On Rules and Methods: Neural Representations of Complex Rule Sets and Related Methodological Contributions

---

Dissertation
zur Erlangung des akademischen Grades

Doktor rerum naturalium
(Dr. rer. nat.)

eingereicht an der
Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin

von
Kai Görgen, M.Sc.

Präsidentin der Humboldt-Universität zu Berlin
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Lebenswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin
Prof. Dr. Bernhard Grimm

Gutachter/innen
1. Prof. Dr. John-Dylan Haynes
2. Prof. Dr. Benjamin Blankertz
3. Prof. Dr. Felix Blankenburg

Tag der mündlichen Prüfung: 31. Oktober 2019

Cite as: Görgen, K. (2019). *On Rules and Methods: Neural Representations of Complex Rule Sets and Related Methodological Contributions* (Doctoral Dissertation, Humboldt-Universität zu Berlin, Berlin, Germany). https://doi.org/10.18452/20711

Acknowledgements

I am most grateful to all those people and institutions without whom this thesis would probably not have been written.

To John-Dylan Haynes for the opportunity to conduct the work in his lab, for trust and scientific freedom to work on what I considered interesting, for interesting discussions, scientific advice, and numerous creative solutions for funding.

To Benjamin Blankertz and Felix Blankenburg for evaluating this thesis as external reviewers, and to my whole examination committee (that in addition included Martin Rolfs, Anna Kuhlen, and John-Dylan Haynes) for an unforgettable defence.

To my co-authors and friends of the work that made it into this thesis: Carlo Reverberi, Doris Pischedda, Martin Hebart, and Carsten Allefeld.

To my co-authors and friends of work that did not make it to the final version of this thesis: Sven Dähne, Felix Bießmann, Stefan Haufe, Frank Meinecke, and Benjamin Blankertz.

Thank you very, very much for all the years of great collaboration that I really enjoyed a lot.

To various people within and outside the lab, for interesting scientific discussions, especially, but not exclusively, Achim Meyer, Thomas Christophel, Thorsten Kahnt, Radek Cichy, Jakob Heinzle, Anna Kuhlen, Carsten Bogler, Martin Weygandt, David Wisniewski, Robert Deutschländer, Joram Zoch, Dan Birman, Felix Töpfer, and Riccardo Barbieri.

To all master and lab-rotation students who helped at various stages.

To all participants who took part in our experimental studies, for their heroic patience.

To the funding bodies that provided the financial support: BCCN Berlin and the GRK 1589 Sensory Computation in Neural Systems, which both also provided plenty of very helpful, non-financial support; BCAN Berlin; TU Berlin; HU Berlin; Charité Universitätsmedizin Berlin.

To Klaus Obermayer, Vanessa Casagrande, Robert Martin, Margret Franke, Brigitte Krätke-Mann, and Camilla Groiss for their great help to navigate bureaucratic waters and the numerous creative solutions for problems, even such that came on short notice.

To Carsten Allefeld, Doris Pischedda, Corinna Pehrs, Martina Michalikova, and Lieven Schenk, for their helpful critique, comments, and suggestions to the various versions of this thesis.

To other scientific colleagues and good friends for their various kinds of support, especially Peter König, Johannes Höhne, Torsten Betz, and Niklas Wilming.

To those, who enthusiastically followed the work all the way from its start, crossed fingers, shared all the ups and downs, and were eagerly looking forward to its end.

Especially to those, who should be here to see it end, but are not here anymore. Maat et jut. Mer verjesse üch nit.

To everyone else who deserves thanks whom I unintentionally forgot, with deep apologies.

To my parents and brother.

And to Jona. For everything.

Abstract

Where and how does the brain represent complex rule sets? This thesis presents a series of three empirical studies that directly address this question. An additional methodological study investigates the employed analysis method and the experimental design. The empirical studies address the initial question by using multivariate pattern analysis (MVPA) of functional magnetic resonance imaging (fMRI) data from healthy human participants. The methodological study has been inspired by the empirical work. Its impact and application range, however, extend well beyond the empirical studies of this thesis.

The empirical studies (Study 1–3) investigate how the brain represents different features of complex rule sets: Where are cues and rules represented, and are these represented independently? Where are compound rules (i.e. rules consisting of multiple rules) represented, and are these composed from their single rule representations? Where are rules from different hierarchical levels represented, and is there a hierarchy-dependent functional gradient within ventro-lateral prefrontal cortex (VLPFC)? Where is the order of rule-execution represented, and is it represented as a separate higher-level rule? All empirical studies employ information-based functional mapping (using a "searchlight" approach) to localise representations of rule set features brain-wide and spatially unbiased. Rule sets consist of one or multiple stimulus-response mapping rules coupling visual stimuli with specific responses, which were instructed by visual cues during the experiments.

Across all empirical studies, the different rule set features were represented in local, spatially distributed activity patterns. Visual cues were mainly represented in visual occipital areas, independent of the rules they were instructing. The rules, by contrast, were represented in VLPFC throughout all studies. Two of the studies showed additional representations in parietal cortex, while the third study found rule representations in temporal cortex and dorsal striatum. Finally, rule order was represented in dorsal striatum and dorsal premotor cortex. Within VLPFC, anatomical locations of rule representations did not differ significantly for different rule types (single or compound, low or high level hierarchical rules).

One core finding of this thesis is that compound rules were represented compositionally in VLPFC, i.e. their representing activity patterns were similar to those of their constituting single rules, and vice versa. Although single and compound rules were also represented in parietal cortex, representations did not seem to be compositional here, suggesting compositional coding as a specific property of the neural code of prefrontal cortex (PFC). A second core finding is that, in contrast to our initial hypothesis, we did not find any evidence for a functional gradient in VLPFC. This directly contradicts popular theories that postulate such hierarchical-topographical organisation of PFC. The findings of this thesis moreover challenge further assumptions on the neuro-cognitive architecture of rule representations like (1) flexible allocation of resources in PFC, (2) representation and execution of complex rules sets within one fronto-parietal multiple-demand network, or (3) one single region containing all important information about rule sets. Instead, our results support the idea that representations of different components of complex rule sets are distributed across different brain regions.

The methodological study (Study 4) introduces "The Same Analysis Approach (SAA)". SAA allows to detect, avoid, and eliminate confounds and other errors in experimental design and analysis, especially mistakes caused through malicious experiment-specific design-analysis interactions. SAA is relevant for MVPA, but can also be applied in other fields, both within and outside of neuroscience.

Zusammenfassung

Wo und wie werden komplexe Regelsätze im Gehirn repräsentiert? Drei empirische Studien in dieser Doktorarbeit untersuchen diese Frage experimentell. Eine weitere methodische Studie liefert Beiträge zur Weiterentwicklung der genutzten Analysemethode sowie des Experimentaldesigns. Die empirischen Studien nutzen multivariate Musteranalyse (MVPA) funktioneller Magnetresonanzdaten (fMRT) gesunder Probanden. Die Fragestellungen der methodischen Studie wurden durch die empirischen Arbeiten inspiriert. Wirkung und Anwendungsbreite der entwickelten Methode gehen jedoch über die Anwendung in den empirischen Studien dieser Arbeit hinaus.

Die empirischen Studien (Studien 1–3) untersuchen, wie das Gehirn verschiedene Merkmale komplexer Regelsätze repräsentiert: Wo werden Hinweisreize und Regeln repräsentiert, und sind deren Repräsentationen voneinander unabhängig? Wo werden Regeln repräsentiert, die aus mehreren Einzelregeln bestehen, und sind Repräsentationen der zusammengesetzten Regeln eine Kombination der Repräsentationen der Einzelregeln? Wo sind Regeln verschiedener Hierarchieebenen repräsentiert, und gibt es einen hierarchieabhängigen Gradienten im ventrolateralen präfrontalen Kortex (VLPFC)? Wo wird die Reihenfolge der Regelausführung repräsentiert und wird sie als separate Regel höherer Ebene repräsentiert? Alle empirischen Studien verwenden informationsbasiertes funktionales Mapping (mit Hilfe des „Searchlight"-Ansatzes), um Repräsentationen verschiedener Elemente komplexer Regelsätze hirnweit und räumlich unverzerrt zu lokalisieren. Die untersuchten Regelsätze bestehen aus einer oder mehreren Stimulus-Response-Zuordnungsregeln, die visuelle Reize mit spezifischen Reaktionen koppeln, welche während der Experimente durch visuelle Hinweisreize instruiert wurden.

Für alle in den empirischen Studien untersuchten Regelsatzelemente konnten Repräsentationen in Form lokaler, räumlich verteilter Aktivitätsmustern nachgewiesen werden. Die visuellen Hinweisreize wurden hauptsächlich in visuellen Bereichen des Okzipitallappens repräsentiert. Ihre Repräsentationen waren dabei unabhängig von denen der instruierten Regeln. Regelrepräsentation zeigte sich wiederum in allen empirischen Studien im VLPFC. Zwei der Studien fanden zusätzliche Regelrepräsentationen im parietalen Kortex, eine dritte fand diese hingegen im temporalen Kortex und dorsalen Striatum. Repräsentationen der Ausführungsreihenfolge wurde im dorsalen Striatum und dorsalen prämotorischen Kortex nachgewiesen. Im VLFPC unterschieden sich die anatomischen Positionen von Regelrepräsentationen hierbei nicht signifikant für verschiedene Regeltypen (Einzelregeln oder zusammengesetzte, Regeln hierarchisch niedriger oder höherer Stufen).

Ein Kernergebnis dieser Arbeit ist, dass sich im VLPFC Repräsentationen zusammengesetzter Regeln kompositionell (d. h. als Kombination) aus den Repräsentationen ihrer Einzelregeln zusammensetzen. Obwohl sich auch im parietalen Kortex Repräsentationen einzelner und zusammengesetzter Regeln fanden, schienen sich diese nicht kompositionell zusammenzusetzen, was kompositionelle Codierung als eine spezifische Eigenschaft des neuronalen Codes des präfrontalen Kortex (PFC) nahelegt. Ein zweiter zentraler Befund dieser Arbeit ist, dass entgegen unserer ursprünglichen Hypothese unsere Studien keine Hinweise auf einen funktionellen Gradienten in VLPFC lieferten. Dies steht in direktem Widerspruch zu aktuellen einflussreichen Theorien, die eine hierarchisch-topographische Organisation des präfrontalen Kortex (PFC) postulieren. Die Ergebnisse dieser Arbeit stellen darüber hinaus weitere Annahmen zur neurokognitiven Architektur von Regelrepräsentationen in Frage, speziell (1) die Idee der flexiblen Zuweisung von Ressourcen innerhalb des PFC, (2) dass ein einzelnes fronto-parietales „Multiple-Demand"-Netzwerk alle Eigenschaften komplexer Regelsätze repräsentieren und ihre Verarbeitungen durchführen würde, sowie (3) die Existenz einer einzelnen Region, die alle wichtigen Informationen über Regelsätze enthielte. Stattdessen stützen unsere Ergebnisse die Theorie, dass verschiedene Komponenten von komplexen Regelsätzen in verschiedene Gehirnregionen repräsentiert werden.

Komplementierend zu den empirischen Studien ist eine methodische Studie (Studie 4) Teil dieser Arbeit. Diese präsentiert „The Same Analysis Approach (SAA)", ein Ansatz zur Erkennung und Behebung experimentspezifischer Fehler, besonders solcher, welche aus Design–Analyse–Interaktionen entstehen. SAA ist für MVPA relevant, aber auch in weiteren Bereichen innerhalb sowie außerhalb der Neurowissenschaften anwendbar.

Content

Acknowledgements i
Abstract iii
Zusammenfassung v
Table of Figures viii
Abbreviations ix
List of Contributing Publications xi
1 Introduction 1
1.1 Scientific background 2
1.2 Research aims 4
2 Methods 11
2.1 A short history of MVPA 11
2.2 Analysis pipeline of empirical work 14
2.3 Cross-validation and cross-set decoding 15
3 Studies 19
3.1 Study 1: Compositionality of rule representations in human prefrontal cortex 21
3.2 Study 2: Distributed representations of rule identity and rule order in human frontal cortex and striatum 23
3.3 Study 3: Neural representations of hierarchical rule sets: The human control system represents rules irrespective of the hierarchical level they belong to 24
3.4 Study 4: The Same Analysis Approach (SAA) – Practical protection against the pitfalls of novel neuroimaging analysis methods 27
4 General Discussion 31
4.1 General insights and implications from the empirical studies 31
4.2 Compositional versus non-compositional coding 32
4.3 Distributed coding versus a single task set region 35
4.4 A functional gradient within PFC? 38
4.5 Discussion of methodological study 38
4.6 Open issues and future directions 40
5 References 43
Appendix A Selbstständigkeitserklärung ..... A
Appendix B Full Publication Record ..... B
Appendix C Original Publications (Full text) C

Table of Figures

Figure 1.1 – Empirical research questions 8
Figure 1.2 – Methodological contribution and typical MVPA pipeline 8
Figure 2.1 – Compound rules and the cue trick 17
Figure 2.2 – Different validation schemes 17
Figure 3.1 – Basic experimental paradigm 19
Figure 3.2 – Neural representations of different organisation principles of complex rule sets 26

Abbreviations

BOLD	Blood Oxygen Level Dependent
DA	Decoding Accuracy
DLPFC	Dorso-Lateral Prefrontal Cortex
EEG	Electroencephalography
FFA	Fusiform Face Area
fMRI	Functional Magnetic Resonance Imaging
GLM	General Linear Model
HRF	Haemodynamic Response Function
LDA	Linear Discriminant Analysis
MD(N)	Multiple Demand (Network)
MEG	Magnetoencephalography
MVPA	Multivariate Pattern Analysis
PET	Position Emission Tomography
PFC	Prefrontal Cortex
PPA	Parahippocampal Place Area
ROI	Region of Interest
SAA	The Same Analysis Approach
SPM	Statistical Parametric Mapping
SVM	Support Vector Machine
TDT	The Decoding Toolbox
VLPFC	Ventro-Lateral Prefrontal Cortex
X→Y	Rule “if X, then Y”

List of Contributing Publications

This dissertation is based on the following research articles:

Study 1

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

Study 2

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

Study 3

Pischedda, D.*, Görgen, K.*, Haynes, J.-D., & Reverberi, C. (2017). Neural Representations of Hierarchical Rule Sets: The Human Control System Represents Rules Irrespective of the Hierarchical Level to Which They Belong. *The Journal of Neuroscience*, *37*(50), 12281–12296. doi:10.1523/jneurosci.3088-16.2017

Study 4

Görgen, K., Hebart, M. N., Allefeld, C., & Haynes, J.-D. (2018). The Same Analysis Approach: Practical protection against the pitfalls of novel neuroimaging analysis methods. *NeuroImage*, *180*, 19–30. doi:10.1016/j.neuroimage.2017.12.083

*: These authors contributed equally to the manuscript

The original publications can be found in Appendix C. The print version of this thesis contains the full text of all publications. The electronic version only contains references to the publications due to copyright restrictions by the publisher that prohibit to include them in electronic versions (Study 1) or because including them might lead to incompatible copyrights (Studies 2-4) between this part and the publications.

1 Introduction

Modern life would be impossible without the omnipresent application of rules. Road traffic, work, usage of tools from knives to mobile phones, serious communication, or small talk: The concurrent application of a multitude of rules is one of the main pillars of human cognition and decision making (Bunge & Wallis, 2008; Miller & Cohen, 2001; Miller et al., 1960; Monsell, 2003).

Rules play a key role in *cognitive control*, our ability to guide thoughts and behaviour to reach goals in a flexible manner (Allport et al., 1994; Meiran, 2000; Miller et al., 1960; Monsell & Driver, 2000). How important this ability is for daily life becomes especially apparent when it is disturbed, such as in patients with lesion of the frontal cortex who have problems to organise their daily life (e.g. Milner, 1963; Shallice & Burgess, 1991). The ability to use cognitive rules has been repeatedly linked to general intelligence (e.g. Burgess et al., 2011; Duncan, 1993; Duncan et al., 2000; Duncan, 2005), but is still subject of ongoing discussions (e.g. Alvarez & Emory, 2006; Nyhus & Barceló, 2009; cf. Fuster, 2015).

It has been clear since long that humans and other animals represent rules, i.e. keep in memory which rules to perform (Allport et al., 1994; Lashley, 1951; Meiran, 2000; Norman, 1981; Wylie & Allport, 2000). The assumption was that rules would be represented in the brain (Bianchi, 1922; Luria, 1966, 1973; Milner, 1963; Pavlov, 1927). This assumption also played an essential part in neural and behavioural models on cognitive rule use (Fuster, 1989; Miller et al., 1960; Norman & Shallice, 1980, 1986; Stuss & Benson, 1986; see Section 1.1). However, directly demonstrating rule representations in the brain has remained difficult for long.

It took until the turn of the millennium for the first direct demonstration of neural representations of rules in monkeys to arrive (Hoshi et al., 1998, 2000; Wallis et al., 2001; White & Wise, 1999), and until only short before work on this thesis started for demonstrations in humans (Bode & Haynes, 2009; Haynes et al., 2007; Sakai & Passingham, 2003, 2006; Section 1.1). Still, many questions surrounding rule representations remained unexplored, especially questions on representations of sets of multiple rules (Section 1.2).

