43. doi:10.1037/a0021098

Hanson, S. J., & Halchenko, Y. O. (2008). Brain Reading Using Full Brain Support Vector Machines for Object Recognition: There Is No "Face" Identification Area. *Neural Computation*, *20*(2), 486–503.

Harshbarger, T. B., & Song, A. W. (2008). Differentiating Sensitivity of Post-Stimulus Undershoot under Diffusion Weighting: Implication of Vascular and Neuronal Hierarchy. *PLoS ONE*, *3*(8), e2914. doi:10.1371/journal.pone.0002914

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., & Bießmann, F. (in press). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*. doi:10.1016/j.neuroimage.2013.10.067

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, *293*(5539), 2425–2430.

Haynes, J.-D., & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience*, *8*(5), 686–691.

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, *7*(7), 523–534.

Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., & Passingham, R. E. (2007). Reading Hidden Intentions in the Human Brain. *Current Biology*, *17*(4), 323–328.

Heeger, D. J., & Ress, D. (2002). What does fMRI tell us about neuronal activity? *Nature Reviews Neuroscience*, *3*(2), 142–151.

Henson, R. N. A. (2004). Analysis of fMRI time series. In R. S. J. Frackowiak (Ed.), *Human brain function* (2nd ed.). San Diego, CA, US: Elsevier Science.

Henson, R. N. A., & Friston, K. J. (2006). Convolution models for fMRI. In K. J. Friston, J. Ashburner, S. Kiebel, T. E. Nichols, & W. D. Penny (Eds.), *Statistical parametric mapping: the analysis of functional brain images* (1st ed., pp. 178–192). London, UK: Elsevier.

Henson, R. N. A., Rugg, M. D., & Friston, K. J. (2001). The choice of basis functions in event-related fMRI. *NeuroImage*, *13*(6), 149. doi:10.1016/S1053-8119(01)91492-2

Hoshi, E., Shima, K., & Tanji, J. (1998). Task-dependent selectivity of movement-related neuronal activity in the primate prefrontal cortex. *Journal of Neurophysiology*, *80*(6), 3392–3397.

Hoshi, E., Shima, K., & Tanji, J. (2000). Neuronal activity in the primate prefrontal cortex in the process of motor selection based on two behavioral rules. *Journal of Neurophysiology*, *83*(4), 2355–2373.

Houdé, O., Zago, L., Mellet, E., Moutier, S., Pineau, A., Mazoyer, B., & Tzourio-Mazoyer, N. (2000). Shifting from the Perceptual Brain to the Logical Brain: The Neural Impact of Cognitive Inhibition Training. *Journal of Cognitive Neuroscience*, *12*(5), 721–728. doi:10.1162/089892900562525

Huettel, S. A., McKeown, M. J., Song, A. W., Hart, S., Spencer, D. D., Allison, T., & McCarthy, G. (2004). Linking Hemodynamic and Electrophysiological Measures of Brain Activity: Evidence from Functional MRI and Intracranial Field Potentials. *Cerebral Cortex*, *14*(2), 165–173. doi:10.1093/cercor/bhg115

Huettel, S. A., Song, A. W., & McCarthy, G. (2009). *Functional magnetic resonance imaging* (2nd ed.). Sunderland, MA, US: Sinauer Associates.

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks*, *13*(4-5), 411–430.

Jersild, A. T. (1927). Mental set and shift. *Archives of Psychology*, *14*(89), 5–82.

Johansen-Berg, H., Dawes, H., Guy, C., Smith, S. M., Wade, D. T., & Matthews, P. M. (2002). Correlation between motor improvements and altered fMRI activity after rehabilitative therapy. *Brain*, *125*(12), 2731–2742. doi:10.1093/brain/awf282

Johnson-Laird, P. N. (2006). *How we reason*. Oxford, NY, US: Oxford University Press.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hove, UK: Lawrence Erlbaum Associates.

Johnson-Laird, P. N., & Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review*, *109*(4), 646–678. doi:10.1037/0033-295X.109.4.646

Johnson-Laird, P. N., Byrne, R. M., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review*, *99*(3), 418–439. doi:10.1037/0033-295X.99.3.418

Jonides, J., & Mack, R. (1984). On the cost and benefit of cost and benefit. *Psychological Bulletin*, *96*(1), 29–44.

Kamitani, Y., & Sawahata, Y. (2010). Spatial smoothing hurts localization but not information: Pitfalls for brain mappers. *NeuroImage*, *49*(3), 1949–1952.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679–685.

Kandel, E. R. (2000). Two Opposing Views Have Been Advanced on the Relationship Between Brain and Behavior. In *Principles of neural science* (4th ed., pp. 5–6). New York: McGraw-Hill, Health Professions Division.

Kant, I. (1785). *Grundlegung zur Metaphysik der Sitten*. New York, NY, US: Harper.

Kellenbach, M. L., Brett, M., & Patterson, K. (2003). Actions Speak Louder Than Functions: The Importance of Manipulability and Action in Tool Representation. *Journal of Cognitive Neuroscience*, *15*(1), 30–46. doi:10.1162/089892903321107800

Kennerley, S. W., Sakai, K., & Rushworth, M. F. S. (2004). Get your ADInstruments Research Organization of Action Sequences and the Role of the Pre-SMA. *Journal of Neurophysiology*, *91*(2), 978–993.

Kerns, J. G., Cohen, J. D., MacDonald, A. W., Cho, R. Y., Stenger, V. A., & Carter, C. S. (2004). Anterior Cingulate Conflict Monitoring and Adjustments in Control. *Science*, *303*(5660), 1023–1026. doi:10.1126/science.1089910

Kimberg, D. Y., Aguirre, G. K., & D'Esposito, M. (2000). Modulation of task-related neural activity in task-switching: an fMRI study. *Cognitive Brain Research*, *10*(1-2), 189–196. doi:10.1016/S0926-6410(00)00016-1

Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review*, *107*(4), 852–884.

Knauff, M., Mulack, T., Kassubek, J., Salih, H. R., & Greenlee, M. W. (2002). Spatial imagery in deductive reasoning: a functional MRI study. *Cognitive Brain Research*, *13*(2), 203–212. doi:10.1016/S0926-6410(01)00116-1

Koechlin, E., Basso, G., Pietrini, P., Panzer, S., & Grafman, J. (1999). The role of the anterior prefrontal cortex in human cognition. *Nature*, *399*, 148–151. doi:10.1038/20178

Koechlin, E., & Hyafil, A. (2007). Anterior Prefrontal Function and the Limits of Human Decision-Making. *Science*, *318*(5850), 594–598. doi:10.1126/science.1142995

Koechlin, E., Ody, C., & Kouneiher, F. (2003). The Architecture of Cognitive Control in the Human Prefrontal Cortex. *Science*, *302*(5648), 1181–1185. doi:10.1126/science.1088545

Koechlin, E., & Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*, *11*(6), 229–235. doi:10.1016/j.tics.2007.04.005

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1-2), 273–324. doi:10.1016/S0004-3702(97)00043-X

Kouider, S., & Dehaene, S. (2007). Levels of processing during non-conscious perception: a critical review of visual masking. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 857–875. doi:10.1098/rstb.2007.2093

Kouider, S., & Dehaene, S. (2009). Subliminal Number Priming Within and Across the Visual and Auditory Modalities. *Experimental Psychology (formerly "Zeitschrift Für Experimentelle Psychologie")*, *56*(6), 418–433.

Kouider, S., & Dupoux, E. (2001). A functional disconnection between spoken and visual word recognition: evidence from unconscious priming. *Cognition*, *82*(1), B35–B49. doi:10.1016/S0010-0277(01)00152-4

Kouneiher, F., Charron, S., & Koechlin, E. (2009). Motivation and cognitive control in the human prefrontal cortex. *Nature Neuroscience*, *12*(7), 939–945. doi:10.1038/nn.2321

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, *103*(10), 3863–3868.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(4), 1–28. doi:10.3389/neuro.06.004.2008

Kroger, J. K., Sabb, F. W., Fales, C. L., Bookheimer, S. Y., Cohen, M. S., & Holyoak, K. J. (2002). Recruitment of Anterior Dorsolateral Prefrontal Cortex in Human Reasoning: a Parametric Study of Relational Complexity. *Cerebral Cortex*, *12*(5), 477–485. doi:10.1093/cercor/12.5.477

Krüger, G., Kastrup, A., & Glover, G. H. (2001). Neuroimaging at 1.5 T and 3.0 T: Comparison of oxygenation-sensitive magnetic resonance imaging. *Magnetic Resonance in Medicine*, *45*(4), 595–604. doi:10.1002/mrm.1081

Kuhl, B. A., Dudukovic, N. M., Kahn, I., & Wagner, A. D. (2007). Decreased demands on cognitive control reveal the neural processing benefits of forgetting. *Nature Neuroscience*, *10*(7), 908–914. doi:10.1038/nn1918

Lachenbruch, P. A., & Mickey, M. R. (1968). Estimation of Error Rates in Discriminant Analysis. *Technometrics*, *10*(1), 1–11. doi:10.1080/00401706.1968.10490530

Lauterbur, P. C. (1973). Image Formation by Induced Local Interactions: Examples Employing Nuclear Magnetic Resonance. *Nature*, *242*(5394), 190–191.

Lea, R. B. (1995). On-line evidence for elaborative logical inferences in text. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(6), 1469–1482. doi:10.1037/0278-7393.21.6.1469

Lindsay, D. S., & Jacoby, L. L. (1994). Stroop process dissociations: The relationship between facilitation and interference. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(2), 219–234.

Liston, C., Miller, M. M., Goldwater, D. S., Radley, J. J., Rocher, A. B., Hof, P. R., … McEwen, B. S. (2006). Stress-Induced Alterations in Prefrontal Cortical Dendritic Morphology Predict Selective Impairments in Perceptual Attentional Set-Shifting. *The Journal of Neuroscience*, *26*(30), 7870–7874. doi:10.1523/JNEUROSCI.1184-06.2006

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature*, *453*(7197), 869–878.

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

Logothetis, N. K., & Wandell, B. A. (2004). Interpreting the BOLD Signal. *Annual Review of Physiology*, *66*(1), 735–769. doi:10.1146/annurev.physiol.66.082602.092845

Lu, C.-H., & Proctor, R. W. (1995). The influence of irrelevant location information on performance: A review of the Simon and spatial Stroop effects. *Psychonomic Bulletin & Review*, *2*(2), 174–207.

Lund, T. E., Madsen, K. H., Sidaros, K., Luo, W.-L., & Nichols, T. E. (2006). Non-white noise in fMRI: Does modelling have an impact? *NeuroImage*, *29*(1), 54–66. doi:10.1016/j.neuroimage.2005.07.005

MacDonald, A. W., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the Role of the Dorsolateral Prefrontal and Anterior Cingulate Cortex in Cognitive Control. *Science*, *288*(5472), 1835–1838. doi:10.1126/science.288.5472.1835

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*(2), 163–203.

Mahmoudi, A., Takerkart, S., Regragui, F., Boussaoud, D., & Brovelli, A. (2012). Multivoxel Pattern Analysis for fMRI Data: A Review. *Computational and Mathematical Methods in Medicine*, *2012*, 1–14.

Mansfield, P. (1977). Multi-planar image formation using NMR spin echoes. *Journal of Physics C: Solid State Physics*, *10*(3), L55–L58.

Martin, A. (2007). The Representation of Object Concepts in the Brain. *Annual Review of Psychology*, *58*(1), 25–45. doi:10.1146/annurev.psych.57.102904.190143

Meiran, N. (1996). Reconfiguration of processing mode prior to task performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1423–1442. doi:10.1037/0278-7393.22.6.1423

Meiran, N. (2000). Modeling cognitive control in task-switching. *Psychological Research*, *63*(3-4), 234–249. doi:10.1007/s004269900004

Mesulam, M.-M. (1998). From sensation to cognition. *Brain*, *121*(6), 1013–1052. doi:10.1093/brain/121.6.1013

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227–234. doi:10.1037/h0031564

Miller, E. K. (2000). The prefrontal cortex and cognitive control. *Nature Reviews Neuroscience*, *1*(1), 59–65. doi:10.1038/35036228

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*(1), 167–202. doi:10.1146/annurev.neuro.24.1.167

Mion, M., Patterson, K., Acosta-Cabronero, J., Pengas, G., Izquierdo-Garcia, D., Hong, Y. T., … Nestor, P. J. (2010). What the left and right anterior fusiform gyri tell us about semantic memory. *Brain*, *133*(11), 3256–3268. doi:10.1093/brain/awq272

Misaki, M., Kim, Y., Bandettini, P. A., & Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage*, *53*(1), 103–118. doi:10.1016/j.neuroimage.2010.05.051

Mitchell, T. M. (2010). *Machine learning*. New York, NY, US: McGraw-Hill.