In this thesis, I present *four studies*: *Three empirical studies* (Reverberi, Görgen et al., 2012a; Reverberi*, Görgen* et al., 2012b; Pischedda*, Görgen* et al., 2017; Sections 3.1–3.3; Figure 1.1) that investigate where and how the brain represents components of complex rule sets[1] and *one methodological study* (Görgen et al., 2018; Section 3.4; Figure 1.2) that advances empirical and statistical methodology. The empirical studies investigate how the brain represents cognitive rules in situations that require the concurrent application of multiple rules at once. The methodological study applies – but is not limited – to "multi-variate pattern analysis" (MVPA; Edelman et al., 1998; Haxby et al., 2001; Haynes & Rees, 2005a, 2006; Kamitani & Tong, 2005; Kriegeskorte et al., 2006; see Section 1.2), the key methodology used in the empirical studies.

This thesis is structured as follows: The remainder of this chapter lays out the scientific background for the studies of the thesis, current at the time when work on this thesis started (Section 1.1; newer work is discussed later, especially in Chapter 4). I then derive the research questions that motivate the presented studies (Section 1.2). Chapter 2 introduces MVPA and further key design decisions for the empirical work. Chapter 3 summarises the studies that are the core of this thesis. Chapter 4 closes the thesis with an overarching discussion. The published articles of the original studies are contained in Appendix C.

* Equal contribution

[1] Throughout this thesis, the term "*rule set*" denotes all rule-relevant components of a task set; see Section 1.1.

1.1 Scientific background

Much research on rule use takes place in the context of *"task set"* (Monsell, 1996, 2003; Sakai, 2008; von Kries, 1895; sometimes called "anticipatory set", e.g. Fuster, 1984, or "preparatory set", e.g. Fuster, 1989, p. 116). The term "task set" refers to the neurocognitive configuration that is necessary to perform a task (Monsell, 2003; Sakai, 2008; von Kries, 1895), a central part of which is to *represent* all required task-related information. One important component to represent is the *rules* that should be applied, which determine which *responses* should be performed to which *stimuli* under which *conditions* (for other task components, see e.g. Gopher et al., 2000).

Rule representations are core parts of theories on cognitive control (Bunge & Wallis, 2008; Monsell & Driver, 2000) and computational models thereof (e.g. Botvinick, 2007; Cooper & Shallice, 2006, 2000; Gilbert & Shallice, 2002; Lebiere & Anderson, 1993; Rumelhart & Norman, 1982). The representations can be either explicit, e.g. as schema (Norman, 1981; Norman & Shallice, 1980, 1986), chunks (Anderson, 1976, 1983, 1993), or managerial knowledge units (Grafman, 1995); or implicit, e.g. as vector space representations distributed across multiple units, e.g. neurons (Duncan, 2001; Miller & Cohen, 2001; Rumelhart & Norman, 1982).

It is clear from behavioural observations that humans (Allport et al., 1994; Lashley, 1951; Meiran, 2000; Norman, 1981; Wylie & Allport, 2000), monkeys (Stoet & Snyder, 2003), and other animals such as pigeons (Honig & Dodd, 1983) or bees (Giurfa et al., 2001) represent rules. These representations are maintained even when no direct cues in the environment indicate which rules should be performed (e.g. Allport et al., 1994; Jersild, 1927; Meiran, 2000; Monsell, 1996; Rogers & Monsell, 1995) and active processes are at work that prepare a forthcoming task (Allport et al., 1994; Gopher et al., 2000; Meiran, 2000; Monsell, 1996; Rogers & Monsell, 1995). Behavioural work e.g. on attention slips in everyday tasks (Norman, 1981; Shallice & Burgess, 1991) furthermore demonstrates that behaviour often does not rely on single rules, but instead on *hierarchically* organised sets of rules, goals, and intentions (Monsell & Driver, 2000; Badre, 2008; Koechlin et al., 2003; Fuster, 1989; Botvinick, 2008; Fuster, 2001; Miller et al., 1960; Norman & Shallice, 1980, 1986; Christoff et al., 2009; Koechlin & Summerfield, 2007; O'Reilly, 2010; Petrides, 2005).

Since long, scientists ascribe rule application and representation to the brain, especially to *prefrontal cortex* (PFC; e.g. Bianchi, 1922; Ferrier, 1876, pp. 287–8; Luria, 1966, 1973; Milner, 1963; Pavlov, 1927; von Kries, 1895). An exception are basic reflexes that reside in the spinal cord (Sherrington, 1906). Theories routinely used this conjecture to set up models of rule use and cognitive control (Duncan, 2001; Fuster, 1989; Grafman, 1995; Miller & Cohen, 2001; Miller et al., 1960; Norman & Shallice, 1980, 1986; Stuss & Benson, 1986). However, direct measurements of neural representations of rules were not possible for a long time, despite intensive research from e.g. behavioural observations in human patients (Burgess et al., 2000; Diamond, 1990; Luria, 1966, 1973; Milner, 1963; Petrides, 1985b, 1990, 1997; Petrides & Milner, 1982; Rowe et al., 2007; Shallice & Burgess, 1991), ablation in monkeys (Buckley et al., 2009; Bussey et al., 2002; Canavan et al., 1989; Diamond & Goldman-Rakic, 1989; Dias et al., 1997; Fuster & Alexander, 1970; Gaffan et al., 2002, 2002; Gaffan & Harrison, 1988, 1989; Parker & Gaffan, 1998; Passingham, 1993; Petrides, 1982, 1985a, 1991b, 1991a, 1996, 2000; Wise et al., 1996), during development (e.g. Goldman et al., 1970; Goldman & Galkin, 1978), electrophysiological recordings in monkeys (Fuster, 1973; Goldman-Rakic, 1987; Kubota & Niki, 1971; Niki & Watanabe, 1979; Passingham, 1993; Quintana et al., 1988; Rao et al., 1997; Thorpe et al., 1983; Yajeya et al., 1988), or non-invasive methods in healthy human subjects using electroencephalography (EEG; Brass et al., 2005; Düzel et al., 1999; Rushworth et al., 2002, 2005), position emission tomography (PET; Frith et al., 1991; Owen et al., 1996; Toni & Passingham, 1999), or functional magnetic resonance imaging (fMRI; e.g. Banich et al., 2000;

Baron et al., 2010; Baron & Osherson, 2011; Bengtsson et al., 2009; Brass et al., 2002, 2005; Brass & Cramon, 2004; Bunge et al., 2002, 2003; Cavina-Pratesi et al., 2006; Cole et al., 2010; Crone et al., 2006, p. 006; D'Esposito et al., 1999a; MacDonald et al., 2000; Postle et al., 1999; Reverberi et al., 2007, 2010; Rowe et al., 2000; Ruge & Wolfensteller, 2010; Schumacher et al., 2003, 2007) .

It took until the turn of the millennium to directly demonstrate rule-representing neurons in monkeys (Asaad et al., 2000; Gail & Andersen, 2006; Genovesio et al., 2005; Hoshi et al., 1998, 2000; Muhammad et al., 2006; Stoet & Snyder, 2004; Wallis et al., 2001; Wallis & Miller, 2003b; White & Wise, 1999). In humans, localisation of rule representation even took until shortly before work on this thesis started. At the time work on this thesis started[2], these representations were suggested to be either PFC-wide functional connectivity states (Rowe et al., 2007; Sakai & Passingham, 2003, 2006) or local activation patterns (Bode & Haynes, 2009; Haynes et al., 2007). Another early study (Li et al., 2007) also used local activation patterns to study rule use, but went a different way. Instead of demonstrating rule-representations directly, the study investigated whether applying different rules would change how much and which task-related information could be extracted from the brain, and found that the ability to retrieve task-related information indeed changes when participants apply different rules.

The key innovation in these studies was the application of novel data analysis methods that enabled demonstration of rule representations in humans (Edelman et al., 1998; Haxby et al., 2001; Cox & Savoy, 2003; Haynes & Rees, 2005a, 2005b, 2006; Kamitani & Tong, 2005; Polyn et al., 2005; Kriegeskorte et al., 2006; Norman et al., 2006; Sakai & Passingham, 2003). This allowed to overcome a major limitation of traditional analysis methods such as the general linear model (GLM[3]; Friston et al., 1991, 1995; Penny et al., 2011; Worsley & Friston, 1995) that could only associate increases or decreases of large-scale activity[4] across participants to experimental conditions.

Specifically, one research group (Rowe et al., 2007; Sakai & Passingham, 2003, 2006) demonstrated rule-selective changes of correlations between the activity of frontopolar cortex (FPC, BA[5] 10, the most anterior region within PFC), dorsolateral PFC (DLPFC, BA 46), and premotor cortex (PM, BA 6), depending on which of three tasks participants were instructed to perform (either phonological, semantic, or visual judgments of visually presented words). Interestingly, differences in correlation were already present during task preparation, i.e. when participants knew which rule to perform but before they applied them. Importantly, conventional fMRI analysis did not show any difference in brain activity during this period, as no region showed an activation increase or decrease between the three rules. They further found that the strength of these correlations predicted the strength of DLPFC and PM activity as well as reaction time during performance. From these, the authors concluded that the large-scale "functional connectivity" brain states would represent individual rules, not only in a simple passive, maintaining manner that would just store rule-related information (e.g. the presented cue), but in a manner that actively prepares rule execution.

[2] As stated above, this introduction contains work until 2010/2011. Newer work is discussed later, especially in Chapter 4.

[3] Note that what neuroscientists call "general linear model (GLM)" is known as "linear model (LM)" in statistics; "GLM" in statistics refers to the "general*ised* linear model" (Nelder & Wedderburn, 1972), a generalisation of the linear model to non-Gaussian data (see e.g. Fahrmeir et al., 2009).

[4] More precisely, increased *activation* refers in fMRI studies typically to an increase of the Blood-Oxygen-Level-Dependent (BOLD) signal. While the exact nature of the relation between neural activity and the BOLD signal is still under debate (see e.g. Boynton, 2011; Heeger et al., 2000), a wealth of evidence suggests good agreement between both signals, typically related via the so-called "hemodynamic response function" (HRF; e.g. Bießmann et al., 2010; Logothetis, 2008; Logothetis et al., 2001; Mukamel et al., 2005; Privman et al., 2007). For the purpose of the studies presented in this thesis, the results do not depend on the exact relation underlying both signals, because the results of the studies only demonstrate presence of information about certain experimental variables in the fMRI signal in parts of the brain and relations between the fMRI response patterns. They do not make claims about the exact physiological or neuronal processes underlying it (for a discussion of the topic, see e.g. Beckett et al., 2012; Freeman et al., 2011; Kamitani & Sawahata, 2010; Kriegeskorte et al., 2010; Op de Beeck, 2010a, 2010b).

[5] Brodmann area (BA): cytoarchitectural defined location from the anatomical brain atlas originally defined by Korbinian Brodmann, see (e.g. Brodmann, 1909; Petrides & Pandya, 1984, 1999, 2002)

A limitation of this functional connectivity approach was that it only allowed to detect representation of qualitatively different types of rules. For example, the rules employed in these studies used stimuli from different sensory domains (phonological, semantic, visual judgments) that are processed by largely different brain areas. It thus seemed unlikely that the same approach would allow for more fine-grained differentiation between tasks from the same type of rule that would rely on the same set of brain regions, such as between different stimulus-response mappings for abstract and concrete words.

Employing a different approach, another group (Bode & Haynes, 2009; Haynes et al., 2007) demonstrated that distinguishing rules of the same type is indeed possible. The authors employed cross-validated decoding, a "multi-variate" (or "multi-voxel") "pattern analysis" method (MVPA; Edelman et al., 1998; Haxby et al., 2001; Cox & Savoy, 2003; Haynes & Rees, 2005a, 2005b, 2006; Kamitani & Tong, 2005; Polyn et al., 2005; Kriegeskorte et al., 2006; Norman et al., 2006) that had been previously employed to distinguish e.g. visual stimuli from brain activity (Haynes & Rees, 2005a; Kamitani & Tong, 2005). In their work, the authors demonstrated that local patterns of fMRI activity provide information about which of two covert intentions, add or subtract, participants had in mind (Haynes et al., 2007), or which of two individual stimulus-response associations, mapping two visual patterns to left and right button presses, participants were instructed to perform (Bode & Haynes, 2009).

One specific limitation of the work by Haynes, Bode and colleagues (Bode & Haynes, 2009; Haynes et al., 2007) was that it did not dissociate rule representations from representations or brain activity related to their instructions (for work in monkey, see e.g. Stoet & Snyder, 2004; White & Wise, 1999), i.e. brain activity due to cue images shown as explicit instruction (Bode & Haynes, 2009) or brain activity underlying the free choice which rule to use (Haynes et al., 2007). Thus, further work was required to dissociate both.

A general limitation of this and all other previous work on rule representation, in both monkeys and humans, was that only representations of simple tasks were investigated that required the application of one single rule in each trial. However, often exactly this ability – to employ multiple rules and combine them in different ways – enables cognitive flexibility (e.g. Badre, 2008; Bunge & Wallis, 2008; Cole et al., 2010; Miller et al., 1960).

Examples for interesting questions that require more complex tasks to investigate include how the brain *combines* representations of multiple rules that should be applied concurrently (e.g. "if there is a tomato, press left; if there is a banana, press right"), how multiple rules are executed in a specific *order* (e.g. "first, check if there is a tomato, and if so, press left; next, check if you see a banana, and if so, press right"), or how *hierarchically structured* rule sets (e.g. Badre, 2008; Koechlin et al., 2003), in which higher-level rules influence the application of lower-level rules, are composed and applied (see Figure 1.1).

1.2 Research aims

This thesis addresses the questions of how compositional, ordered, and hierarchically structured rule sets are represented in the human brain, and how rule representations (or other task-set components) can be dissociated from related information, e.g. the cue used to instruct which rule to use.

Regarding the first question on *compositionality* of rule representations: Behaviourally, it is clear that humans routinely compose new rules from individual elements, or create sets of rules from multiple individual rules, often in combinations they have never performed before (e.g. Cole et al., 2010; Monsell, 1996; Ruge & Wolfensteller, 2010). A parsimonious hypothesis to account for how the brain enables this flexibility would be that it employs a compositional code. For example, the representation of a given rule would be composed by combining representations of its composing parts (e.g. combining "if you see a

tomato" with "then press the left button") or representations of rule sets by combining representations constituting rules (e.g. use rule A and rule B) and potential further representations of how to employ them (e.g. use rule A before rule B) (e.g. Cole et al., 2010). Indeed, compositionality assumptions for brain activity are abundant in cognitive neuroscience (see e.g. Baron et al., 2010; Baron & Osherson, 2011; Friston et al., 1995; Kay et al., 2008; Mitchell et al., 2008; Naselaris et al., 2011; Penny et al., 2011). Most models on rule use conform to this compositionality hypothesis, typically taking individual elements (e.g. model neurons) to represent individual rules, rule components, etc. These would be combined by simple co-activation (e.g. Cooper & Shallice, 2006; Norman & Shallice, 1980, 1986; Rumelhart & Norman, 1982, 1982; but see e.g. Botvinick, 2007; Botvinick & Plaut, 2004; Rigotti et al., 2010).

However, shortly before the work on this thesis started, a series of three studies on neural representations of rule-relevant two-object sequences in monkeys challenged this compositionality hypothesis (Siegel et al., 2009; Warden & Miller, 2007, 2010). In the studies, monkeys had to remember identity and presentation order of two consecutively shown objects to perform one of two tasks. In contrast to the prediction from the compositionality hypothesis that the activity of neurons to a sequence of two objects should be the sum of the activity of its individual components (objects, order, tasks), the observed activity was in most cases a complex combination of those. For example, some neurons fired selectively to the occurrence of a first object, but stopped as soon as a second object was shown. Other neurons only represented a second object if a specific first object was shown, or showed other complex mixtures between presented objects, presentation order, and task (see also Sigala et al., 2008).

If rule representations in the human brain behaved similarly, this would require refinement of theories and models on mechanisms underlying human rule use (such as e.g. Botvinick & Plaut, 2004; Eliasmith et al., 2012; Fusi et al., 2016; Rigotti et al., 2010, 2013). Testing this, i.e. if rule representations in the human brain are compositional or complex mixtures, poses the *first research aim* of this thesis.

Regarding the second question on representation of rule *order*, the three studies mentioned above (Siegel et al., 2009; Warden & Miller, 2007, 2010) also suggest a hypothesis. Especially, the finding that the different factors object identity, object order, and current task jointly influence neural activity in the same neural substrate makes two predictions: First, rules and rule order should be jointly represented in the same areas, at least in lateral PFC (monkey area 46), the area investigated by these studies. Second, a method that allows to distinguish rule identity should also allow to distinguish rule order, and vice versa, again at least in this area. As far as I know, no previous work on processing or representing order of cognitive rules exists, but only work as the above that studied neural processes underlying memorising sequences of other content, such as presentation order of objects, locations, or movements, mainly in monkeys (see e.g. Averbeck et al., 2003, 2006; Barone & Joseph, 1989; Dragoi & Buzsáki, 2006; Mushiake et al., 2006; Ninokura et al., 2003; Petrides & Milner, 1982; Siegel et al., 2009; Warden & Miller, 2007, 2010; Yin, 2009). Thus, testing whether rule order can be detected in the human brain, locating where it is represented, and testing if rule order and rule identity are represented together, poses the *second research aim* of this thesis.