Mitchell, T. M., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M., & Newman, S. (2004). Learning to Decode Cognitive States from Brain Images. *Machine Learning*, *57*(1-2), 145–175. doi:10.1023/B:MACH.0000035475.85309.1b

Momennejad, I., & Haynes, J.-D. (2012). Human anterior prefrontal cortex encodes the "what" and "when" of future intentions. *NeuroImage*, *61*(1), 139–148. doi:10.1016/j.neuroimage.2012.02.079

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. doi:10.1016/S1364-6613(03)00028-7

Monti, M. M. (2011). Statistical Analysis of fMRI Time-Series: A Critical Review of the GLM Approach. *Frontiers in Human Neuroscience*, *5*(28), 1–13. doi:10.3389/fnhum.2011.00028

Monti, M. M., & Osherson, D. N. (2012). Logic, language and the brain. *Brain Research*, *1428*, 33–42. doi:10.1016/j.brainres.2011.05.061

Monti, M. M., Osherson, D. N., Martinez, M. J., & Parsons, L. M. (2007). Functional neuroanatomy of deductive inference: A language-independent distributed network. *NeuroImage*, *37*(3), 1005–1016. doi:10.1016/j.neuroimage.2007.04.069

Monti, M. M., Parsons, L. M., & Osherson, D. N. (2009). The boundaries of language and thought in deductive inference. *Proceedings of the National Academy of Sciences*, *106*(30), 12554–12559. doi:10.1073/pnas.0902422106

Moonen, C. T. W., & Bandettini, P. A. (Eds.). (1999). *Functional MRI*. Berlin, Heidelberg, DE: Springer-Verlag.

Moseley, M. E., & Glover, G. H. (1995). Functional MR imaging. Capabilities and limitations. *Neuroimaging Clinics of North America*, *5*(2), 161–191.

Mourão-Miranda, J., Bokde, A. L. W., Born, C., Hampel, H., & Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. *NeuroImage*, *28*(4), 980–995.

Mourão-Miranda, J., Reynaud, E., McGlone, F., Calvert, G., & Brammer, M. (2006). The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. *NeuroImage*, *33*(4), 1055–1065. doi:10.1016/j.neuroimage.2006.08.016

Muhammad, R., Wallis, J. D., & Miller, E. K. (2006). A Comparison of Abstract Rules in the Prefrontal Cortex, Premotor Cortex, Inferior Temporal Cortex, and Striatum. *Journal of Cognitive Neuroscience*, *18*(6), 974–989. doi:10.1162/jocn.2006.18.6.974

Mukamel, R., Gelbard, H., Arieli, A., Hasson, U., Fried, I., & Malach, R. (2005). Coupling Between Neuronal Firing, Field Potentials, and fMRI in Human Auditory Cortex. *Science*, *309*(5736), 951–954. doi:10.1126/science.1110913

Muller, K.-R., Mika, S., Ratsch, G., Tsuda, K., & Scholkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, *12*(2), 181–201. doi:10.1109/72.914517

Murray, E. A., Bussey, T. J., & Wise, S. P. (2000). Role of prefrontal cortex in a network for arbitrary visuomotor mapping. In W. X. Schneider, A. M. Owen, & J. Duncan (Eds.), *Executive Control and the Frontal Lobe: Current Issues* (pp. 114–129). Berlin, Heidelberg, DE: Springer-Verlag.

Naccache, L., & Dehaene, S. (2001a). The Priming Method: Imaging Unconscious Repetition Priming Reveals an Abstract Representation of Number in the Parietal Lobes. *Cerebral Cortex*, *11*(10), 966–974.

Naccache, L., & Dehaene, S. (2001b). Unconscious semantic priming extends to novel unseen stimuli. *Cognition*, *80*(3), 215–229.

Nachev, P., Wydell, H., O'Neill, K., Husain, M., & Kennard, C. (2007). The role of the pre-supplementary motor area in the control of action. *NeuroImage*, *36*(Supplement 2), T155–T163. doi:10.1016/j.neuroimage.2007.03.034

Nakamura, K., Sakai, K., & Hikosaka, O. (1998). Neuronal activity in medial frontal cortex during learning of sequential procedures. *Journal of Neurophysiology*, *80*(5), 2671–2687.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9*(3), 353–383.

Nee, D. E., & Brown, J. W. (2012a). Dissociable Frontal-Striatal and Frontal-Parietal Networks Involved in Updating Hierarchical Contexts in Working Memory. *Cerebral Cortex*, *23*(9), 2146–2158. doi:10.1093/cercor/bhs194

Nee, D. E., & Brown, J. W. (2012b). Rostral–caudal gradients of abstraction revealed by multi-variate pattern analysis of working memory. *NeuroImage*, *63*(3), 1285–1294. doi:10.1016/j.neuroimage.2012.08.034

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, *106*(3), 226–254. doi:10.1037/0096-3445.106.3.226

Newstead, S. E., Handley, S. J., Harley, C., Wright, H., & Farrelly, D. (2004). Individual differences in deductive reasoning. *The Quarterly Journal of Experimental Psychology Section A*, *57*(1), 33–60. doi:10.1080/02724980343000116

Nichols, T. E., Brett, M., Andersson, J., Wager, T., & Poline, J.-B. (2005). Valid conjunction inference with the minimum statistic. *NeuroImage*, *25*(3), 653–660. doi:10.1016/j.neuroimage.2004.12.005

Nichols, T. E., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical Methods in Medical Research*, *12*(5), 419–446. doi:10.1191/0962280203sm341ra

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, *15*(1), 1–25. doi:10.1002/hbm.1058

Niessing, J., Ebisch, B., Schmidt, K. E., Niessing, M., Singer, W., & Galuske, R. A. W. (2005). Hemodynamic Signals Correlate Tightly with Synchronized Gamma Oscillations. *Science*, *309*(5736), 948–951. doi:10.1126/science.1110948

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430.

Noveck, I. A., Goel, V., & Smith, K. W. (2004). The Neural Basis of Conditional Reasoning with Arbitrary Content. *Cortex*, *40*(4-5), 613–622. doi:10.1016/S0010-9452(08)70157-6

O'Doherty, J., Deichmann, R., & Dolan, R. J. (2003). Dissociating valence of outcome from behavioral control in human orbital and ventral prefrontal cortices. *The Journal of Neuroscience*, *23*(21), 7931–7939.

O'Reilly, R. C., Noelle, D. C., Braver, T. S., & Cohen, J. D. (2002). Prefrontal Cortex and Dynamic Categorization Tasks: Representational Organization and Neuromodulatory Control. *Cerebral Cortex*, *12*(3), 246–257. doi:10.1093/cercor/12.3.246

Ogawa, S., & Lee, T.-M. (1990). Magnetic resonance imaging of blood vessels at high fields:In vivo andin vitro measurements and image simulation. *Magnetic Resonance in Medicine*, *16*(1), 9–18. doi:10.1002/mrm.1910160103

Ogawa, S., Lee, T.-M., Nayak, A. S., & Glynn, P. (1990). Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magnetic Resonance in Medicine*, *14*(1), 68–78.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113. doi:10.1016/0028-3932(71)90067-4

Op de Beeck, H. P. (2010). Against hyperacuity in brain reading: Spatial smoothing does not hurt multivariate fMRI analyses? *NeuroImage*, *49*(3), 1943–1948.