Regarding the third question concerning *hierarchically* organised rule sets (Badre, 2008; Fuster, 1989; Koechlin et al., 2003; Lashley, 1951; Luria, 1966; Miller et al., 1960), behavioural observations clearly demonstrate that many real-world tasks require the application of hierarchically structured rule sets (Botvinick & Bylsma, 2005; Lashley, 1951; Norman, 1981). Examples include mental slips such as forgetting to add tea during tea preparation, and only recognising the error when pouring water instead of tea into a cup (Reason, 1979), or omitting words or letters during typing a text (Lashley, 1951). Both would not be possible if the processes would be strictly linear, i.e. if finishing one action would be required to start the next (for more examples, see e.g. Cooper & Shallice, 2000; Norman, 1981; Reason,

1991; Shaffer, 1978). Computational models of hierarchical behaviour are typically also constructed hierarchically (Botvinick, 2007; Cooper & Shallice, 2006, 2000; Frank & Badre, 2011; Miller et al., 1960; O'Reilly et al., 2002), although Botvinick & Plaut (2004) demonstrated that this is not necessary.

A number of neurocognitive theories map this hierarchical organisation of behaviour to the brain by postulating one (Badre & D'Esposito, 2009; Bunge & Zelazo, 2006; Christoff et al., 2009; Christoff & Gabrieli, 2000; Frank & Badre, 2011; Fuster, 1989, 2001; Kim et al., 2011; Koechlin et al., 2003; Koechlin & Summerfield, 2007; Wise et al., 1996; Wood & Grafman, 2003) or multiple (O'Reilly, 2010; Petrides, 2005) axes along which different hierarchical levels should reside (most within PFC, but see Koechlin & Jubault, 2006 for a proposal of hierarchical organisation in posterior frontal cortex).

However, apart from the general tenet that increasingly complex or abstract rules relate to different regions along some axis, these theories disagree wildly. While some theories propose that discrete areas along the axes perform different cognitive functions (e.g. Petrides, 2005), others postulate cognitive "gradients" along which rules are organised by increasing complexity, typically locating lower to higher rules from posterior to anterior locations (Bunge & Zelazo, 2006; Christoff et al., 2009; Christoff & Gabrieli, 2000; Frank & Badre, 2011; Fuster, 1989, 2001; Kim et al., 2011; Koechlin et al., 2003; Koechlin & Summerfield, 2007; O'Reilly, 2010; O'Reilly et al., 2002; Wise et al., 1996; Wood & Grafman, 2003). Among gradient theories, some postulate that different areas *process* information according to increasingly complex rules (e.g. Christoff et al., 2009; Christoff & Gabrieli, 2000; Kim et al., 2011; Petrides, 2005), while others hold that they also *represent* increasingly complex rules (e.g. Badre, 2008; Badre & D'Esposito, 2007; Koechlin & Summerfield, 2007). Major differences also exist in the proposed topographical layouts for gradients, including differences in which brain regions play a role (see e.g. Badre, 2008).

Gradient theories further disagree on the principle that defines what makes some rules (or processes) more complex than others. Proposals include "cross-temporal contingencies", i.e. joint processing of events of increasing temporal distance (Fuster, 1989, 2001); increasing temporal or contextual distance of representation of such events (e.g. Frank & Badre, 2011; Koechlin & Summerfield, 2007); "relational complexity", according to which higher-level functions operate on output of the next lower-level (Christoff et al., 2009; Christoff & Gabrieli, 2000); increasingly abstract task switching processes (Kim et al., 2011); or increasing distance of information to the final decision during reasoning (e.g. Badre & D'Esposito, 2007). Still others postulate different criteria along different axes (Kouneiher et al., 2009; O'Reilly, 2010).

Even the general question whether gradients within PFC exist at all is still heavily debated. Especially John Duncan (Duncan, 2001, 2006, 2010) dismisses the idea of functional specialisation and hierarchical gradients within PFC altogether. Instead, he postulates that the PFC would constitute a "multiple-demand" network where all regions would work together during all kinds of tasks that require cognitive control, and in which neural resources would be distributed as needed (similar to memory in a conventional computer; for similar ideas, see e.g. Dehaene & Naccache, 2001).

Compared to the amount of attention that the gradient theories received, empirical data to test these are scarce: only few studies directly tested predictions of gradient hypotheses (Badre et al., 2009, 2010; Badre & D'Esposito, 2007; Badre & Frank, 2011; Christoff et al., 2009; Koechlin et al., 2003; Kouneiher et al., 2009) and all tested only predictions from theories that were invented by one or more of their authors. Indeed, the first independent study that explicitly tested two competing theories, published shortly after work on this thesis had started, failed to provide evidenced for both tested theories (Reynolds et al., 2012); instead the authors created a new theory to explain their data.

A specific shortcoming of all previous studies that empirically investigated gradient theories was that no study had measured *neural representations* of rules from different levels of hierarchical rule sets (i.e. to measure which *specific* rules are currently active as opposed to general activation differences for rules from different levels; see Chapters 2 and 3), leaving "representational gradient" theories essentially untested. Thus, testing whether cognitive gradients – especially "representational gradients" – exist, poses the *third research aim* of this thesis.

The three empirical studies that I present in this thesis follow these research aims. They examine how and where the brain represents rules and other components for rule sets that employ the organisation principles introduced above (Figure 1.1):

- Creating compound rules from simple rules (Study 1; Section 3.1)
- Specifying a temporal order for rule execution (Study 2; Section 3.2)
- Organizing rules in cognitive hierarchies (Study 3; Section 3.3)

For this, all studies decompose representations of rule sets using MVPA (Section 1.1 and Chapter 2) to *localise* their components and understand *how* their components work together.

Study 1 (Reverberi, Görgen et al., 2012a; Section 3.1) investigates rule sets composed of either single rules (e.g. "if you see a tomato, press the left button", Figure 1.1, panel a) or double rules that are combinations of two single rules ("if you see a tomato, press the left button; if you see a banana, press the right button", Figure 1.1, panel b), and asks whether the neural representations of double rules are a combination of the representations of their constituting single rules, i.e. if their neural code is compositional. The study also dissociates representations of rules from related information, such as their instructing cues. It confirms rule representations in parietal cortex and ventrolateral PFC. Evidence for compositional coding is only found in the latter.

Study 2 (Reverberi*, Görgen* et al., 2012b; Section 3.2) investigates rule sets that contain double rules as in Study 1, but introduces a specific order for the application (e.g. "first apply the banana rule, then apply the tomato rule", Figure 1.1, panel c). The study localises representations of rule order and especially asks if rules and their order are co-localised, specifically in lateral PFC. The results show rule representation in dorsolateral PFC, slightly below yet overlapping with the results from Study 1. Representations are also found in posterior cortices, but in temporal instead of parietal cortex (Section 4.6.1). Representations of order are localised in striatum and premotor cortex, but – in conflict with the hypothesis derived above – not in lateral PFC.

Study 3 (Pischedda*, Görgen* et al., 2017; Section 3.3) investigates hierarchical rule sets composed of two double rules from different levels of a cognitive hierarchy (e.g., lower-level rule: as in Study 1; higher level-rule: "if you see a star in the background, apply the lower-level rule only to pictures in blue frames; otherwise apply it to all pictures"; Figure 1.1, panel d). The study investigates where rules from different hierarchical levels are represented, and tests whether representation locations depend on their hierarchical level, and if so, whether their locations lie along one of the proposed gradients within PFC (see above). Results again confirm representations of rules in lateral PFC and parietal cortex, again in VLPFC as in Study 1. In conflict with gradient theories, the results show no evidence for differences in representation location of rules from different levels.

* Equal contribution

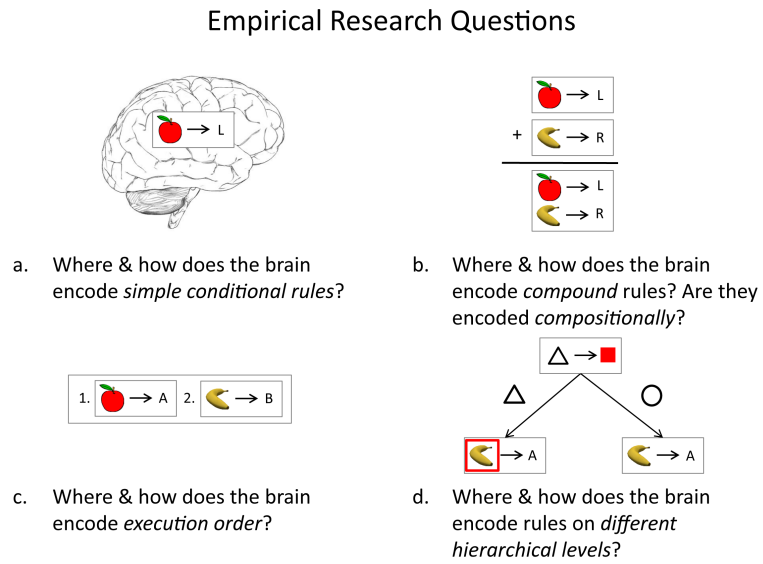

Figure 1.1 – Empirical research questions. Overview of the central research questions of the empirical *Studies 1–3* (Sections 3.1–3.3) of this thesis: a.) Where and how are representations of simple conditional rules located, and can they be dissociated from other rule-related information (*Study 1*); b.) Are representations of compound rules, i.e. rules composed of two simple rules, combinations of the representations of their composing simple rules, and if so, is this the case for all locations that represent rules (*Study 1*); c.) Where and how does the brain represent the order in which multiple rules should be executed (*Study 2*); d.) Where and how does the brain represent rules from different hierarchical levels, and are they represented at different locations (*Study 3*). *Brain image panel a. by Bert Verhelst, GNU 1.2, commons.wikimedia.org/wiki/File:Hersenen.png.*

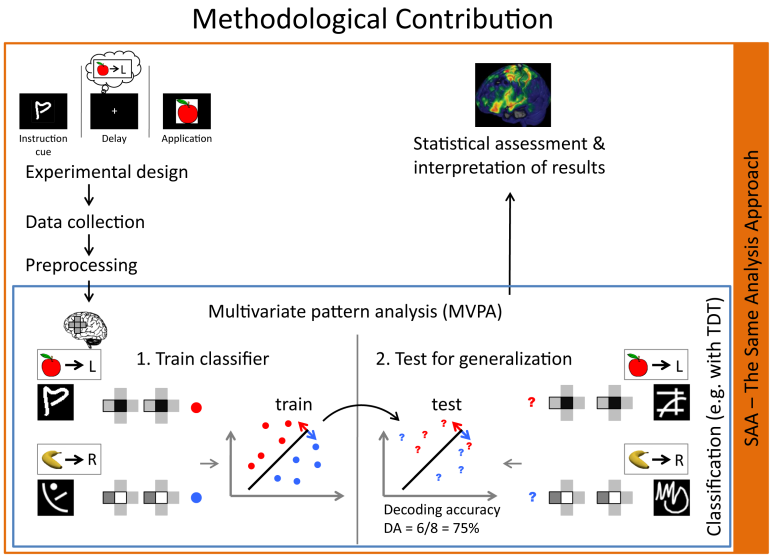

Figure 1.2 – Methodological contribution and typical MVPA pipeline. Scope of the methodological contribution of this thesis projected onto the experimental pipeline employed in the empirical work. *Study 4*, "The Same Analysis Approach" (SAA; Section 3.4) is a practical approach for confound detection, correction, and avoidance, encompassing the full experimental pipeline (consisting of experimental design, data collection, preprocessing, MVPA, statistical assessment, and interpretation of results).
The pipeline itself depicts the general experimental procedure from the empirical studies (see Chapter 2): fMRI data from a delay period (during which the participant is ready to apply a rule that was instructed by a visual cue) serve to train a classifier to distinguish different rules. Successful classification on independent fMRI data demonstrates rule information in the data. The "cue trick" (same rules instructed by different cues for training and test, see Figure 2.1) allows to conclude that information is rule- (and not cue-) specific.

The methodological study (Study 4; Görgen et al., 2018; Section 3.4) complements the empirical work by investigating statistical and empirical methodology surrounding the key analysis method of the experimental studies, MVPA (see Figure 1.2). It has been motivated by a number of unexpected difficulties we had to overcome during data analysis of the empirical projects, and especially by the insight that many of these were the result of two interrelated problems: 1. that conventional design principles (Fisher, 1935) not necessarily safeguard experiments against pitfalls when novel or complex analysis methods are employed; and 2. that conventional control analyses (such as *t*-/*F*-tests, ANOVAs; see e.g. Coolican, 2009) can fail to detect such pitfalls (see Görgen et al., 2018).

In search for a remedy, we developed "The Same Analysis Approach (SAA)" (Study 4; Görgen et al., 2018; Section 3.4), a general framework to systematically detect, avoid, and eliminate confounds and other analysis errors in experimental designs and analysis pipelines. For that, we tackled a very general problem: How can unknown analysis problems, such as unintended confounds or programming errors, be detected? The core insight – also reflected by the name – was that the same analysis that is used for data analysis needs also to be employed to detect and avoid potential confounds.

The paper lays out the principles of SAA and demonstrates its application on two novel, unintuitive problems, a "design-analysis mismatch" between counterbalancing and cross-validation and linear decoding of a non-linear effect. Although originally SAA was developed for application scenarios in the area of MVPA, its scope and application range go well beyond the applications in the empirical work of this thesis, and extend to applications both within and outside neuroscience. In general, SAA joins a growing body of literature that aims towards improving MVPA by providing a better understanding of both its merits and pitfalls (Allefeld, Görgen et al., 2016; Etzel et al., 2013; Etzel & Braver, 2013; Haufe, Meinecke, Görgen et al., 2014; Haynes, 2015; Hebart & Baker, 2018; Mumford et al., 2012, 2015; Noirhomme et al., 2014; Schreiber & Krekelberg, 2013; Todd et al., 2013; Woolgar et al., 2014).[6]

In this first chapter, I have motivated the research conducted in this thesis by laying out its scientific background current at the time when the work on this thesis had started (as stated above, newer work is presented later in the thesis, especially in Chapter 4). In the next chapter (Chapter 2), I introduce MVPA (the key analysis method for the empirical Studies 1–3 and major subject of the methodological Study 4) and important design decisions that we took to decompose complex task sets.

[6] The reader might have noticed that the introduction of the empirical work took much more space in this introduction than that of the methodological work. This is not because the methodological work is less important, but because it only takes very few words to say why methodological work is important: Research on sound methodology is important, because without sound methodology, there is no sound empirical work. Because the scientific background necessary to make the empirical research questions comprehendible to the readers is much more complex, this required much more space.

2 Methods

Cross-validated decoding (Haynes & Rees, 2005a; Kamitani & Tong, 2005), one type of MVPA ("multi-variate" or "multi-voxel pattern analysis"; Edelman et al., 1998; Haxby et al., 2001; Cox & Savoy, 2003; Polyn et al., 2005; Kamitani & Tong, 2005; Haynes & Rees, 2005a, 2005b, 2006; Haynes et al., 2007; Kriegeskorte et al., 2006; Norman et al., 2006), is the key innovation that allows the empirical work in this thesis to investigate neural representations of rules and other task-set components (Studies 1–3; Reverberi, Görgen et al., 2012a; Reverberi*, Görgen* et al. 2012b; Pischedda*, Görgen* et al., 2017; Section 3.1–3.3). MVPA methods are also a major application target for the methodological work in Study 4 (Görgen et al., 2018; Section 3.4).

In this chapter, I first describe the development of MVPA and introduce important methodological tools, such as cross-validation to get unbiased generalisation estimates (Efron & Tibshirani, 1995) or the "searchlight" approach (also "information-based mapping", Kriegeskorte et al., 2006; Haynes et al., 2007) that allows space-resolved analyses (Section 2.1). Next, I explain the main analysis pipeline that we employ in our empirical work (Section 2.2), and "cross-set validation", a specific methodological choice we made that enabled us to separate rule- and cue-related information (Section 2.3).

2.1 A short history of MVPA

In general, the term MVPA subsumes a conglomerate of different analysis methods that only share the major idea to analyse patterns of brain activity. This is in contrast to traditional analysis methods, most notably the general linear model (e.g. Friston et al., 1991, 1995; Penny et al., 2011; Worsley & Friston, 1995), that analyses general activation or deactivation of individual voxels or regions. Apart from multi-variate decoding (Cox & Savoy, 2003; Haxby et al., 2001; Haynes & Rees, 2005b, 2005a; Kamitani & Tong, 2005), other important MVPA methods include shape-space visualisation via multi-dimensional scaling (Edelman et al., 1998), multi-variate encoding and inverse encoding models (Huth et al., 2012; Kay et al., 2008; Kok et al., 2013; Mitchell et al., 2008; Naselaris et al., 2011; Sprague et al., 2014; Thirion et al., 2006), partial least squares (McIntosh et al., 1996), or representational similarity analysis (Kriegeskorte et al., 2008).[7]

A seminal paper by Shimon Edelman and colleagues (Edelman et al., 1998) was presumably the first to employ this idea to investigate patterns of neural responses (see also e.g. Friston et al., 1996 for another early multivariate methodological approach). A major motivation was the recent discovery of neural maps in inferior temporal cortex of monkeys, in which neurons that selectively fired for similar stimulus properties were organized in vertical "columns" (Fujita et al., 1992; Tanaka, 1992, 1996; Tanaka et al., 1991). In their work, Edelman and colleagues wondered whether the clustered activity within these columns would be sufficiently large to be measured with fMRI in humans. More specifically, they hypothesised that similar objects would produce similar fMRI activity patterns, following own (e.g. Edelman, 1995) and others' (e.g. Hinton, 1984) ideas on distributed representations and similarity relations. To test their hypothesis, they recorded fMRI data while images of different objects were shown to human participants. The critical difference to previous work was that – during data analysis – they did not follow the standard procedure to analyse each voxel independently, but instead compared the *similarity of the activity pattern of many voxels* between the different images directly (technically, they

* Equal contribution

[7] The MVPA work discussed here comes from work in humans, which indeed used the term MVPA exclusively for a long time. Some studies in monkey independently developed and performed multivariate analysis techniques as well (Averbeck et al., 2003, 2006; Siegel et al., 2009; Sigala et al., 2008). Because both fields nearly never mentioned work from the other field (for an exception, see Haynes et al., 2007), development remained largely separated.

calculated the pairwise correlation between the activation of selected voxels that was elicited by the presented images). They then used multi-dimensional scaling (Torgerson, 1952) to visually arrange pictures of the shown objects in two dimensional plots, such that the distances between the plotted objects approximated the Euclidean distances between their neural "voxel-space representations". They found that objects that participants judged to be more similar (measured by employing multi-dimensional scaling to behavioural data, see Shepard, 1980) were also placed closer in voxel-space, while dissimilar objects tended to be further apart. Thus, instead of analysing the data *univariately* for each voxel, they used the data of multiple voxels at the same time to perform a *multivariate* analysis.

Despite their result, nearly all fMRI studies continued to use conventional mass-univariate analysis. The next major step in the development of multivariate analysis in neuroimaging came from work by Haxby and colleagues (2001). Instead of creating similarity maps, these authors wondered whether patterns of brain activity would be sufficient to distinguish different categories of objects, i.e. to *decode* which object category a person sees even for objects for which no specific brain region had been identified. Indeed, the authors showed that images of objects from different categories (faces, houses, cats, bottles, scissors, shoes, chairs, and scrambled images) can be distinguished from their elicited brain activity.

Few further studies employing multivariate analysis to fMRI data followed, and the terms "MVPA" (for either "multi-voxel pattern analysis", Norman et al., 2006; or "multi-variate pattern analysis", Polyn et al., 2005) and "decoding" (Kamitani & Tong, 2005) emerged. A next major methodological step to improve decoding performance was to go beyond simple correlations to calculate the similarity between voxel-activity patterns. Instead, Cox and Savoy (2003) introduced two popular methods from machine learning to neuroscience that are routinely still used today: Linear Discriminant Analysis (LDA; Fisher, 1936) and Support Vector Machines (SVMs; Cortes & Vapnik, 1995; see e.g. Müller et al., 2001).[8]

While initial multivariate fMRI studies nearly exclusively investigated object representations, Kamitani and Tong (2005) started to apply MVPA to investigate cognitive states that were not directly stimulus related, but instead related to attended properties of the same physical stimuli. They did so by asking observers to attend to specific features of stimuli (consisting of two overlapping oriented gratings), and showed that these attended features could be decoded from the measured brain activity. Haynes and Rees (2005a) demonstrated that "invisible" stimuli (also oriented gratings, that were not consciously perceived using a backward masking paradigm) elicit brain activity that allowed to successfully decode the identity of these stimuli. A short time later, Haynes and Rees (2005b) demonstrated that another purely perceptual phenomenon, the spontaneously fluctuating perception of bi-stable stimuli, could also be predicted from local patterns of brain activity (for a contemporary review of this and other early MVPA work, see Norman et al., 2006).

A major problem of all MVPA studies until then was how to choose which brain locations should be used during decoding. For this, early MVPA studies typically selected voxels that demonstrated univariate effects in a related contrast, e.g. by taking the voxels that showed maximal differences among responses to stimulus categories (Haxby et al., 2001) or between scrambled and proper images (Edelman et al., 1998; see also Cox & Savoy, 2003; Haynes & Rees, 2005a; Polyn et al., 2005). This bears a high risk of unintentional "double dipping" or circular analyses (Vul et al., 2009; Kriegeskorte et al., 2009; Button, 2019), if the contrasts to define the areas overlap with the classification analysis. An alternative voxel selection criterion was to employ separate data from functional localiser runs for voxel selection (e.g. Cox & Savoy, 2003; Haynes et al., 2007; Kamitani & Tong, 2005). In some cases, voxel selection was further restricted to voxels from predefined regions of interest (ROIs), such as "visual areas" for visual tasks (e.g. Haynes & Rees, 2005a, 2005b; Kamitani & Tong, 2005). This beard the danger to bias the outcome of

[8] http://www.svms.org/history.html lists historical contributions (back to Fisher, 1936) that led to the development of SVMs.

studies, because they could only show that information was present in *the preselected region* but not elsewhere, thereby creating systematic location-specific biases. Finally, classification in a region of interest only informs about information contained within the region as a whole, but does not inform about potential location-specific differences within this region. This did not only fail to harness one of the largest advantages of fMRI above invasive techniques, namely its potential to measure from all locations of the brain at once, but also led to the pitfall of interpreting classifier weights to localise the origin of information (see Haufe, Meinecke, Görgen et al., 2014).

A clever solution to overcome this limitation was the introduction of the "searchlight" approach that allows whole-brain "information-based mapping" (Kriegeskorte et al., 2006) in a spatially unbiased fashion (Haynes et al., 2007). The main idea is to perform MVPA in not only one or few preselected regions, but systematically for *all* locations within the brain. Specifically, the authors suggested to calculate MVPA in small spheres ("searchlights") around each voxel. The result of all searchlight analyses are then collected in a new image, the "information map", that contains for each voxel an estimate of how much information about the condition of interest exists in its local surrounding.

To draw inference about a group of multiple subjects, Kriegeskorte and colleagues (2006) suggested to combine information maps from multiple subjects by first performing searchlight analyses for each subject individually, next to bring these into a common stereotactic space, and to then employ conventional group level analysis methodology (e.g. Friston et al., 1991; Penny et al., 2011). Haynes and colleagues (Haynes et al., 2007), who were the first to employ group level analysis for searchlight decoding, reached the same goal by performing classification already on spatially normalised images.[9]

An initial motivation behind the searchlight approach and decoding in general was that the brain would represent many different kinds of information in different local maps, such as the object map in inferior temporal cortex that motivated the study of Edelman et al. (1998) discussed above, and that fMRI would allow to readout these maps. Although this hypothesis is under heavy discussion ever since (Beckett et al., 2012; Freeman et al., 2011; Kamitani & Sawahata, 2010; Kriegeskorte et al., 2010; Op de Beeck, 2010a, 2010b; see e.g. Haynes, 2015), employing the searchlight approach has turned out to be a very effective method to map information in the brain. The advantage of this approach is that it allows to detect where in the brain information is present without any a-priori anatomical hypothesis. In analogy to the term *mass-univariate* analysis that I used above to distinguish it from a single *multivariate* analysis (e.g. a single ROI analysis), the multivariate searchlight approach could thus be called a *mass-multivariate* analysis. Indeed, recent methodological work (e.g. Weichwald et al., 2015) see content and interpretability of information maps similar to result maps of mass-univariate analysis, for example because interpretation rules for single decoding models (Weichwald et al., 2015) or internal parameters (Haufe, Meinecke, Görgen et al., 2014) do not apply.

An important step during decoding is to quantify information content in data. In MVPA practice, this is often done by estimating *generalisation performance*, i.e. how well a classifier would perform on novel data, which is also common practice in machine learning (see e.g. Bishop, 2006). Other methods to quantify information content include other multivariate distance measures such as the Mahalanobis distance (Mahalanobis, 1936; see e.g. Kriegeskorte et al., 2006) or cross-validated MANOVA (Allefeld & Haynes, 2014). Different methods exist again to estimate generalisation performance, e.g. employing a classifier on a test set of data that was separated from the data used for training (employed e.g. in Cox & Savoy, 2003; Polyn et al., 2005); bootstrap (Efron & Tibshirani, 1995; employed e.g. in Carlson et al., 2003); or cross-validation (Efron, 1983; Efron & Tibshirani, 1995; employed e.g. in Bode & Haynes, 2009;

[9] Today, both approaches are used. Personal experience and reports from colleagues suggest that both procedures yield very similar results, although I am not aware of any systematic study on this topic.

Haynes et al., 2007; Haynes & Rees, 2005b, 2005a; Kamitani & Tong, 2005). When the work on this thesis started, cross-validated decoding had become quasi-standard for MVPA, so we employed it as well (although this might not have optimal power, see e.g. Allefeld & Haynes, 2014; Rosenblatt et al., 2016).

In summary, cross-validated decoding using the searchlight approach provides what is needed to investigate neural representations of complex rule sets. The method allows to *identify and localise* representations of experimental conditions that do not differ in large-scale activations, but are contained in local activation patterns instead. Its potential has been demonstrated in numerous experiments from different areas in cognitive neuropsychology (for reviews, see e.g. Haxby et al., 2014; Tong & Pratte, 2012), such as the work cited above, and has been the key for localisation of representations of simple rules (Bode & Haynes, 2009; Haynes et al., 2007) prior to the work within this thesis.

2.2 Analysis pipeline of empirical work

The empirical studies in this thesis (Studies 1–3; Reverberi, Görgen et al., 2012a; Reverberi*, Görgen* et al. 2012b; Pischedda*, Görgen* et al., 2017; Sections 3.1–3.3) employ cross-validated decoding for data analysis. Captured in single expressions, they employ *stratified leave-one-run-out cross-validated searchlight decoding* and *stratified cross-set MVPA searchlight decoding analysis* on run-wise finite impulse response (FIR) regression coefficient estimates from fMRI data (the terms are explained in the following). The general goal of the analysis is to infer *where* and *how* rule sets are represented in the human brain. The focus of the empirical studies is to decode representations of rule sets during maintenance, i.e. the period during which participants know which rules to apply and are ready to apply them, but before the target stimuli appear to which the rules should be applied.[10]

Before decoding, the raw fMRI data[11] are preprocessed[12] and time-resolved regression coefficient images[13] are created for each condition, run, and participant. These images are then analysed using cross-validated searchlight decoding[14] to calculate information maps that quantify how well local activity patterns distinguish between the experimental conditions of interest, such as between two different rules.

* Equal contribution

[10] Investigating the delay period is different to most conventional and other MVPA (e.g. Woolgar et al., 2011) fMRI studies (but see Bunge et al., 2003). The motivation is that during the delay, participants represent the rule set, but cannot present any other task relevant information (such as target images, etc.). In contrast, when the target screen appears, rule sets can be resolved, and thus only the applying rules or even parts thereof (e.g. the response) may be maintained. Work in monkeys (e.g. Warden & Miller, 2007; Sigala et al., 2008) and a recent study in humans (Hebart et al., 2018) suggests that task set information changes its representational form in different task phases.

[11] All analysis are performed on fMRI data that are recorded from healthy participants (aged 18-30, male and female) while they retrieve, prepare, and apply different rule sets to target images (see Chapter 3).

[12] Preprocessing includes *slice-time correction* to remove differences in acquisition time between different image slices, *realignment* to correct for head movement during scanning, and *high-pass filtering* to correct for slow non-physiological drifts in saturation.

[13] For this, *first level analysis* statistical parametric mapping (SPM; Friston et al., 1995; Worsley & Friston, 1995) is calculated for each participant using full-brain general linear models (GLMs). All models include separate regressors for different rules. Study 1 (Section 3.1) and Study 2 (Section 3.2) also contain separate regressors for the same rules that were instructed by different cues (to conduct the "cue trick"; see details of each experiment for the exact conditions). Different to the standard procedure to use regressors that are convolved with the hemodynamic response function (HRF), we calculated finite impulse response (FIR) models (Henson, 2004) that estimate the BOLD activity of each voxel in consecutive 2s time bins after condition onsets. This procedure creates full brain images ("*beta* images") with voxel-wise correlation coefficients between activity of each voxel and the corresponding regressor, yielding one image per condition, run, FIR bin, and participant.

[14] Searchlight decoding works as follows: Each voxel of the brain serves as centre voxel for one decoding analysis. For each centre voxel v_i, the first-level correlation coefficients within a sphere (radius 4-5 voxels in the studies within this thesis) around v_i are extracted for the classes of interest and serve as data vectors for the classification procedure. Cross-validation (Cox & Savoy, 2003; Efron, 1983; Efron & Tibshirani, 1995) then serves to get an unbiased estimate of how well a classifier that was trained on all recorded data would predict the class of newly recorded data not used for training. For this, the data is repeatedly split into independent training and validation sets according to a given cross-validation scheme (here: "stratified leave-one-run-out" or "stratified cross-set", see Section 2.3). Separate classifiers, in our experiments always linear SVM with fixed regularisation parameter C = 1 (Müller et al., 2001), are trained on each training set to predict which data belong to which class. The classifier is then validated on the left-out validation set. The measure we use to quantified classifier performance is "decoding accuracy" (DA), i.e. the percentage of samples from the validation data for which the classifier predicts the class label correctly. The individual DAs of the different folds are then averaged, yielding the final cross-validation decoding accuracy estimate that is then written at location v_i in a new resulting information map. This procedure is repeated for each participant and often for multiple analyses.

The resulting information maps[15] are then spatially normalized to a standard brain space (Montreal National Institute; MNI) and submitted to a group level statistical test ("second level analysis" in SPM). This then assesses voxel-wise (or ROI-wise) information by testing where decoding accuracy (DA) across participants is significantly higher than expected by chance (i.e. 50% for 2 conditions, 100%/n for n conditions; see Allefeld, Görgen et al., 2016 for interpretation and alternative second-level approaches). Multiple comparison correction[16] is employed to prevent an increase of the false positive rate (alpha level) from performing multiple individual tests. Preprocessing, first-level coefficients estimation, and the final group level statistics after decoding are performed using SPM (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK) and MATLAB (The Mathworks, Inc.). For decoding, Study 1 used custom code, and Studies 2 and 3 employed "TDT – The Decoding Toolbox" (Hebart*, Görgen* et al., 2015).

2.3 Cross-validation and cross-set decoding

A specific feature in our studies is the use of a special cross-validation scheme called "cross-set" decoding (or "cross-classification", e.g. Kaplan et al., 2015; Hebart & Baker, 2018). Cross-validation (Efron, 1983; Efron & Tibshirani, 1995) is a method to get an unbiased generalisation estimate for a model, i.e. an estimate how well a model (e.g. a classifier) that is trained on some given data would perform on new data from the same sample. Typically, cross-validation proceeds by iteratively splitting all data into training and test set according to a certain cross-validation scheme, e.g. by taking all data from each experimental run[17] as test data once ("leave-one-run-out cross-validation"; Figure 2.2, panel a). A model (in decoding: a classifier) is trained on the training set data of each split, and its performance is then validated on the test set data. Averaging all test performances yields the final cross-validated generalisation estimate. Although not initially designed for that purpose, when used for classification, the estimate can be used to test if class-specific information exists in the data sample, by testing if the classification performance is significant above chance level (e.g. Haxby et al., 2001; Haynes & Rees, 2005a; Kriegeskorte et al., 2006).[18]

"Cross-set decoding", in contrast, performs training and testing across different sets of data (Figure 2.2, panel b). These are typically similar in some aspects but differ in others. For example, we used cross-classification to distinguish the same rules that were instructed by different cues ("cue trick" or "double-coding scheme"; Figure 2.1; for application of the double coding scheme in monkey neuropsychology using ANOVA, see e.g. Wallis et al., 2001; Stoet & Snyder, 2004; for psychophysics in human, see e.g.

[15] The exact number of information maps differs between analyses (see original articles of Studies 1–3), depending on the number of conditions that allow the same contrast and whether decoding was performed for each time point (e.g. visual control analysis in Study 1) or across the selected time window (e.g. rule decoding in Study 1).

[16] We employ two different types of multiple comparison corrections that allow different spatial inferences. The first type applies to statistical tests that perform spatially unbiased inference on searchlight accuracy maps and is achieved by employing topological inference (Penny et al., 2011; Taylor & Worsley, 2007; Friston, 2009) through family-wise error correction on cluster level (FWEc) to correct for the large number of inferences (one per voxel) for a specified experiment-wise false positive level (all our studies use the standard level $\alpha = 0.05$). A recent publication (Eklund et al., 2016) demonstrates that (in contrast to the overly simplified general conclusion of that study, see e.g. Brown & Behrmann, 2017) this control is valid when single-voxel thresholds of $p < 0.001$ or below are employed (e.g. Flandin & Friston, 2017; Kessler et al., 2017; Nichols, 2016), which is the case in our studies. The second type applies to analyses that perform more spatially specific inferences using DA averages in ROIs by employing Bonferroni correction to correct for testing multiple ROIs by setting $\alpha = 0.05/n$ (for n ROIs). In our studies, ROIs are either anatomical ROIs from standard anatomical atlases, functional ROIs from independent analyses, or functional-anatomical ROIs retrieved using leave-one-participant out cross-validation to ensure non-circularity of statistical inference (Kriegeskorte et al., 2009; Vul et al., 2009). In general, ROI analyses are considered more sensitive than full-brain searchlight analyses due to less correction for multiple comparison (the number of tested ROIs is typically much smaller than the number of voxels that need to be corrected in full-brain searchlight analyses). This however comes to the cost that ROIs only allow inference at the locations where they are, and can thus create a bias for the investigated spatial locations. Reasons for employing ROI analyses especially include cases in which prior evidence exists that certain regions are specifically involved in the process under investigation. Combining both, searchlight analysis and ROI analysis, allows for both, spatially unbiased inference as well as employing priori knowledge for more sensitive, spatially specific inference.