Parsons, L. M., & Osherson, D. N. (2001). New Evidence for Distinct Right and Left Brain Systems for Deductive versus Probabilistic Reasoning. *Cerebral Cortex*, *11*(10), 954–965. doi:10.1093/cercor/11.10.954

Passingham, R. E., Toni, I., & Rushworth, M. F. S. (2000). Specialisation within the prefrontal cortex: the ventral prefrontal cortex and associative learning. In W. X. Schneider, A. M. Owen, & J. Duncan (Eds.), *Executive Control and the Frontal Lobe: Current Issues* (pp. 103–113). Berlin, Heidelberg, DE: Springer-Verlag.

Pasupathy, A., & Miller, E. K. (2005). Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature*, *433*(7028), 873–876. doi:10.1038/nature03287

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-Driven Computations in Speech Processing. *Science*, *298*(5593), 604–607. doi:10.1126/science.1072901

Penny, W. D., & Holmes, A. P. (2004). Random-Effects Analysis. In R. S. J. Frackowiak (Ed.), *Human brain function* (2nd ed., pp. 843–850). San Diego, CA, US: Elsevier Science.

Pereira, F., & Botvinick, M. (2011). Information mapping with pattern classifiers: A comparative study. *NeuroImage*, *56*(2), 476–496. doi:10.1016/j.neuroimage.2010.05.026

Pereira, F., & Gordon, G. (2006). The support vector decomposition machine. *Proceedings of the 23rd International Conference on Machine Learning*, 689–696. doi:10.1145/1143844.1143931

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage*, *45*(1), S199–S209.

Pessiglione, M., Schmidt, L., Draganski, B., Kalisch, R., Lau, H., Dolan, R. J., & Frith, C. D. (2007). How the Brain Translates Money into Force: A Neuroimaging Study of Subliminal Motivation. *Science*, *316*(5826), 904–906. doi:10.1126/science.1140459

Petrides, M. (1985). Deficits on conditional associative-learning tasks after frontal- and temporal-lobe lesions in man. *Neuropsychologia*, *23*(5), 601–614. doi:10.1016/0028-3932(85)90062-4

Petrides, M. (1990). Nonspatial conditional learning impaired in patients with unilateral frontal but not unilateral temporal lobe excisions. *Neuropsychologia*, *28*(2), 137–149. doi:10.1016/0028-3932(90)90096-7

Petrides, M. (1997). Visuo-motor conditional associative learning after frontal and temporal lesions in the human brain. *Neuropsychologia*, *35*(7), 989–997. doi:10.1016/S0028-3932(97)00026-2

Petrides, M. (2002). The Mid-ventrolateral Prefrontal Cortex and Active Mnemonic Retrieval. *Neurobiology of Learning and Memory*, *78*(3), 528–538. doi:10.1006/nlme.2002.4107

Petrides, M. (2005). The Rostral-Caudal Axis of Cognitive Control within the Lateral Frontal Cortex. In S. Dehaene, J.-R. Duhamel, M. C. Hauser, & G. Rizzolatti (Eds.), *From Monkey Brain To Human Brain: A Fyssen Foundation Symposium* (pp. 293–314). Cambridge, MA, US: The MIT Press.

Petrides, M., & Baddeley, A. (1996). Specialized Systems for the Processing of Mnemonic Information within the Primate Frontal Cortex [and Discussion]. *Philosophical Transactions of the Royal Society, London, Series B*, *351*(1346), 1455–1462. doi:10.1098/rstb.1996.0130

Poldrack, R. A., Mumford, J. A., & Nichols, T. E. (2011). *Handbook of functional MRI data analysis*. New York, NY, US: Cambridge University Press.

Prado, J., Chadha, A., & Booth, J. R. (2011). The Brain Network for Deductive Reasoning: A Quantitative Meta-analysis of 28 Neuroimaging Studies. *Journal of Cognitive Neuroscience*, *23*(11), 3483–3497. doi:10.1162/jocn_a_00063

Prado, J., Kaliuzhna, M., Cheylus, A., & Noveck, I. A. (2008). Overcoming perceptual features in logical reasoning: An event-related potentials study. *Neuropsychologia*, *46*(11), 2629–2637. doi:10.1016/j.neuropsychologia.2008.04.017

Prado, J., Van Der Henst, J.-B., & Noveck, I. A. (2010). Recomposing a fragmented literature: How conditional and relational arguments engage different neural systems for deductive reasoning. *NeuroImage*, *51*(3), 1213–1221. doi:10.1016/j.neuroimage.2010.03.026

Quintana, J., & Fuster, J. M. (1999). From Perception to Action: Temporal Integrative Functions of Prefrontal and Parietal Neurons. *Cerebral Cortex*, *9*(3), 213–221. doi:10.1093/cercor/9.3.213

Rader, A. W., & Sloutsky, V. M. (2002). Processing of logically valid and logically invalid conditional inferences in discourse comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(1), 59–68. doi:10.1037/0278-7393.28.1.59

Rainer, G., Rao, S. C., & Miller, E. K. (1999). Prospective Coding for Objects in Primate Prefrontal Cortex. *The Journal of Neuroscience*, *19*(13), 5493–5505.

Rasmussen, J. (1982). Human errors. A taxonomy for describing human malfunction in industrial installations. *Journal of Occupational Accidents*, *4*(2-4), 311–333.

Reason, J. T. (1990). *Human error*. Cambridge, NY, US: Cambridge University Press.

Reverberi, C., Bonatti, L. L., Frackowiak, R. S. J., Paulesu, E., Cherubini, P., & Macaluso, E. (2012). Large scale brain activations predict reasoning profiles. *NeuroImage*, *59*(2), 1752–1764. doi:10.1016/j.neuroimage.2011.08.027

Reverberi, C., Cherubini, P., Frackowiak, R. S. J., Caltagirone, C., Paulesu, E., & Macaluso, E. (2010). Conditional and syllogistic deductive tasks dissociate functionally during premise integration. *Human Brain Mapping*, *31*(9), 1430–1445. doi:10.1002/hbm.20947

Reverberi, C., Cherubini, P., Rapisarda, A., Rigamonti, E., Caltagirone, C., Frackowiak, R. S. J., … Paulesu, E. (2007). Neural basis of generation of conclusions in elementary deduction. *NeuroImage*, *38*(4), 752–762. doi:10.1016/j.neuroimage.2007.07.060

Reverberi, C., Görgen, K., & Haynes, J.-D. (2012a). Compositionality of Rule Representations in Human Prefrontal Cortex. *Cerebral Cortex*, *22*(6), 1237–1246. doi:10.1093/cercor/bhr200

Reverberi, C., Görgen, K., & Haynes, J.-D. (2012b). Distributed Representations of Rule Identity and Rule Order in Human Frontal Cortex and Striatum. *The Journal of Neuroscience*, *32*(48), 17420–17430. doi:10.1523/JNEUROSCI.2344-12.2012

Reverberi, C., Pischedda, D., Burigo, M., & Cherubini, P. (2012). Deduction without awareness. *Acta Psychologica*, *139*(1), 244–253. doi:10.1016/j.actpsy.2011.09.011

Reverberi, C., Shallice, T., D'Agostini, S., Skrap, M., & Bonatti, L. L. (2009). Cortical bases of elementary deductive reasoning: Inference, memory, and metadeduction. *Neuropsychologia*, *47*(4), 1107–1116. doi:10.1016/j.neuropsychologia.2009.01.004

Reynolds, J. R., O'Reilly, R. C., Cohen, J. D., & Braver, T. S. (2012). The Function and Organization of Lateral Prefrontal Cortex: A Test of Competing Hypotheses. *PLoS ONE*, *7*(2), e30284. doi:10.1371/journal.pone.0030284

Rips, L. J. (1988). Deduction. In *The psychology of human thought* (pp. 116–152). New York, NY, US: Cambridge University Press.