* Equal contribution

[17] fMRI experiments are often divided in multiple runs, during which participants perform the experimental task on multiple trials (the empirical Studies 1–3 have 6 runs of ca. 10 mins each). Between runs, fMRI recording is switched off to have short breaks.

[18] Alternatives exist to all analysis components, cross-validation, classification, and second level inference (see e.g. Allefeld & Haynes, 2014; Allefeld, Görgen et al., 2016; Hebart & Baker, 2018; Rosenblatt et al., 2016), but as these are the most common choice, we employ them as well.

Kleinsorge, 2012). In this case, significant generalisation performance allows to conclude that the information about the similar aspect is present in the data (e.g. rule-related influences) and generalises across sets, and to exclude that the different aspect (e.g. cue-related influences) confounded the inference. For example, if the goal is to separate rule-related (similar aspect) and cue-related (different aspect) representations (Figure 2.2, panels b, c), successful rule classification trained and tests on data that contained different cue images allows to infer that the data contain rule-specific information that is independent of information about the cues. Early work that employed cross-classification by others includes testing for generalisation between visual stimulation and perception (Kamitani & Tong, 2005) or imagination (Stokes et al., 2009), testing for temporal reappearance (Polyn et al., 2005), or common coding schemes between anticipating and receiving a reward (Kahnt et al., 2010; for more application examples, see Kaplan et al., 2015).

Because in our experiments two visually unrelated cues were used to instruct each rule sets (for example, Cue1 and Cue2 to instruct Rule1, Cue2 and Cue3 to instruct Rule2, see Figure 2.2), the data of two different rule sets Rule1 and Rule2 can be separated into four different splits (Figure 2.2, panel b; these are: Split 1: train Rule1 using Cue1 vs. Rule2 using cue Cue3, test Rule1 using Cue2 vs. Rule2 using Cue4, split 2: vice versa; split 3: train Rule1 using Cue1 vs. Rule2 using Cue4, test Rule1 using Cue2 vs. Rule2 using Cue3, split 4: vice versa).

For the empirical Studies 1 and 2 (Sections 3.1, 3.2), we employed set-wise cross-validation in which classifiers were trained and validated on all data from all runs (Figure 2.2, panel b). Based on insights we gained while working on "The Same Analysis Approach" (Section 3.4) we now highly recommend to combine set-wise with run-wise cross-validation in future studies (Figure 2.2, panel c). Although the combination is computationally slightly more expensive because it needs additional cross-validation steps (number of sets x runs compared to only number of sets), it can prevent unexpected effects that are caused by temporal proximity between measurements and seems to create more stable estimates.

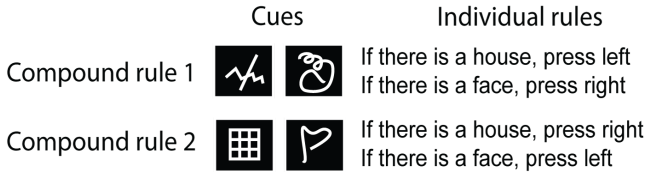

Figure 2.1 – Compound rules and the cue trick. Two elementary features of the empirical studies in this thesis (Studies 1–3) are *1.* decoding between symmetrical compound rules that contain the same building elements and only differ in the individual rules that connect these (e.g. the compound rule 1 and 2 in the figure contain the same parts [house, face, left, right] but connect these in different ways), and *2.* using the "cue trick" (that requires two different symbols to instruct the same rule in different trials) to disambiguate neural representations of a specific cue image from representations of the instructed cue. This allows cross-set decoding (Section 2.3; Figure 2.2) during which a classifier is trained to distinguish rules from data instructed with one set of cues, after which its performance is assessed on data from the other set of cues. Successful classification demonstrates presence of rule-specific neural patterns.

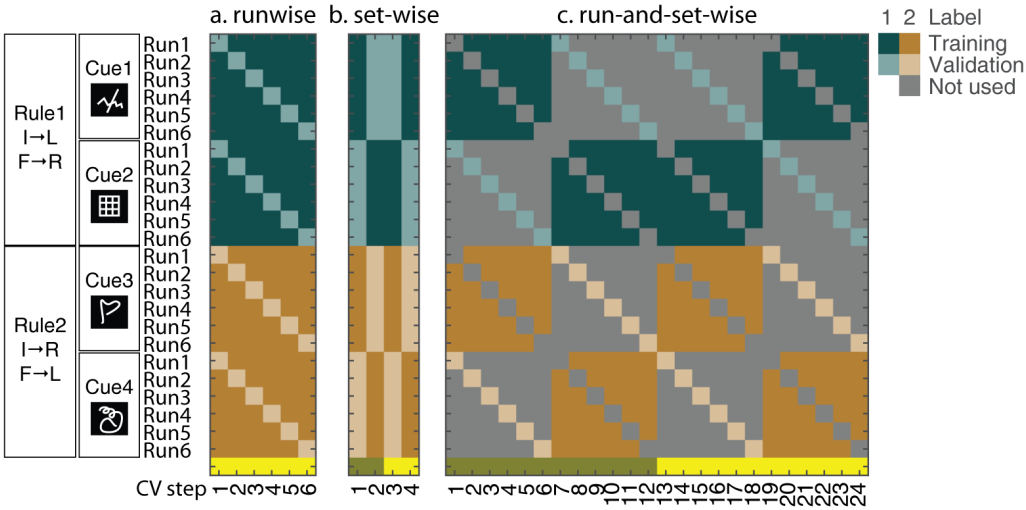

Figure 2.2 – Different validation schemes. a. Standard run-wise ("leave-one-run-out") cross-validation and b. set-wise ("cross-set") validation have been employed in the empirical studies of this work. c. Run-and-set wise validation is recommended for future work (see explanation in text).

3 Studies

In this chapter, I provide an overview about the four studies that are at the core of this thesis (Appendix C). The empirical studies (Studies 1–3; Reverberi, Görgen et al., 2012a; Reverberi*, Görgen* et al. 2012b; Pischedda*, Görgen* et al., 2017; Sections 3.1–3.3) investigate properties of neural coding that underlie how the brain represents rules and rule sets. Specifically, they investigate which *compositionality principles* the brain applies in these representations, i.e. if the brain composes representation of complex rule sets of representations of their constituting parts. For that, all studies employed multivariate pattern analysis (MVPA) of functional magnetic resonance imaging (fMRI) data (Figure 1.2) using searchlight decoding (Kriegeskorte et al., 2006; Haynes et al., 2007) to allow spatially unbiased full-brain space-resolved analysis of local activation patterns (see Chapter 2). The theoretical study (Study 4; Görgen et al., 2018; Section 3.4) introduces "The Same Analysis Approach", a pragmatic approach to systematically detect, avoid, and eliminate confounds and other analysis errors in studies employing MVPA or other complex analysis methods. The theoretical study drew inspiration from the empirical studies in this thesis, which in turn benefited from the developed methods and insights.

All empirical studies in this thesis employ a task-cueing paradigm (e.g. Sudevan & Taylor, 1987; Meiran, 1996; for reviews and other paradigms see e.g. Monsell, 2003; Sakai, 2008) depicted in Figure 3.1. Before the experiments, participants learn to associate visual cues (arbitrary symbols) to different rules. To dissociate effects of cues and rules during data analysis, two different cues were learned for each rule (see "cue trick", Section 3.1 and Figure 2.1). Participants then trained to perform the experimental task. During the experiment, participants repeatedly performed the following procedure: Each experimental trial started with an instruction screen that showed one or two cues. Participants had to recall the associated rule(s) from these cues that formed the rule set for this trial. This rule set then had to be maintained in memory during a delay phase of several seconds. Finally, a target screen appeared, showing one or more images to which the participants had to apply the rules as fast and accurate as possible.

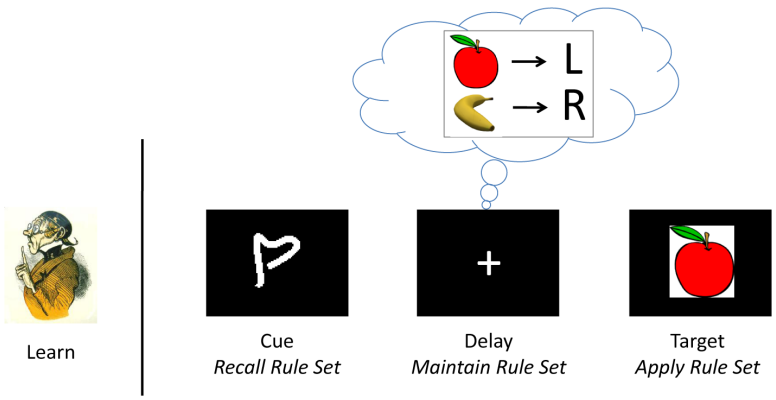

Figure 3.1 – Basic experimental paradigm. Before the experiment, participants learn to associate visual cues to rules and train to conduct the experiment. During the experiment, participants repeatedly perform the same task in successive trials: First, one or more previously learned cues appear and participants have to recall the associated rule(s) (here: "If you see an apple, press the left button; if you see a banana, press the right button"). This rule set needs to be maintained in memory during a delay period of several seconds (the period for which we analysed data, see main text). Finally, a target image occurs (here: apple) to which the participants need to apply the rule set and perform the resulting action (here: left button press). *Source left image: "Lehrer Lämpel" by Wilhelm Busch, GNU-FDL.*

* Equal contribution

The difference between the three empirical studies lies in the rule sets that the participants had to apply. The critical analysis in all experiments is to test whether and where the brain contains information about the identity of rule set components (e.g. which specific rule a participant had in mind) during the delay period[19], i.e. during the time in which participants knew which rule set to apply, but not to which target stimuli, because the target screen had not appeared yet.

The most prominent difference between conventional mass-univariate (see Chapter 2) and MVPA studies on rules is the choice of the experimental conditions. Conventional studies typically use different *types* of tasks to detect differences in *activation*. For example, studies on rules use tasks that are more or less difficult (e.g. Bunge et al., 2003), that involve more or less rules (e.g. Brass & Cramon, 2004), that use bivalent or univalent responses (e.g. Crone et al., 2006), etc. (see e.g. Bunge, 2004). In contrast, most MVPA studies on rules, including those I present in this thesis, use the same type of task, and compare different rules or other features of rule sets within that task (e.g. Haynes et al., 2007; Momennejad & Haynes, 2012, 2013; Soon et al., 2008; Wisniewski et al., 2014). An example is the contrast between two single rules such as "If you see a tomato, press the left button" and "If you see a banana, press the right button" (Study 1; Reverberi, Görgen et al., 2012a; 3.1). This will probably not result in differences in overall activation in a region, but specific *patterns* of individual voxel activations might still differentiate the identity of the rules.[20] Thus, our studies do not use experimental conditions of different *types* to detect differences in *activation;* instead, they differ in specific *content* to detect content-specific *information* (see e.g. Haynes, 2015; Hebart & Baker, 2018).

[19] A difference to most conventional (but see Bunge et al., 2003) and MVPA (e.g. Woolgar et al., 2011) studies; see Footnote 10.

[20] We employed a number of additional experimental measures to make sure that participants indeed represented the rule set during the delay, that the representation did stay constant during the course of the fMRI session, and that data from different participants was compatible. **Extensive training:** To learn cue meaning by heart and to overlearn rule application for stable representations, participants underwent extensive training before participating in the experiment during which they first learned the rules, and then trained to apply them until their responses were both fast and precise. The training typically took part on two of the three days prior to data collection. Data was only recorded if participants successfully performed the training. This was done for two main reasons: First, to ensure that participants learned the rules, and second that they indeed prepared to apply the rule sets in the delay period (i.e. that the rule set was implemented for application, and not just e.g. mentally rehearsed). The delay period in the training was also very short which directly required that participants had the rules present in mind before application. This should avoid that participants change their mental strategies during the time of fMRI data recording (e.g. Cole et al., 2010; Gigerenzer & Gaissmaier, 2011; Gläscher et al., 2010). **Catch trials:** To avoid simplified strategies, we furthermore employed catch trials to avoid that participants changed their strategy during fMRI recording in Studies 2 and 3 (Sections 3.2, 3.3). Catch trials are critical for the experiment because they force and allow verifying that participants indeed perform the task as they should, even though the neural data is typically not suitable for analysis. For example, we randomly delivered *short catch trials* in which the delay period was only maximally 1s long, instead of 3-5s in experimental trials. This allowed to check that participants had the rules ready for application (i.e. represented them) during the whole delay period. We also included *unbalanced catch trials* that contained rule combinations that did not contain balanced rules, i.e. both conditions and both responses (e.g. "instrument→A and instrument→B"). This was to make sure that the participants really maintained both rules, and not represented rule sets in a reduced form by remembering one of the rules (e.g. "instrument→A") and then retrieving the other rule during target execution. For details, see the method sections of the empirical Studies 1–3. **Extensive behavioural control analysis:** To verify that our experimental design decision had the intended effects, we conducted extensive behavioural control analysis on reaction times and errors rates, both during pilot experiments and on the behavioural data collected during scanning. We initially followed the standard MVPA practice to employ statistical tests typically used in psychology (e.g. the t-test; Student, 1908). However, because we had doubts that these would detect potential problems caused by the loss of signs at the decoding level, we performed additional analyses that employed the same analysis methods as the main analysis from Study 2 (Section 3.2) onward. Specifically, we correlated across participants *1)* decoding accuracies from the result cluster ROIs and *2)* absolute differences in reaction times and errors rates. We found no evidence for correlations between decoding performance and the tested behavioural measures. The obtained results speak against criticism of our studies that had been voiced later (Todd et al., 2013), which hypothesised that differences in reaction times – and not differences in rule representation – would explain our results. A generalisation of this together with further insights from this and other data analyses projects have led to the methodological Study 4 ("The Same Analysis Approach"; Section 3.4) of this thesis. **Standardized training procedures and homogenous group of participants:** Two further measures were taken with the goal to make strategies and other mental processes of the participants, as well as their fMRI BOLD signals, as similar as possible. First, we recruited participants of similar age, background, and education (mostly university students, age 18–30 years). While this makes the sample not representative to the general population, it ensures that the study is comparable to most other studies (e.g. Henrich et al., 2010). One reason to have participants of roughly the same age is that age has been reported to alter the hemodynamic response function, with younger participants showing stronger BOLD signals (e.g. D'Esposito et al., 2003; Garrett et al., 2017; Hesselmann et al., 2001; Ross et al., 1997). A homogenous sample thus reduces the variance of the fMRI data between participants, and thus might increase statistical power (e.g. D'Esposito et al., 1999b, 2003). As second measure, we used highly automated computer-based training procedures to avoid experimenter effects (Rosenthal, 1963, 1966, 2009) as good as possible. The training was done in different steps that were typically performed until participants reached certain criteria, and instructions were typically presented on the screen. Of course, participants could always ask the experimenter if they had any question.

3.1 Study 1: Compositionality of rule representations in human prefrontal cortex

Study 1 of this thesis (Reverberi, Görgen et al., 2012a) was designed to investigate the first research aim introduced in Chapter 1, to test whether neural representations of a "compound rule" (a rule that consists of two individual rules, e.g. "if you see a house, press left; if you see a face, press right", short: "house→left; face→right") is composed of the neural patterns of the individual rules ("house→left" and "face→right"; e.g. Cole et al., 2010; Ruge & Wolfensteller, 2010) or not (Warden & Miller, 2007, 2010; Sigala et al., 2008). A second aim of the study was to dissociate neural representations of the instructing cues (the visual symbols that instruct participants which rules to use) and the instructed rules (e.g. Wallis et al., 2001; Wallis & Miller, 2003b for work in monkeys).[21]

To dissociate rules and cues, we first identified *rule-independent representations of cue images*. For this, we split the data for each rule by the cue that was used to instruct the rule (each rule was instructed by two rules; half of the trials were instruct by the one cue, the other half by the other cue; see Figure 2.1). We then employed standard cross-validated searchlight decoding (Section 2.2) to test where the brain contained information about which of the two cues was used to instruct the rule.

Next, we identified *cue-independent representations of rules* by employing the "cue trick": Cross-classification (Section 2.3) was used to distinguish between pairs of rules (e.g. "house→left" vs. "face→right") by training a classifier on data from trials that used one set of visual cues to instruct the two rules, and testing its prediction performance on data that used another set of visual cues to instruct the same rules (Figure 2.1). Because the only commonality between the trials used for training and testing were the instructed rules, but not the cue images that were used as instruction, successful classification can only be achieved when neural representations are rule- but not cue-specific.

To exclude that successful classification can be achieved using the triggering condition ("house" vs. "face") or the consequence ("left" vs. right") alone, we performed the main cross-classification analysis on the two symmetrical (or "orthogonal", Sakai, 2008) compound rules, i.e. rules composed of two individual rules that together contain both triggering conditions ("left" and "right") and both consequences ("house" and "face") and only differ in the mapping between both (i.e. rule 1: "house→left; face→right" vs. rule 2: "house→right; face→left"). Because both compound rules contained the same triggering conditions and consequence, the only difference between both are the rules that map antecedents to consequences, and thus, successful classification should only be possible if the neural activity depends on this mapping rule.[22]

Finally, we tested the *compositional hypothesis* by employing cross-classification to decode compound rules from simple rules (and vice versa). For that, data from e.g. two individual rules that were part of different double rules were used to train a classifier (e.g. individual rule 1: "house→left" vs. individual rule 2: "house→right"). The classifier then classified data from the two compound rules (e.g. compound rule 1:

[21] Depending on a variety of factors, psychological and neural processes can differ substantially even for very similar (sometimes even equal) rules (e.g. Gigerenzer & Gaissmaier, 2011; Gläscher et al., 2010). The basic type of rule that we use in all experiments of this thesis are *stimulus-response mapping rules* (S-R rules) of the form "if you see an item of category X, then press Y", where Y is either "press the left/right button" (Study 1 and 3; Sections 3.1, 3.3) or "the side where the letter A/B occurs" (Study 2; Section 3.2). We typically combine two of these basic rules into one "compound rule" to ensure that observed results are indeed caused by the full mapping between stimuli and responses, and not stimuli or responses separately (see Section 3.1). In Study 3, we additionally employed higher-level modifier rules that had the same logical form as the lower-level S-R rules, which specified to which images the lower-level rules had to be applied (see Section 3.3).

[22] An additional caveat we had to exclude was that participants would use mental shortcuts to present double rules, such as to only remember the first part ("house→right, else other side" vs. "house→left, else other side"). In this case, representations of "right" and "left" would have been sufficient to distinguish the double rules. The following measures were taken to prevent this: First, the allowed response time was considerably short so that participants had no time to think about rules, but had to prepare responses in advance. Second, images of a third category were shown, which required no response. Further evidence that participants presented the full double rules and did not use mental shortcuts comes from the results of the next set of analyses, in which we demonstrated that compound rules can be decoded using the constituting individual rules and vice versa. This would not have been possible had participants used mental shortcuts.

"house→left; face→right" vs. compound rule 2: "house→right; face→left"). Were the neural representations of the compound rules composed of the representations of the composing individual single rules, (as e.g. hypothesised by Cole et al., 2010; Ruge & Wolfensteller, 2010), the classifier should classify "compound rule 1" as "single rule 1" (because both contain "house→left") and "compound rule 2" as "single rule 2" (because both contain "house→right"). If instead a non-compositional, independent neural code would be used to create the compound rule representations (as in Warden & Miller, 2007, 2010; Sigala et al., 2008), classification performance should be at chance level.

We found a) *rule-independent representations of the instructing cue* in visual and parietal cortex (left inferior and right middle occipital gyrus, BA 18 and BA 19; left superior parietal lobe, BA 7), b) *cue-independent representations of rules* in parietal cortex (left BA 7/40) and right ventro-lateral prefrontal cortex (VLPFC; BA 46, 47), and c) *evidence for compositional coding* of compound rules only in right VLPFC (BA 46, 47), but not in parietal cortex (Reverberi, Görgen et al., 2012a, their Figure 3; this thesis, Figure 3.2, panel 2).

The results thus confirmed previous hypotheses that both parietal cortex and VLPFC would represent specific rules (e.g. Bode & Haynes, 2009), but go beyond that by demonstrating that these representations indeed depend on the rule to be used, not the visual cue that was used as instruction. The fact that compositional coding of rule representations was only found in VLPFC but not parietal cortex, whereas conversely representations of visual cues were only found in parietal cortex but not VLPFC, led us to speculate that parietal cortex and VLPFC might fulfil different *roles* in rule processing: Parietal cortex – located between visual cortex and PFC – might "convert" the visual cue information into its associated rule information, which would then be transmitted to VLPFC to execute the rules.

Two studies (Cole et al., 2011; Woolgar et al., 2011) have independently and simultaneously investigated questions similar to those investigated here (Reverberi, Görgen et al., 2012a). Like us, Cole and colleagues (2011) used searchlight decoding to investigate compositionality of rule representations. However, they focused on a different level than we did. Rather than investigating compositionality of representations that results from combining *multiple* individual rules as we did, they investigated compositionality of the representations that results from using different *components* of single individual rules. For that, they decoded components of rules that were made up of three components (component 1: relevant dimension, e.g. "is it sweet?"; component 2: decision rule, e.g. "are the images the same"; component 3: response button, e.g. "left index finger"). They then showed that these components can be decoded individually, even for combinations that participants never saw before. Like us, they found compositionality of the decision rule component only in PFC, with a focus on right VLPFC. A follow-up study (Cole et al., 2016) suggested that these rule representations are indeed relevant for behaviour by demonstrating that the strength of the patterns that represent rules (measured as decoding accuracy of single trials) discriminates successful from erroneous application. One critical distinction between their and our study is the investigated task phase. [23] While they decode representations during rule application, we decode representations during rule maintenance before application.

Woolgar and colleagues (2011) also investigated compositionality of different task-critical components, localising representations of stimuli, rules, and responses. Especially, they also dissociated representations of rules and cues. For that, they even employed the same approach that we also used: searchlight decoding and the "cue trick". Using a double-coding scheme and cross-classification (see Section 3.1) they identify representations of rules that were cue-independent, and using cross-validated decoding between cues of the same rules allowed to identify rule-independent cue representations. Their

[23] This difference might be critical, as work in monkeys on the relation of representations of objects between different task phases (Sigala et al., 2008; Warden & Miller, 2007, 2010) found that these representations change their format during different task phases. Recent work in humans (Hebart et al., 2018) supports this view, calling for further investigation of this issue (see also Section 4.4).

results largely agree with ours: They also found rule representations in VLPFC and parietal cortex, and representations of their cues in visual cortex. Major differences between their study and ours lie in the analysed task phase (they analysed during the application phase, as Cole et al., 2011, above; we analysed the maintenance phase before that; see Footnote 23), the employed rules (they used spatial mappings; we used categorisation), and that they did not investigate compositionality of multiple rules.

3.2 Study 2: Distributed representations of rule identity and rule order in human frontal cortex and striatum

In Study 2 (Reverberi*, Görgen* et al., 2012b), we investigated the second research aim of this thesis (see Section 1.2), i.e. if the "compositionality principle" hypothesised in Study 1 (Reverberi, Görgen et al., 2012a; 3.1) would also hold for more complex rule sets. For that, we extended the compound rule sets from Study 1 by the component "rule order". The general experimental paradigm and analysis methodology were the same as in Study 1. Participants were again instructed to perform two individual rules A and B, but in addition were also instructed to perform either A before B, or B before A.

Thus, the specific research questions in Study 2 of this thesis were:

1. Where and how does the brain represent execution order of rules?
2. Is the neural code of rule sets that contain execution order compositional?

To answer these questions, we tested where rule *identity* and rule *order* could be decoded in the brain. Because the rules were very similar to the ones used in Study 1, we also tested whether the results from Study 1 would be replicated.

Specifically, we tested predictions of two recent tenets of theories on the neurocognitive architecture underlying rule processing: First, we tested the hypothesis that the fronto-parietal network, and specifically VLPFC, would contain all necessary information for a specific task (Duncan, 2001; Miller & Cohen, 2001). If this was the case, this network should also contain information on rule order. Second, we tested a group of "gradient theories" of PFC organisation (Badre, 2008; Badre & D'Esposito, 2009; Koechlin et al., 2003), according to which rules at different levels of a cognitive hierarchy reside along a rostro-caudal gradient within PFC, such that progressively higher-level rules would be represented increasingly more anterior and influence lower-level rules more posterior. Because an order for rules is a particular case of a higher-level rule ("First perform rule A, then rule B") with regard to the lower-level rules that are ordered (rules A and B), we hypothesised to find representations of rule order residing higher up along such a gradient. A couple of similar theories exist that mainly differ in the exact location of the gradient and the cognitive features that define the hierarchical level of rules.

The results however spoke *against* most of our expectations: While information on rule identity was encoded in right VLPFC (BA 47), replicating results from Study 1, no information on rule order was evident in VLPFC. Instead, putamen and dorsal premotor cortex contained representations of rule order, both regions typically involved in sequence learning and arbitrary rule learning (e.g. Averbeck et al., 2006; Badre et al., 2010; Mushiake et al., 2006; Yin, 2009, 2010). In part, our findings are still compatible with the proposed compositionality principle, because representations of rule order and rule identity could be decomposed. The findings however *contradict* our expectations as well as theories on PFC function mentioned above (Duncan, 2001; Miller & Cohen, 2001) that predict that rule order should also have been available in VLPFC. Instead, our findings suggests that VLPFC might not be a general region that contains all necessary task set information, but might be conceived to be part of a larger system of specialised brain areas that cooperate to conduct more complex tasks (Frank & Badre, 2011).

* Equal contribution

In addition to the main results, the study produced two interesting side findings: First, the unexpected finding that information on rule identity was *not* evident in parietal cortex, as others and we had observed before (e.g. Bode & Haynes, 2009; Reverberi, Görgen et al., 2012a; Woolgar et al., 2011) and later (e.g. Pischedda*, Görgen* et al., 2017), but in temporal cortex (left temporal pole, BA 21; right posterior temporal lobe, BA 20, 21, 37). A potential explanation for this difference might be a seemingly small change compared to e.g. Study 1: Instead of using direct button presses as response ("press left", "press right"), participants had to remember the letters "A" and "B" as response for the rules. The letters appeared after a delay (from which we analysed the neural data) together with the target stimuli to which the participants had to apply the rule. Either "A" appeared on the left of the target and "B" on the right, or vice versa. Participants then had to press the button on the side where the letter appeared that they remembered (e.g. if the answer to a rule was "A", and "A" was on the left side of the target image, they had to press the left button). Using "A" and "B" made sure that participants could not anticipate which button to press before the target screen appeared, making motor preparation impossible[24]. This change arouse as consequence of a precursor study where we investigated the same question and observed preparatory motor signals that confounded the interpretation of rule order decoding results (Görgen, 2010). As we did not expect any major difference in representations, we were surprised to find such large-scale differences, with different brain areas mutually exclusively representing seemingly similar content (but see Sakai & Passingham, 2003).

The second interesting side finding was that we already tested an hypothesis for a potential confound that an influential paper one year later voiced as strong critique of our and others work: the idea that results would *not* arise from differences between rule representations, but from differences in reaction times, potentially reflecting difference in difficulty (Todd et al., 2013). In contrast to Todd and colleagues however, our analysis did not provide any evidence for this hypothesis (see also Woolgar et al., 2014), suggesting that our results were not suspect to the hypothesised difficulty confound. Both side findings, the discrepancy between "A"/"B" and "left"/"right" response rules as well as the reaction time control analysis to test for differences in difficulty later led to the development of "The Same Analysis Approach" (Study 4; Görgen et al., 2018; 3.4).

3.3 Study 3: Neural representations of hierarchical rule sets: The human control system represents rules irrespective of the hierarchical level they belong to

The previous study (Study 2; Reverberi*, Görgen* et al., 2012b; 3.2) confirmed our research hypothesis that the brain would represent rule order as an individual rule set feature and that the neural code of rule sets containing rule order would be compositional. However, we were surprised about the localisation of the representations. We had started the experiment with the hypothesis that the brain would represent rule order as a higher-level feature in a cognitive rule hierarchy, and – following recent "gradient theories" of prefrontal cortex organisation (Badre, 2008; Badre & D'Esposito, 2009; Koechlin et al., 2003) – had expected that rule order would have been represented anterior to rules themselves (Badre et al., 2009, 2010; Badre & D'Esposito, 2007; Badre & Frank, 2011; Koechlin et al., 2003; see Section 1.1). Because we found no representation about rule order anterior to the representations of rules in PFC (indeed, we did not find any evidence for the presence of rule order in PFC at all), we were not able to test this hypothesis. A potential explanation for why we did not find rule order representations in PFC might be that rule order

* Equal contribution

[24] This measure was necessary because even though we balanced the number of occurrences in which participants had to apply the first and the second rule, participants still prepared to apply the first rule (as can be seen from faster reaction time and less errors for the first rule, see Görgen, 2010). This activity most likely caused strong decoding accuracies in motor cortices, and also prevented clear interpretation of significant information on rule order in other areas (because these might have been caused by motor preparation, as well). In contrast, when using symbols as responses, participants could not prepare any motor response, and thus had no real benefit of preparing to apply one rule over the other.

is simply represented differently than rules of different levels because it is a temporal feature, which might explain the disagreement between the initial hypothesis and our neuroscientific findings.

We thus designed a further study (Pischedda*, Görgen* et al., 2017) to again test the gradient hypothesis (research aim three of this thesis, see Section 1.2), this time employing a rule hierarchy in which higher-level rules influenced the application conditions of lower-level rules. The design conforms to a number of abstraction definitions (Badre, 2008, 2013; Badre & D'Esposito, 2007, 2009; Botvinick, 2008; Petrides, 2005), as we discuss in the paper. The lower-level rules were compound rules similar to those from both previous studies. As in Study 1 (Reverberi, Görgen et al., 2012a; 3.1), the rules assign left and right button presses to specific target categories, for example "If you see a banana, press the left button". Application of the higher-level rules specified to which images the lower-level rules had to be applied. For example, "If you see a square, apply the lower-level rules only to images in blue frames". To minimize potential confounding alternative explanations, we implemented a number of measures to keep rules from both levels as similar as possible: The logical structure of rules from both levels was equivalent, the trial structure of all analysed trials was the same, and the instructing cues for both levels was randomly assigned to each participant from the same pool of symbols. A distinguishing feature of our study was that we measured representations of higher- and lower-level rules while participants yet only knew rules from that respective level, and were still waiting for the instruction for the other-level rule. This was to prevent any influences of the other-level rule to better allow localisation of representations of rules from the different levels (cf. Nee & Brown, 2012). Still, we made sure that participants already implement the rules of the level that had been instructed, and were thus able to identify *pure* representations of rules of each level *in the absence* of information of rules of the other level. As in Study 2 (Reverberi*, Görgen* et al., 2012b; 3.2), we hypothesised that higher-level rules would be represented more anteriorly in PFC than lower-level rules.

The results only partially confirmed our hypotheses: Both lower- and higher-level rules could be decoded from local patterns of brain activity, in prefrontal (e.g. VLPFC, BA 46, 47) and parietal (e.g. superior and inferior parietal lobe, BA 7 and BA 40) areas (Figure 3.2, panel 4). Apart from that, the results did not match our initial hypotheses. Especially, the results did not provide any evidence that different regions would encode higher- and lower-level rules, except in motor and premotor cortex (precentral gyrus, BA 6), where only lower-level rules were represented (this likely reflects that only lower-level rules contained motor information and is not speaking to gradient theories). This result directly contradicts the gradient hypothesis (Badre, 2008; Badre & D'Esposito, 2009; Koechlin et al., 2003) and instead suggests that a common brain network represents rules from both levels. As in Study 2, we performed extensive analyses to test whether the findings could have been caused by reaction time differences between conditions (Todd et al., 2013). We again found no evidence for this hypothesis. Indeed, Bayesian correlation analyses between reaction times and fMRI decoding performance even provided moderate evidence against that hypothesis.

Concurrent and independent from us, Nee and Brown (2012) performed a similar study that also tested the gradient hypothesis by decoding rules from two different cognitive levels. In contrast to us, however, they found large-scale differences between representations of their lower- and higher-level rules. As we discuss in our paper (Pischedda*, Görgen* et al., 2017), we see a number of alternative explanations that potentially explain the discrepancies between their and our results, especially the fact that they did not manipulate lower and higher rules independently. The similar designs of both studies that caused their different outcomes might also be fruitful to exploit in future studies. Specifically, the design of the two studies could serve as starting points to test which design features (if any) are critical to elicit rule

* Equal contribution

representations that are organised along a functional gradient, by systematically testing (i.e. experimentally manipulating) the differences between both.

In summary, the empirical Studies 1–3 of this thesis (Reverberi, Görgen et al., 2012a; Reverberi*, Görgen* et al. 2012b; Pischedda*, Görgen* et al., 2017) demonstrate that individual rule representations can be reliably decoded from local patterns of brain activity in prefrontal cortex as well as from further regions. The exact locations might depend on the specific types of rules that are represented (see Study 2; Reverberi*, Görgen* et al., 2012b; Section 3.2). While we demonstrated different types of compositionality of complex rule sets, we did not find evidence for the popular hypothesis that rules from different levels would be represented along a functional gradient in PFC (Badre, 2008; Badre & D'Esposito, 2009; Koechlin et al., 2003). Different reasons for this discrepancy between our experimental results and this hypothesis are conceivable, including differences in the operationalisation of the rule hierarchy or differences between the task periods that were investigated in different studies. Direct comparisons between these alternatives offer fruitful research questions for further empirical investigation.