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. Cambridge, MA, US: The MIT Press.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, *124*(2), 207–231. doi:10.1037/0096-3445.124.2.207

Rosenfeld, A., & Kak, A. C. (1982). *Digital picture processing* (2nd ed.). New York, NY, US: Academic Press.

Rowe, J. B., Sakai, K., Lund, T. E., Ramsoy, T., Christensen, M. S., Baare, W. F. C., … Passingham, R. E. (2007). Is the Prefrontal Cortex Necessary for Establishing Cognitive Sets? *The Journal of Neuroscience*, *27*(48), 13303–13310. doi:10.1523/JNEUROSCI.2349-07.2007

Rugg, M. D., & Wilding, E. L. (2000). Retrieval processing and episodic memory. *Trends in Cognitive Sciences*, *4*(3), 108–115. doi:10.1016/S1364-6613(00)01445-5

Rushworth, M. F. S., Croxson, P. L., Buckley, M. J., & Walton, M. E. (2008). Ventrolateral and Medial Frontal Contributions to Decision-Making and Action Selection. In S. A. Bunge & J. D. Wallis (Eds.), *Neuroscience of rule-guided behavior* (pp. 129–157). Oxford, NY, US: Oxford University Press.

Rushworth, M. F. S., Hadland, K. A., Paus, T., & Sipila, P. K. (2002). Role of the Human Medial Frontal Cortex in Task Switching: A Combined fMRI and TMS Study. *Journal of Neurophysiology*, *87*(5), 2577–2592. doi:10.1152/jn.00812.2001

Rushworth, M. F. S., Johansen-Berg, H., Göbel, S. M., & Devlin, J. T. (2003). The left parietal and premotor cortices: motor attention and selection. *NeuroImage*, *20*(Supplement 1), S89–S100. doi:10.1016/j.neuroimage.2003.09.011

Rushworth, M. F. S., Noonan, M. P., Boorman, E. D., Walton, M. E., & Behrens, T. E. (2011). Frontal Cortex and Reward-Guided Learning and Decision-Making. *Neuron*, *70*(6), 1054–1069. doi:10.1016/j.neuron.2011.05.014

Rushworth, M. F. S., Walton, M. E., Kennerley, S., & Bannerman, D. (2004). Action sets and decisions in the medial frontal cortex. *Trends in Cognitive Sciences*, *8*(9), 410–417. doi:10.1016/j.tics.2004.07.009

Sakai, K. (2008). Task Set and Prefrontal Cortex. *Annual Review of Neuroscience*, *31*(1), 219–245. doi:10.1146/annurev.neuro.31.060407.125642

Sakai, K., & Passingham, R. E. (2003). Prefrontal interactions reflect future task operations. *Nature Neuroscience*, *6*(1), 75–81. doi:10.1038/nn987

Sakai, K., & Passingham, R. E. (2004). Prefrontal Selection and Medial Temporal Lobe Reactivation in Retrieval of Short-term Verbal Information. *Cerebral Cortex*, *14*(8), 914–921. doi:10.1093/cercor/bhh050

Sakai, K., & Passingham, R. E. (2006). Prefrontal Set Activity Predicts Rule-Specific Neural Processing during Subsequent Cognitive Performance. *The Journal of Neuroscience*, *26*(4), 1211–1218. doi:10.1523/JNEUROSCI.3887-05.2006

Schild, H. H. (1990). *MRI made easy: (... well almost)*. Berlin, DE: Schering AG.

Schumacher, E. H., Cole, M. W., & D'Esposito, M. (2007). Selection and maintenance of stimulus–response rules during preparation and performance of a spatial choice-reaction task. *Brain Research*, *1136*, 77–87. doi:10.1016/j.brainres.2006.11.081

Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M., & Pessoa, L. (2008). Measuring consciousness: relating behavioural and neurophysiological approaches. *Trends in Cognitive Sciences*, *12*(8), 314–321. doi:10.1016/j.tics.2008.04.008

Shallice, T., Burgess, P., & Robertson, I. (1996). The Domain of Supervisory Processes and Temporal Organization of Behaviour [and Discussion]. *Philosophical Transactions of the Royal Society, London, Series B*, *351*(1346), 1405–1412. doi:10.1098/rstb.1996.0124

Sheth, S. A., Mian, M. K., Patel, S. R., Asaad, W. F., Williams, Z. M., Dougherty, D. D., … Eskandar, E. N. (2012). Human dorsal anterior cingulate cortex neurons mediate ongoing behavioural adaptation. *Nature*, *488*(7410), 218–221. doi:10.1038/nature11239

Shima, K., Mushiake, H., & Tanji, J. (1996). Role for cells in the presupplementary motor area in updating motor plans. *Proceedings of the National Academy of Sciences*, *93*(16), 8694–8698.

Sichel, J. L., & Chandler, K. A. (1969). The Color-Word Interference Test: The Effects of Varied Color-Word Combinations Upon Verbal Response Latency. *The Journal of Psychology*, *72*(2), 219–231. doi:10.1080/00223980.1969.10543502

Smith, E. E., & Jonides, J. (1999). Storage and Executive Processes in the Frontal Lobes. *Science*, *283*(5408), 1657–1661. doi:10.1126/science.283.5408.1657

Smith, S. M. (2004). Overview of fMRI analysis. *British Journal of Radiology*, *77*(no. suppl 2), S167–S175.

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, *44*(1), 83–98. doi:10.1016/j.neuroimage.2008.03.061

Sohn, M.-H., Albert, M. V., Jung, K., Carter, C. S., & Anderson, J. R. (2007). Anticipation of conflict monitoring in the anterior cingulate cortex and the prefrontal cortex. *Proceedings of the National Academy of Sciences*, *104*(25), 10330–10334. doi:10.1073/pnas.0703225104

Sohn, M.-H., Ursu, S., Anderson, J. R., Stenger, V. A., & Carter, C. S. (2000). The role of prefrontal cortex and posterior parietal cortex in task switching. *Proceedings of the National Academy of Sciences*, *97*(24), 13448–13453. doi:10.1073/pnas.240460497

Soon, C. S., Brass, M., Heinze, H.-J., & Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, *11*(5), 543–545. doi:10.1038/nn.2112

Squire, L. R., Wixted, J. T., & Clark, R. E. (2007). Recognition memory and the medial temporal lobe: a new perspective. *Nature Reviews Neuroscience*, *8*(11), 872–883. doi:10.1038/nrn2154

Stelzer, J., Chen, Y., & Turner, R. (2013). Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control. *NeuroImage*, *65*, 69–82.

Stenning, K., & van Lambalgen, M. (2008). *Human Reasoning and Cognitive Science*. Cambridge, MA, US: The MIT Press.

Stevenson, R. J., & Over, D. E. (1995). Deduction from Uncertain Premises. *The Quarterly Journal of Experimental Psychology Section A*, *48*(3), 613–643. doi:10.1080/14640749508401408

Stoet, G., & Snyder, L. (2008). Task-Switching in Human and Nonhuman Primates: Understanding Rule Encoding and Control from Behavior to Single Neurons. In S. A. Bunge & J. D. Wallis (Eds.), *Neuroscience of rule-guided behavior* (pp. 227–253). Oxford, NY, US: Oxford University Press.

Stoet, G., & Snyder, L. H. (2004). Single Neurons in Posterior Parietal Cortex of Monkeys Encode Cognitive Set. *Neuron*, *42*(6), 1003–1012. doi:10.1016/j.neuron.2004.06.003

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. doi:10.1037/h0054651

Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain: 3-dimensional proportional system: an approach to cerebral imaging*. Stuttgart, NY, US: Thieme Georg Verlag.

Tanji, J., & Hoshi, E. (2001). Behavioral planning in the prefrontal cortex. *Current Opinion in Neurobiology*, *11*(2), 164–170. doi:10.1016/S0959-4388(00)00192-6

Thoenissen, D., Zilles, K., & Toni, I. (2002). Differential Involvement of Parietal and Precentral Regions in Movement Preparation and Motor Intention. *The Journal of Neuroscience*, *22*(20), 9024–9034.

Toni, I., Schluter, N. D., Josephs, O., Friston, K. J., & Passingham, R. E. (1999). Signal-, Set- and Movement-related Activity in the Human Brain: An Event-related fMRI Study. *Cerebral Cortex*, *9*(1), 35–49. doi:10.1093/cercor/9.1.35

Triantafyllou, C., Hoge, R. D., Krueger, G., Wiggins, C. J., Potthast, A., Wiggins, G. C., & Wald, L. L. (2005). Comparison of physiological noise at 1.5 T, 3 T and 7 T and optimization of fMRI acquisition parameters. *NeuroImage*, *26*(1), 243–250. doi:10.1016/j.neuroimage.2005.01.007

Tulving, E., Kapur, S., Craik, F. I., Moscovitch, M., & Houle, S. (1994). Hemispheric encoding/retrieval asymmetry in episodic memory: positron emission tomography findings. *Proceedings of the National Academy of Sciences*, *91*(6), 2016–2020. doi:10.1073/pnas.91.6.2016

Vapnik, V. (1998). The Support Vector Method of Function Estimation. In J. A. K. Suykens & J. Vandewalle (Eds.), *Nonlinear Modeling: Advanced Black-Box Techniques* (pp. 55–85). Boston, MA, US: Springer US.

Visser, M., Jefferies, E., & Lambon Ralph, M. A. (2010). Semantic Processing in the Anterior Temporal Lobes: A Meta-analysis of the Functional Neuroimaging Literature. *Journal of Cognitive Neuroscience*, *22*(6), 1083–1094. doi:10.1162/jocn.2009.21309

Viswanathan, A., & Freeman, R. D. (2007). Neurometabolic coupling in cerebral cortex reflects synaptic more than spiking activity. *Nature Neuroscience*, *10*(10), 1308–1312. doi:10.1038/nn1977

Wagner, A. D., Bunge, S. A., & Badre, D. (2004). Cognitive Control, Semantic Memory, and Priming: Contributions from Prefrontal Cortex. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (3rd ed., pp. 709–725). Cambridge, MA, US: MIT Press.

Wallis, J. D., Anderson, K. C., & Miller, E. K. (2001). Single neurons in prefrontal cortex encode abstract rules. *Nature*, *411*, 953–956.

Wallis, J. D., & Miller, E. K. (2003). Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. *European Journal of Neuroscience*, *18*(7), 2069–2081. doi:10.1046/j.1460-9568.2003.02922.x

Walton, M. E., Devlin, J. T., & Rushworth, M. F. S. (2004). Interactions between decision making and performance monitoring within prefrontal cortex. *Nature Neuroscience*, *7*(11), 1259–1265. doi:10.1038/nn1339

Weygandt, M., Blecker, C. R., Schäfer, A., Hackmack, K., Haynes, J.-D., Vaitl, D., … Schienle, A. (2012). fMRI pattern recognition in obsessive–compulsive disorder. *NeuroImage*, *60*(2), 1186–1193. doi:10.1016/j.neuroimage.2012.01.064

White, I. M., & Wise, S. P. (1999). Rule-dependent neuronal activity in the prefrontal cortex. *Experimental Brain Research*, *126*(3), 315–335. doi:10.1007/s002210050740

Wise, S. P., Murray, E. A., & Gerfen, C. R. (1996). The Frontal Cortex-Basal Ganglia System in Primates. *Critical Reviews™ in Neurobiology*, *10*(3-4), 317–356. doi:10.1615/CritRevNeurobiol.v10.i3-4.30

Wittgenstein, L. (1921). Logisch-Philosophische Abhandlung. In *Annalen der Naturphilosophie* (Vols. 1-3/4, Vol. XIV, pp. 185–308). Leipzig, DE: Verlag Unesma.