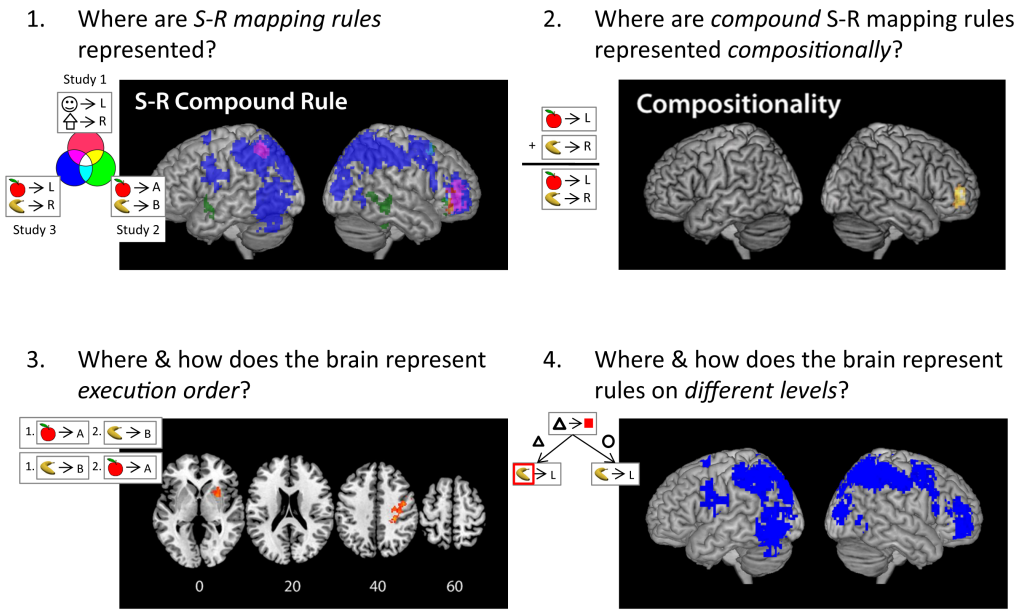

Figure 3.2 – Neural representations of different organisation principles of complex rule sets. The figure provides an overview about the central results of experimental studies of this thesis. **1.** Regions containing representations of stimulus response (S-R) mapping compound rules from all studies (red: Study 1, Section 3.1; green: Study 2, Section 3.2; blue: Study 3, Section 3.3). Left inlay depicts the tasks: Study 1 used rules that associated houses and faces to left/right button presses; Study 2 rules associated objects to letters that appeared to the left or right of the targets; Study 3 rules associated objects to button presses (lower level) and shapes to background colours (higher level, panel 4). See study descriptions for details. **2.** Compositionality coding of S-R compound rules was only found in right VLPFC (Study 1, Section 3.1). **3.** Regions containing representations of execution order (Study 2, Section 3.2; below brain slices: MNI z-coordinate). **4.** Regions containing higher modifier compound rules (associating shapes and colours) and lower S-R compound rules (Study 3, Section 3.3).

* Equal contribution

3.4 Study 4: The Same Analysis Approach (SAA) – Practical protection against the pitfalls of novel neuroimaging analysis methods

Study 4, the methodological study within this thesis (Görgen et al., 2018), presents "SAA – The Same Analysis Approach", a pragmatic approach to systematically detect, avoid, and eliminate confounds and other analysis errors in studies employing MVPA or other complex analysis methods. It developed from observations I made during empirical and theoretical work, including the empirical Studies 1–3 of this thesis (Reverberi, Görgen et al., 2012a; Reverberi*, Görgen* et al. 2012b; Pischedda*, Görgen* et al., 2017; 3.1–3.3), e.g. the side findings described in Section 3.2.

Why do we need such an approach? A common advice to researchers who employ MVPA is to be cautious. Unfortunately, there is only little advice on what exactly to be cautious for. In classical statistics, application practices for standard analysis methods (such as *t*-tests, *F*-tests, or ANOVAs) have been investigated very well (see e.g. Bortz, 2005; Bortz & Döring, 2002; Coolican, 2009; Cox & Reid, 2000). Standard textbooks on classical statistics do not only describe the merits of each method, but also best practice guidelines for their application, including systematic procedures to avoid numerous pitfalls, confounds, and caveats that have been discovered over the years. In MVPA, this process just seems to start.

A number of recent papers describe particular pitfalls of MVPA analysis, often together with specific solutions. These include surprising changes in distribution when using cross-validation (Jamalabadi et al., 2016; Noirhomme et al., 2014), wrong intuitions about interpretation of the group level *t*-test (Allefeld, Görgen et al., 2016), unexpected effects of task difficulty (Todd et al., 2013), or the problem of circular inferences (Vul et al., 2009; Kriegeskorte et al., 2009; Button, 2019; for more, see e.g. Haynes, 2015; Hebart & Baker, 2018). However, no general approach exists to detect problems (already described or novel) of a concrete experiment, i.e. a concrete experimental setup with fixed design and analysis method.

The aim of SAA is to provide such an approach. Starting with the intention to understand the reasons behind some peculiarities that we repeatedly observed during our analyses, we took a step back, looked at standard MVPA analysis as a whole, and compared it with standard procedures of classical statistical methods (such as *t*-/*F*-test, ANOVAs, e.g. Coolican, 2009; Cox & Reid, 2000). A first particularly interesting observation was that no specific design principle exists for experiments that are employing MVPA. Currently, nearly all MVPA studies use design principles that psychologists and biologists (see e.g. Bortz, 2005; Bortz & Döring, 2002; Coolican, 2009; Cox & Reid, 2000) employed for decades, such as randomisation and balancing, dating back at least to Fisher's fundamental work "The Design of Experiments" (Fisher, 1935). However, these design principles are specifically tailored for experiments that are analysed with specific statistical analysis methods (e.g. *t*-/*F*-test, ANOVAs).

In search for a remedy, we looked at how Fisher developed design principles for different analysis methods in the first place. The fundamental idea behind his work was that the experimental design should ensure to

- Measure the effect of interest as well as possible, and
- Avoid confounding influences from other variables by creating designs such that *even if a covariate indeed would contain information other than the one that should be experimentally investigated, this would not have an effect on the final analysis*.

From that, Fisher created his famous design principles, empowering researchers with tools that guard studies against experimental confounds.

* Equal contribution

Because MVPA methods – and other novel analysis techniques or complex analysis pipelines – differ from these established analyses, Fisher's guards might not work for them as intended. Therefore, other ways might be needed to fulfil Fisher's goals. For this, we propose turning Fisher's procedure around: Instead to *primarily create* designs with specific design principles that should exclude confounds (and then hope this will work), we suggest to *primarily verify* that a given design (created by whatever method) fulfils the design goals for a given analysis method (e.g. MVPA decoding). A core insight that led to this idea was our experience that finding *solutions* to avoid or eliminate confounds is often not too difficult, but that it is often hard to identify *that* a problem exists and *what* its causes are.

To verify if a design fulfils the design goals, we suggest to calculate if only the experimental variable(s) of interest – and no other known variable(s) – will influence the results. A convenient way to achieve this is to conduct *the same analysis* that is used to analyse the main data (e.g. decoding on neural data) also on design variables and other control data (e.g. to substitute any measured neural data by values of design factors, such as the number of a trial [the first trial value is "1", the second "2", etc.] or its reaction time), as well as on synthetic null data (to verify distribution assumptions). A property that makes SAA especially suited to control experimental studies is that it can be used throughout the whole process to detect, avoid, and eliminate confounds and other errors: from the design, to behavioural pre-experiments, to the final data analysis. SAA can thus fulfil the same function as unit testing in computer science (e.g. Myers et al., 2011), to automatically and continuously monitor problems that occur after changes of the experimental design or analysis. SAA further shares its logic with controlling procedures from other scientific fields, such as providing positive and negative controls as in chemistry or molecular biology, where often positive and negative probes are tested alongside the test data (e.g. Fedoroff & Richardson, 2001; Johnson & Besselsen, 2002) or in medicine during skin prick tests for allergy diagnosis (Rusznak & Davies, 1998).

In the paper (Görgen et al., 2018), we first introduce SAA on a motivating example, and then demonstrate its power to detect a wide range of confounds and errors in a number of scenarios where classical design principles systematically fail to control MVPA decoding analyses. These include failures that lead to false positive results (significant outcomes in the absence of a true effect) as well as false negatives (systematic suppression of real effects).

In general, we believe that SAA has the potential to facilitate MVPA studies and to reduce errors, because it provides a systematic approach to verify the experimental paradigm. Although we initially developed it for MVPA studies, SAA is not restricted to MVPA, but can be applied in other fields as well, within neuroimaging (for other analysis methods), but also in other fields that employ complex data analysis methods, such as genetics or machine learning. It will be interesting to see how the community receives SAA, which suggestions will be made to conduct "same" analyses, and which other potential confounds and pitfalls will be detected through the application of SAA.

4 General Discussion

In this thesis, I present three empirical studies that investigate where and how the human brain represents complex rule sets that are composed from different construction principles, as well as one methodological study on empirical design and experimental control principles of the employed methodology.

The three empirical studies (Studies 1–3; Reverberi, Görgen et al., 2012a; Reverberi*, Görgen* et al. 2012b; Pischedda*, Görgen* et al., 2017; Sections 3.1–3.3) investigated this question by decomposing the neural representation of complex rule sets using multivariate pattern analysis (MVPA) on functional magnetic resonance imaging (fMRI) data. The studies provide insights into the neural representations underlying three different construction principles: Composing *compound rules* of multiple individual rules (Study 1; Reverberi, Görgen et al., 2012a; Section 3.1); executing rules in a particular *order* (Study 2; Reverberi*, Görgen* et al., 2012b; Section 3.2); and constructing task sets that contain rules from *different hierarchical levels* (Study 3; Pischedda*, Görgen* et al., 2017; Section 3.3). Figure 3.2 provides a summary of the main empirical findings.

The methodological study (Study 4; Görgen et al., 2018; Section 3.4) provides insights into methodological pitfalls as well as information on how to detect, avoid, and eliminate these, in particular – but not exclusively – for MVPA.

The empirical studies in this thesis are part of a larger endeavour to decompose the neurocognitive architecture of human decision making (Bode & Haynes, 2009; Haynes et al., 2007; Hebart et al., 2018; Heinzle et al., 2012; Kahnt et al., 2010, 2011; Momennejad & Haynes, 2012, 2013; Nee & Brown, 2012, 2013; Soon et al., 2008; Tusche et al., 2010, 2013; Wisniewski et al., 2014). The methodological study connects to further projects to improve experimental methodology for neuroimaging (Allefeld, Görgen et al., 2016; Allefeld & Haynes, 2014; Haufe, Meinecke, Görgen et al., 2014; Hebart*, Görgen* et al., 2015; Soch et al., 2016).

Three studies (Cole et al., 2011; Nee & Brown, 2012; Woolgar et al., 2011) have independently and simultaneously investigated questions similar to those that we investigated in Studies 1 and 3 (Reverberi, Görgen et al., 2012a; Pischedda*, Görgen* et al., 2017; Sections 3.1, 3.3). Differences and similarities between these are discussed in the sections on the respective study (Sections 3.1 and 3.3).

This final chapter discusses aspects overarching the different studies. Discussions specific to each study are contained in the respective papers. General aspects of the experimental studies 1–3 (Reverberi, Görgen et al., 2012a; Reverberi*, Görgen* et al. 2012b; Pischedda*, Görgen* et al., 2017; Sections 3.1, 3.2, 3.3) are discussed first (Sections 4.1–4.4). Overarching aspects of the methodological study 4 (Görgen et al., 2018; Section 3.4) are discussed in Section 4.5. The thesis concludes with a discussion of open issues and suggestion for future work (Section 4.6).

4.1 General insights and implications from the empirical studies

Our studies provide initial evidence for a hypothesis that we put forward in Study 1 (Reverberi, Görgen et al., 2012a; 3.1), which we termed the *compositional principle*. We demonstrated that ventrolateral prefrontal cortex (VLPFC) encodes rules *compositionally* by showing that the neural patterns of compound

* Equal contribution

rules (rules that consist of two individual single rules, see Section 3.1) could be decoded using neural patterns of their individual composing rules. This demonstrates that compound rules and their individual composing rules at least partially elicit similar activity patterns. Study 1 also demonstrates that rules are encoded at least partially independently of the cues that were employed to instruct the rules. In Study 2 (Reverberi*, Görgen* et al., 2012b; 3.2), we replicated this finding and extended the scope of the compositionality principle (Reverberi, Görgen et al., 2012a; 3.1) by demonstrating that rule order was also represented independently of the identity of the instructing cues. These insights are further discussed in Section 4.2.

The most consistent finding across all studies was that information about all investigated rule set features was contained in local activity patterns within different large-scale networks across the brain.[25] These included the identity of rules, their order, and the identity of instructing cues. Classical stimulus-response mapping rules (S-R rules), which we investigated in all empirical studies, were always encoded within right VLPFC, matching all major theories in the field (e.g. Bunge & Wallis, 2008; Duncan, 2001; Fuster, 1989; Miller & Cohen, 2001; Sakai, 2008). A discrepancy between additional representations in posterior cortices, between parietal cortex (found in Studies 1 and 3; Reverberi, Görgen et al., 2012a; Pischedda*, Görgen* et al., 2017; 3.1, 3.3) and temporal cortex and striatum (found in Study 2; Reverberi, Görgen et al., 2012b*; 3.2), gives rise to a hypothesis on the exact role of these cortical areas. This hypothesis and further implications of the results for the current debate on how rule sets are encoded – whether in a general task set region or topographically specific – are discussed in Section 4.3.

The most unexpected and controversial finding from our empirical studies was the absence of evidence for a topographical organisation of rule representations from different hierarchical levels within VLPFC along an anterior-to-posterior gradient, which directly contradicts predictions from a number of popular recent theories on the functional organisation of prefrontal cortex (Badre, 2008; Badre & D'Esposito, 2009; Koechlin et al., 2003; see also Nee & D'Esposito, 2016; Schumacher et al., 2018). Instead, we found in Study 2 (Reverberi*, Görgen* et al., 2012b; 3.2) that rules defining an execution order of other rules (and are thus higher in a cognitive hierarchy) were encoded in dorsal striatum, putamen, and right premotor cortex, and in Study 3 (Pischedda*, Görgen* et al., 2017; 3.3) that full-fledged rules from different levels were encoded in the same areas irrespective of their hierarchical level. Implications of these results are discussed in Section 4.3.

4.2 Compositional versus non-compositional coding

An important question to which our results speak is whether rule sets are represented *compositionally* (e.g. Cole et al., 2010; Ruge & Wolfensteller, 2010), i.e. whether representations of full rule sets are built from the representation of its individual constituent components. Compositional coding of rule set features would explain how the brain represents the almost infinite number of complex rule sets humans can apply.

Previous studies found evidence for both compositional (e.g. Muhammad et al., 2006; Wallis et al., 2001; Wallis & Miller, 2003a, 2003b; Stoet & Snyder, 2004) and non-compositional (Warden & Miller, 2007, 2010) coding in monkeys. For example, a number of single-cell recording studies reported different neurons in monkey brains that represented different aspects of a task independently (e.g. Muhammad et al., 2006; Wallis et al., 2001; Wallis & Miller, 2003a, 2003b; for a summary, see Wallis, 2008). They found that a large portion of neurons were sensitive to the identity of one single rule set feature, which in these

* Equal contribution

[25] More precisely, information about individual features was present *locally* in all reported patches of the specific network. The information was *not distributed across* the network in the sense that activity of the full network or distant parts thereof was necessary to recover it.

studies consisted of cues (the task instructing symbols), rules, targets, and the response that the monkeys should perform. Thus, these neurons encoded the current rule set compositionally. Although the percentage of neurons encoding each respective feature depended on the exact recording location, information on these specific aspects was found throughout a widespread fronto-parietal-temporal network (e.g. Muhammad et al., 2006; e.g. Stoet & Snyder, 2004; Wallis et al., 2001; Wallis & Miller, 2003a, 2003b). On the other hand, other studies (Warden & Miller, 2007, 2010) also found evidence for non-compositional (or at least non-additive) coding of rule sets. In addition to neurons that only encoded the identity of one single rule set feature, those studies also identified neurons that encoded combinations of different aspects, thus implying non-compositional coding. This work also showed that representations of task sets in PFC change drastically whether a monkey kept one or two task-relevant stimuli in mind, and that representations additionally depend on the type of task the monkeys had to perform.

The empirical Studies 1–3 (Reverberi, Görgen et al., 2012a; Reverberi*, Görgen* et al. 2012b; Pischedda*, Görgen* et al., 2017; 3.1, 3.2, 3.3) of this thesis directly identified three types of compositional coding: 1. compositional coding of compound rules and their constituent single rules; 2. compositional coding of cues and the rule set features that were instructed by them (rules and rule order); and 3. compositional coding of rules and their order. The following subsections summarise how our findings support each of these points.

4.2.1 Compositional coding of compound rules

Study 1 (Reverberi, Görgen et al., 2012a; 3.1) provides direct evidence for compositional coding of compound rules from their constituent single rules. To the best of our knowledge, the neural underpinnings of this principle have not been investigated before. Some previous fMRI studies have investigated general activation differences between compound and single rules (see Bunge, 2004; Bunge & Zelazo, 2006), which however can only speak to general involvement, but not to the represented content (see Chapter 2). Other studies have investigated compositional coding in monkeys (e.g. Siegel et al., 2009; Warden & Miller, 2007, 2010), however, not for rules, but for task-relevant stimuli in short-term memory. In Study 1 (Reverberi, Görgen et al., 2012a; 3.1), we showed that right VLPFC represented compound rules compositionally. Specifically, we showed that the identity of compound rules could be predicted by classifiers trained from data of the composing single rules (and vice versa). Interestingly, no evidence for compositional coding was present in parietal areas, although there – like in PFC – single and complex rules could be decoded independently. This suggests that VLPFC, but not parietal cortex, employs compositional coding for compound rules.

4.2.2 Compositional coding of cues and rules/rule order

Our studies also provide direct evidence for compositionality of instructed rule set features. This comes from both, Study 1 (Reverberi, Görgen et al., 2012a; 3.1), where compositionality between cues and the instructed rules was present (see also Woolgar et al., 2011), and Study 2 (Reverberi, Görgen et al., 2012b*; 3.2) where we replicated this finding and additionally demonstrated compositionality between cues and rule order. As stated above, this agrees with findings from monkey neurophysiology, where a number of studies (e.g. Muhammad et al., 2006; Sigala et al., 2008; Wallis et al., 2001; Wallis & Miller, 2003a; Stoet & Snyder, 2004) found separate encoding of cues and rules in a wide fronto-parietal region. In contrast to these studies however, we found a clear separation between regions containing information on cues and on rules. While the neurophysiological studies in monkeys found both rule and cue information in frontal (e.g. Muhammad et al., 2006; Sigala et al., 2008; Wallis et al., 2001; Wallis & Miller, 2003a) and parietal (Stoet & Snyder, 2004) areas, with a considerable fraction of neurons representing both cues and rules,

* Equal contribution

we identified cue information only in parietal areas.[26] One potential explanation for this discrepancy might be that these studies used cues from different modalities, e.g. sounds to instruct different rules instead of visual symbols. Because these tasks require multimodal integration, this might have produced a stronger difference in neural firing compared to the task in our experiment, where both cues and targets were presented visually.