Woolgar, A., Hampshire, A., Thompson, R., & Duncan, J. (2011). Adaptive Coding of Task-Relevant Information in Human Frontoparietal Cortex. *The Journal of Neuroscience*, *31*(41), 14592–14599. doi:10.1523/JNEUROSCI.2616-11.2011

Woolgar, A., Thompson, R., Bor, D., & Duncan, J. (2011). Multi-voxel coding of stimuli, rules, and responses in human frontoparietal cortex. *NeuroImage*, *56*(2), 744–752. doi:10.1016/j.neuroimage.2010.04.035

Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal Autocorrelation in Univariate Linear Modeling of FMRI Data. *NeuroImage*, *14*(6), 1370–1386. doi:10.1006/nimg.2001.0931

Worsley, K. J., Evans, A. C., Marrett, S., & Neelin, P. (1992). A Three-Dimensional Statistical Analysis for CBF Activation Studies in Human Brain. *Journal of Cerebral Blood Flow & Metabolism*, *12*(6), 900–918. doi:10.1038/jcbfm.1992.127

Worsley, K. J., & Friston, K. J. (1995). Analysis of fMRI Time-Series Revisited—Again. *NeuroImage*, *2*(3), 173–181. doi:10.1006/nimg.1995.1023

Yang, Z., Fang, F., & Weng, X. (2012). Recent developments in multivariate pattern analysis for functional MRI. *Neuroscience Bulletin*, *28*(4), 399–408. doi:10.1007/s12264-012-1253-3

Yeung, N., Nystrom, L. E., Aronson, J. A., & Cohen, J. D. (2006). Between-Task Competition and Cognitive Control in Task Switching. *The Journal of Neuroscience*, *26*(5), 1429–1438. doi:10.1523/JNEUROSCI.3109-05.2006

APPENDICES

A. Supplemental materials for study 1

A1. Debriefing questionnaire

Questionnaire

To be administered after the end of the main experiment.

Date _____

Participant ID _____

Age - Sex _____

Education _____

Profession _____

1. Did you notice anything strange or unexpected during the experiment?

2. In particular, did you notice anything strange or unexpected during the presentation of the strings of random characters?

3. Between the two strings of characters sometimes a single capital letter has been hidden. Did you ever notice it? Which letter? How many times?

4. Between the two strings of characters sometimes a geometrical figure has been hidden. Did you ever notice it? Which figure? How many times?

5. Between the two strings of characters sometimes a single digit has been hidden. Did you ever notice it? Which number? How many times?

6. Between the two strings of characters sometimes a symbol (e.g. ! ”) £ ($ % & /) has been hidden. Did you ever notice it? Which symbol? How many times?

Laterality questionnaire:

Writing
Drawing
Throwing
Scissors
Comb
Toothbrush
Knife (without fork)
Spoon
Hammer
Screwdriver
R/10

B. Supplemental materials for study 3

B1. Post-experiment questionnaire

Questionnaire

How did you find the task?

☐ Extremely difficult ☐ Difficult ☐ Fine ☐ Easy ☐ Extremely easy

Which aspect of the task did you find particularly difficult?

_____

_____

How did you solve the task?

_____

_____

Did you use any strategy? Which one?

_____

_____

Did you accept to fail on certain trials (for example, you remembered only the rules about the objects, only the rule about one object or shape and forgot the others, etc.)?

_____

_____

C. Supplemental materials for study 4

C1. Pre-screening questionnaire

Datum: ___/___/2013

Alter: _____

Bildungsstand: _____

Geschlecht: M/W

Haben Sie jemals einen Logikkurs belegt? Ja ☐ Nein ☐

Wenn ja, für wie lange? _____

Wenn ja, was für ein Kurs war das? _____

_____

Im folgenden Fragebogen bitten wir Sie, Sätze wie:
„Es gibt ein rotes Dreieck oder es gibt ein gelbes Quadrat."
zu bewerten.

Dabei gehen Sie *anfangs* davon aus, dass der Satz *wahr* ist. Auf jeden Satz folgt eine Szene mit zwei geometrischen Figuren. Sie müssen entscheiden, ob die Szene mit der Wahrheit des Satzes **vereinbar** ist (also der Satz *wahr* bleibt), oder ob die Szene mit der Wahrheit des Satzes **unvereinbar** ist (also der Satz *falsch* wird). Dann wählen Sie die richtige Antwort aus:

Die Szene ist mit der Wahrheit des Satzes vereinbar ☐
Die Szene ist mit der Wahrheit des Satzes unvereinbar ☐

Bitte fangen Sie jetzt mit dem Fragebogen an! Viel Erfolg!

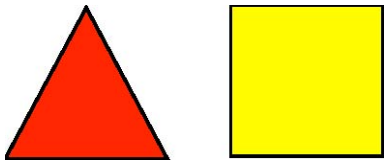
Startzeit: ___:___

Endzeit: ___:___

Nehmen Sie an, der folgende Satz ist wahr:

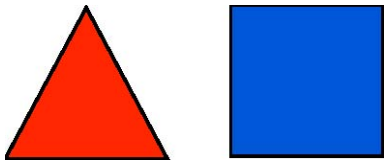
Es gibt ein rotes Dreieck und es gibt ein gelbes Quadrat.

Die Szene:



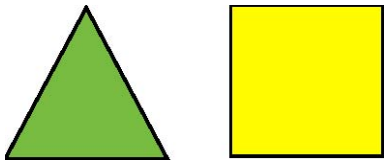
ist mit der Wahrheit des Satzes vereinbar ☐ ist mit der Wahrheit des Satzes unvereinbar ☐

Die Szene:



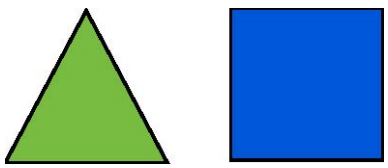
ist mit der Wahrheit des Satzes vereinbar ☐ ist mit der Wahrheit des Satzes unvereinbar ☐

Die Szene:



ist mit der Wahrheit des Satzes vereinbar ☐ ist mit der Wahrheit des Satzes unvereinbar ☐

Die Szene:

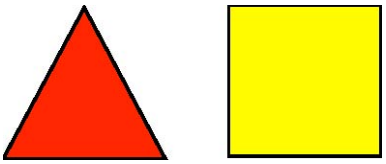


ist mit der Wahrheit des Satzes vereinbar ☐ ist mit der Wahrheit des Satzes unvereinbar ☐

Nehmen Sie an, der folgende Satz ist wahr:

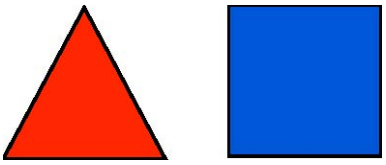
Es gibt ein rotes Dreieck oder es gibt ein gelbes Quadrat.

Die Szene:



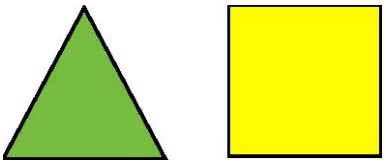
ist mit der Wahrheit des Satzes vereinbar ☐ ist mit der Wahrheit des Satzes unvereinbar ☐

Die Szene:



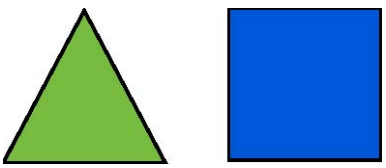
ist mit der Wahrheit des Satzes vereinbar ☐ ist mit der Wahrheit des Satzes unvereinbar ☐

Die Szene:



ist mit der Wahrheit des Satzes vereinbar ☐ ist mit der Wahrheit des Satzes unvereinbar ☐

Die Szene:

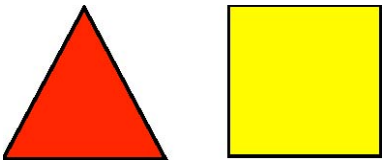


ist mit der Wahrheit des Satzes vereinbar ☐ ist mit der Wahrheit des Satzes unvereinbar ☐

Nehmen Sie an, der folgende Satz ist wahr:

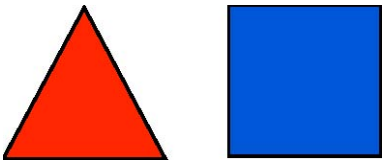
Falls es ein rotes Dreieck gibt, gibt es ein gelbes Quadrat.

Die Szene:



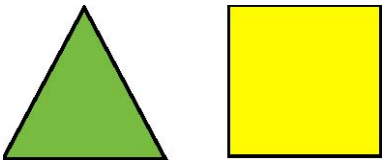
ist mit der Wahrheit des Satzes vereinbar ☐ ist mit der Wahrheit des Satzes unvereinbar ☐

Die Szene:



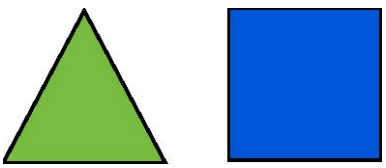
ist mit der Wahrheit des Satzes vereinbar ☐ ist mit der Wahrheit des Satzes unvereinbar ☐

Die Szene:



ist mit der Wahrheit des Satzes vereinbar ☐ ist mit der Wahrheit des Satzes unvereinbar ☐

Die Szene:

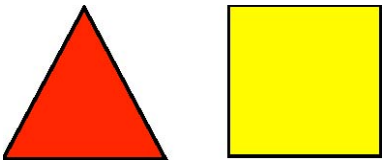


ist mit der Wahrheit des Satzes vereinbar ☐ ist mit der Wahrheit des Satzes unvereinbar ☐

Nehmen Sie an, der folgende Satz ist wahr:

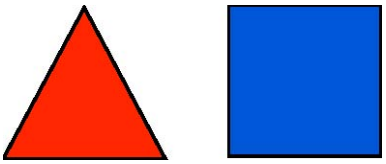
Es gibt kein rotes Dreieck und es gibt ein gelbes Quadrat.

Die Szene:



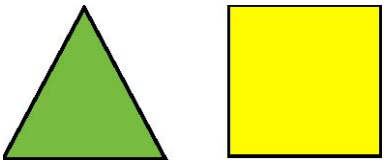
ist mit der Wahrheit des Satzes vereinbar ☐ ist mit der Wahrheit des Satzes unvereinbar ☐

Die Szene:



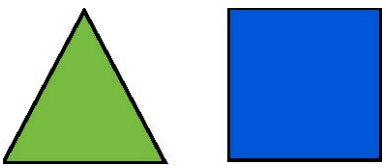
ist mit der Wahrheit des Satzes vereinbar ☐ ist mit der Wahrheit des Satzes unvereinbar ☐

Die Szene:



ist mit der Wahrheit des Satzes vereinbar ☐ ist mit der Wahrheit des Satzes unvereinbar ☐

Die Szene:

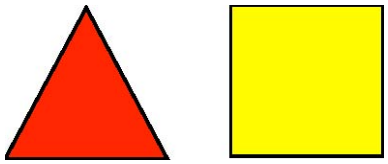


ist mit der Wahrheit des Satzes vereinbar ☐ ist mit der Wahrheit des Satzes unvereinbar ☐

Nehmen Sie an, der folgende Satz ist wahr:

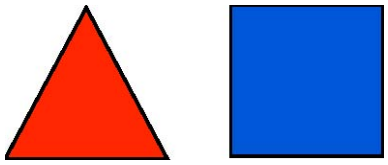
Es gibt kein rotes Dreieck oder es gibt ein gelbes Quadrat.

Die Szene:



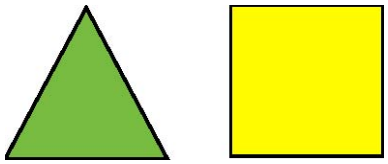
ist mit der Wahrheit des Satzes vereinbar ☐ ist mit der Wahrheit des Satzes unvereinbar ☐

Die Szene:



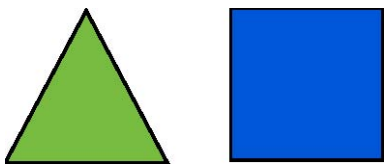
ist mit der Wahrheit des Satzes vereinbar ☐ ist mit der Wahrheit des Satzes unvereinbar ☐

Die Szene:



ist mit der Wahrheit des Satzes vereinbar ☐ ist mit der Wahrheit des Satzes unvereinbar ☐

Die Szene:

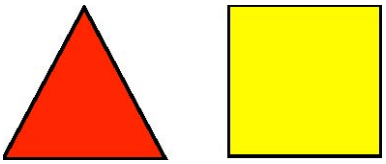


ist mit der Wahrheit des Satzes vereinbar ☐ ist mit der Wahrheit des Satzes unvereinbar ☐

Nehmen Sie an, der folgende Satz ist wahr:

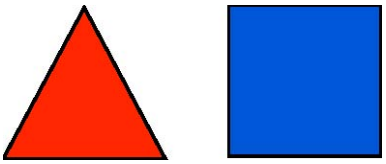
Falls es kein rotes Dreieck gibt, gibt es ein gelbes Quadrat.

Die Szene:



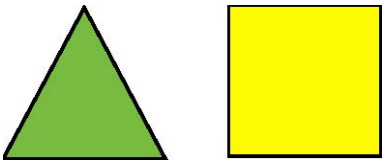
ist mit der Wahrheit des Satzes vereinbar ☐ ist mit der Wahrheit des Satzes unvereinbar ☐

Die Szene:



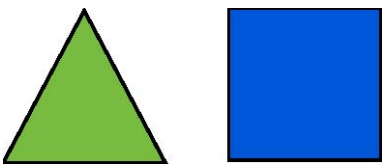
ist mit der Wahrheit des Satzes vereinbar ☐ ist mit der Wahrheit des Satzes unvereinbar ☐

Die Szene:



ist mit der Wahrheit des Satzes vereinbar ☐ ist mit der Wahrheit des Satzes unvereinbar ☐

Die Szene:



ist mit der Wahrheit des Satzes vereinbar ☐ ist mit der Wahrheit des Satzes unvereinbar ☐

C2. Post-experiment questionnaire

Questionnaire

How did you find the task?

☐ Extremely difficult ☐ Difficult ☐ Fine ☐ Easy ☐ Extremely easy

Which aspect of the task did you find particularly difficult?

_____

_____

How did you solve the task?

_____

_____

Did you use any strategy to remember the rules and evaluate the scenes? Which one? Did you change strategy at all?

_____

_____

_____

General Comments

_____

_____

_____