4.2.3 Compositional coding of rules and their order

In Study 2 (Reverberi*, Görgen* et al., 2012b; 3.2), we found compositional coding for task sets that contain two rules and a specific execution order. Specifically, we found a spatial segregation between rule identity and rule order. While representations of rule order were present in right premotor cortex and the dorsal striatum, we did not find any evidence for rule order in PFC, including VLPFC. If rule order indeed were not represented in VLPFC, this would contradict a recent hypothesis that suggests VLPFC as a general controller for task set preparation that controls how posterior brain regions perform the task and how they work together (e.g. Bengtsson et al., 2009; Sakai & Passingham, 2003, 2006). However, although our study provides evidence that VLPFC does not contain information on rule order, it does of course not rule out the possibility completely.

One reason is that negative findings are generally difficult to interpret, because an experimental analysis can always miss an effect that actually exists. In our case, we believe this is unlikely. First, we additionally employed region of interest (ROI) analyses to test for presence of information in VLPFC with higher sensitivity. These did also not show any evidence for presence of information in VLPFC. Second, a region (VLPFC versus striatum) by information (rule identity versus order) interaction effect exists (Study 2; Reverberi*, Görgen* et al., 2012b; 3.2). While a significant interaction does not directly provide evidence for the null hypothesis if one of the tested groups did not show a significant effect for one factor (here rule order in VLPFC), it does demonstrate an inhomogeneity between identity and order information between both brain areas. This means that if information about rule order were present in VLPFC, it would have a significantly smaller effect on activation patterns than in striatum, compared to the effect that rule identity information has in both regions.

A second reason for how VLPFC could represent rule order even though we did not find it is the intriguing possibility that rule order is represented in a different representational *format* that our experimental setup was not able to detect. To our knowledge, rule order as an organising feature of task sets has not been investigated before, but order representations have been investigated in other contexts, such as for neuronal encoding of the order of task relevant objects (Siegel et al., 2009) or for representations of sequences such as spatial locations or movement sequences (Warden & Miller, 2007, 2010). The study of Siegel et al. (2009) suggests that rule order might indeed be opaque to our analysis method. In that study, representations of task sets were investigated while monkeys maintained two objects (not rules) and their order in working memory. While Siegel and colleagues could decode identity and order from the firing of neurons in PFC, they also found that the different objects were represented stronger at different phases of the oscillatory components of the local field potential. In particular, the first object was represented stronger at the beginning of each cycle and the second object was represented stronger at the end. If we assume that human PFC represents rules in a similar way, then the rules in our experiment might also have been represented in fast alternations (and thus information about rule order would have indeed been present in PFC). The temporal precision of the BOLD signal (see Section 2.2) is on the order of seconds, and thus fMRI data will most certainly not allow to identify quickly alternating signals. If this were correct, it would explain why we only found information on rule identity in PFC, but not on rule

[26] In our studies, information about cues was also present in visual areas in occipital cortex. Although we cannot compare our results to the monkey studies directly, because these studies did not record in visual areas and used different modalities to provide cues, overwhelming evidence exists that occipital cortex is sensitive to visual input, and thus information on visual cues should exist in monkey occipital cortex.

* Equal contribution

order: The identity of the two rules would then have been clearly distinguishable from their spatially distributed activity patterns, while their temporal order (encoded by fast temporal alternations between these spatial patterns) would not. Note that such temporal coding does not speak against compositional coding. On the contrary, it actually represents an interesting additional compositionality principle (although one that BOLD fMRI MVPA cannot detect) that could be used to encode information in addition to that encoded in spatial patterns. An interesting question that these considerations lead to is how this temporal order is established in the first place, and whether this is related to the representations of rule order that we found encoded as spatial patterns in striatum (see Womelsdorf & Valiante, 2014 for related considerations).

Taken together, we find that the brain uses a number of different types of compositional coding to represent complex rule sets (Sections 4.2.1, 4.2.2). This is consistent with further work from our group (e.g. Momennejad & Haynes, 2012, 2013; Wisniewski et al., 2014) and others (Cole et al., 2011; Nee & Brown, 2012; Woolgar et al., 2011) that have investigated similar issues (see Sections 3.1, 3.3).

4.3 Distributed coding versus a single task set region

Our results also speak to how the brain distributes representations of complex rule sets (Badre & Frank, 2011; Botvinick, 2008; Crittenden & Duncan, 2014; Duncan, 2001; Fedorenko et al., 2013; Frank & Badre, 2011; Heinzle et al., 2012; Koechlin & Summerfield, 2007; Sakai & Passingham, 2003, 2006). One popular theory postulates that all task-relevant information is represented and processed by one single general network, the fronto-parietal "Multiple Demand" (MD) network (Duncan, 2001, 2013; Crittenden & Duncan, 2014; Fedorenko et al., 2013), which can flexibly allocate resources, similar to memory of standard computers. An alternative theory states that different aspects of complex rule sets are represented in different specialised regions that dynamically collaborate during task execution (Badre & Frank, 2011; Botvinick, 2008; Frank & Badre, 2011; Heinzle et al., 2012; Koechlin & Summerfield, 2007; Sakai & Passingham, 2003). In the latter theory, lateral PFC (especially VLPFC) sticks out as a specialised task set region, a "final common pathway" (Bengtsson et al., 2009), where all task set relevant information come together and which orchestrates the other specialised areas during task performance. The results of our empirical studies speak more towards this second theory of collaboration between regions. They however also restrict that theory, specifically the function of VLPFC as a general task set region (see Section 4.3.2).

4.3.1 Restrictions to dynamic allocation and fronto-parietal multiple demand network

A number of our findings speak against the idea of dynamic allocation of the "multiple demand" (MD) theory (Duncan, 2001, 2013; Crittenden & Duncan, 2014; Fedorenko et al., 2013). First, we found in our empirical studies that all investigated rule set features (cue identity; single, compound, high- and low-level rule identity; and rule order) were represented in local activation patterns. This finding speaks against dynamic allocation in its most general form: Would the brain allocate resources completely flexibly, the same task set features should produce different patterns of brain activity in each trial. The fact that local patterns of brain activity carry information about the identity of specific rule set features requires however that the representations share at least some location specific activity, which makes our findings incompatible with the idea of flexible allocation.

Further evidence against dynamic allocation comes from the observation that in our studies feature representations were never available throughout the entire suggested fronto-parietal network, but were location specific, both in frontal (Studies 1–3; Reverberi, Görgen et al., 2012a; Reverberi*, Görgen* et al., 2012b; Pischedda*, Görgen* et al., 2017; 3.1, 3.3) and posterior regions (Studies 1, 3; Reverberi, Görgen et al., 2012a; Pischedda*, Görgen* et al., 2017; 3.1, 3.3). Within frontal cortex, S-R rules were specifically

* Equal contribution

located in VLPFC. This finding was consistent across all three studies (Figure 3.2, panel 1), and independent of the type of rule (single or compound rule, lower or higher hierarchical levels). The other investigated feature, rule order, was also encoded location-specific in frontal cortex, specifically in premotor cortex. In posterior regions, Studies 1 and 3 (Pischedda*, Görgen* et al., 2017; 3.1, 3.3) found representations of S-R rules in parietal cortex, which is in line with the general topography of the suggested fronto-parietal network. In Study 2 (Reverberi*, Görgen* et al., 2012b; 3.2), however, representations of S-R rules were found in temporal, but not in parietal cortex. If indeed no such information was present in parietal cortex (i.e. if our study did not miss information that was there; see Section 4.2.2 for reasons that speak against a false negative), this would directly challenge the "broad domain generality" assumption of the fronto-parietal network (as suggested in e.g. Fedorenko et al., 2013). In summary, our results speak against the general hypothesis of dynamic allocation and a general front-parietal network as postulated by the MD theory.

Is there an explanation how the discrepancy might arise that parietal (and not temporal) cortex is routinely found in many studies including Study 1 and 3 (Reverberi, Görgen et al., 2012a; Pischedda*, Görgen* et al., 2017; 3.1, 3.3) of thesis (forming the basis for the "broad domain generality" claim of Fedorenko et al., 2013), but that Study 2 (Reverberi*, Görgen* et al., 2012b; 3.2) found rule representations in temporal (and not parietal) cortex? As we suggest in the discussion of Study 2 (Reverberi*, Görgen* et al., 2012b; 3.2), the critical difference between our Study 2 and most other studies might be the required response type: Most studies employ direct motor actions that include a spatial component as response (such as "If X, press the left/right button", "look to the top/bottom", etc.). Study 2 (Reverberi*, Görgen* et al., 2012b; 3.2), in contrast, employs an object search tasks as response ("If X, press the side where A/B appears") that can only be resolved to a spatial motor action when the target image occurs (which is later than the period from which we analyse fMRI data in our empirical studies). Based on this, we hypothesise that parietal cortex might represent rules only if they employ spatial action responses ("If X, press the left/right button"). Temporal cortex, in contrast, might represent rules if they use object search task responses ("If X, press the side were A/B appears"). Similar to Section 4.2.2, empirical support against a potential false negative result (the possibility that we missed rule identity information in parietal cortex) comes from the fact that ample information was present in parietal cortex in other studies. This includes the other two empirical studies in this thesis (Studies 1 and 3; Reverberi, Görgen et al., 2012a; Pischedda*, Görgen* et al., 2017; 3.1, 3.3) and a precursor study that investigated the same topic with the same design but motor action instead of search task responses (Görgen, 2010).

The hypothesis of segregated functional specialisation of parietal and temporal cortex fits well to the "dual-stream" hypothesis from vision (Goodale & Milner, 1992; Milner & Goodale, 2008; Mishkin et al., 1983; Mishkin & Ungerleider, 1982). According to this hypothesis, visual information is split in visual cortex into two streams with different functions that take different routes to frontal cortex. A parietal stream goes "upward" from visual cortex via parietal cortex, and a temporal stream goes "downward" via temporal cortex. The original hypothesis asserted that parietal cortex would process spatial "where" information while temporal cortex would process symbolic "what" information (Mishkin & Ungerleider, 1982; Mishkin et al., 1983). An updated version of the theory asserts that the parietal stream processes "how" things can be manipulated, instead of merely determining "where" they are (Goodale & Milner, 1992; Milner & Goodale, 2008; but see Schenk et al., 2011). Our findings are in line with this idea, but also recast the hypothesis in a slightly more general interpretation: It might be that the two streams not only process "where/how" and "what" information of visual objects, but that they process rules with specific "where/how" or "what" responses. That is, the parietal stream might not only extract "where" an object is located and/or "how" it can be manipulated, and the ventral stream might not only extract "what" an object is, but rules how to use this information are directly associated together with this information.

* Equal contribution

Together with the observation that rule representations were contained in VLPFC in Study 2 (Reverberi*, Görgen* et al., 2012b; 3.2) as in most other studies (including Study 1 and 3; Reverberi, Görgen et al., 2012a; Pischedda*, Görgen* et al., 2017; 3.1, 3.3), this might mean that posterior regions might be recruited because they implement a specific *task*, while VLPFC is a general rule processing region that recruits these areas (see Li et al., 2007 for a similar conclusion). This does *not* mean that specific rules will not be implemented in the posterior areas – on the contrary, our studies suggest that they do indeed contain the rules, i.e. the specific link between triggering conditions and responses. However, the specific rules will only be implemented in posterior areas if they are actually involved in executing these rules. Thus, these findings suggest that also single S-R rules are composed of (at least) two individual parts that are distributed to different brain areas, general rule information in VLPFC, and its specific implementation in posterior brain areas, the location of which depends on the specific type of response. This hypothesis that parietal cortex represents rules that employ "how" responses would explain why the fronto-parietal network is so commonly found in task set research. Although previous studies investigated many different *tasks* and *rules* (see e.g. Fedorenko et al., 2013), they all used similar rule *types* that used motor actions (i.e. "how") responses, which – according to our hypothesis – would create representations in parietal cortex.[27] Because this is a post-hoc hypothesis, further work is required to test this possibility (see Section 4.6.1).

4.3.2 Evidence for distributed coding with restrictions to VLPFC as general controller

Most of our findings speak for the alternative theory of distributed coding, i.e. that a distributed network of specialised regions underlies rule set representation (Badre & Frank, 2011; Botvinick, 2008; Frank & Badre, 2011; Heinzle et al., 2012; Koechlin & Summerfield, 2007; Sakai & Passingham, 2003):

1. Rule set information was never available across the whole fronto-parietal network, but specifically located to certain areas.
2. Information on cues, rules, and rule order were localised in different regions.
3. Information on cues was present mainly in visual cortices.
4. Rule representations existed in VLPFC and in posterior cortices, where representation location depended on rule response type (either parietal cortex for "how" response rules, or temporal cortex for "what" response rules; see Section 4.3.1).
5. Information on rule order was present in striatum and motor cortex.

Still, findings from Study 2 (Reverberi*, Görgen* et al., 2012b; 3.2) speak against one integral part of the theory: The idea that VLPFC would be a *general* controller for task sets (e.g. Bengtsson et al., 2009; Sakai & Passingham, 2003, 2006). Specifically, we found that rule order was most likely not represented within VLPFC, but that it instead was represented in striatum and premotor cortex. If that were true and not an artefact of our analysis method (see Section 4.2.2), important task-set relevant information would be missing in VLPFC. This would mean that VLPFC would not be a completely general task set controller, but would restrict its role to that of a more specific region that would represent and process cognitive rules. If that were the case, it would pose the question which other region might contain all information and coordinate the interaction of the specialised other regions, or – if no such region exists – how the brain would configure this dynamic interplay otherwise.

In summary, the results discussed in Sections 4.3.1 and 4.3.2 suggest that the commonly found "fronto-parietal" network is *not* a "Multiple Demand Network" (Duncan, 2006, 2010; Fedorenko et al., 2013) that contains and processes all kind of task-related information (like random access memory in standard

* Equal contribution

[27] For example, to support their claim, Fedorenko and colleagues (2013) present seven new fMRI experiments that investigate a wide variety of different intelligence tasks; however, all employ action responses (button presses in six tasks, verbal feedback in one task).

computers). Instead, our results suggests that the fronto-parietal network is typically observed because it represents one specific rule type commonly used in most experiments – S-R rules that contain "how" responses (here: specific motor actions). In addition, contrasting S-R mappings containing "how" with "what" responses (here: a letter search task) suggest a specific distinction between the function of parietal and temporal cortex. Finally, localisation of rule order, another feature of complex rule sets, speaks against VLPFC as a general region for task sets (see e.g. Bengtsson et al., 2009; Sakai & Passingham, 2003, 2006), but suggests VLPFC to be a more specialised region that represents and processes rules.

4.4 A functional gradient within PFC?

A final discussion to which our results speak concerns the neurocognitive architecture underlying rule sets *within* PFC. A number of theories suggest that PFC represents features of complex task sets topographically (Badre, 2008; Badre & D'Esposito, 2009; Badre & Nee, 2018; Bahlmann et al., 2015; Koechlin et al., 2003; Koechlin & Summerfield, 2007). Specifically, these theories propose a rostro-caudal gradient within PFC, along which rules at different levels within a cognitive hierarchy would reside. Individual theoretical accounts differ by the exact progression of the proposed axes and by which features define the cognitive gradient, but all theories locate lower rules more posterior within PFC, and progressively higher rules more anterior (see Badre, 2008).

We experimentally tested this core prediction of gradient theories in two of our studies (Studies 2 and 3; Reverberi*, Görgen* et al., 2012b; Pischedda*, Görgen* et al., 2017; 3.2, 3.3). Both investigated representations of rule sets comprising two levels within a cognitive hierarchy. In both studies, the lower level consisted of two complementary S-R rules. The higher control level was a specific execution order for the lower rules in Study 2 (Reverberi*, Görgen* et al., 2012b; 3.2). In Study 3 (Pischedda*, Görgen* et al., 2017; 3.3), explicit higher-level rules instructed to which stimuli the lower rules had to be applied.

As discussed above (Sections 3.2, 4.3.2), in Study 2 (Reverberi*, Görgen* et al., 2012b; 3.2) representations of rule order (in striatum and premotor cortex) were spatially segregated from representations of rule identity (in VLPFC). The fact that rule order was represented outside PFC speak against gradient theories. The results from Study 3 (Pischedda*, Görgen* et al., 2017; 3.3) also speak against gradient theories, because representations of rules from both hierarchical levels showed distinct topographies that did not differ significantly between levels, showing no sign for an anterior-to-posterior gradient. This was particularly surprising because another recent fMRI MVPA study reports evidence in favour of a functional gradient (Nee & Brown, 2012).[28] As discussed in Study 3 (Pischedda*, Görgen* et al., 2017; 3.3), one potential explanation for this discrepancy might be that gradients only arise if both rules are represented at the same time (during the analysed period in Study 3 participants held only rules from one level [lower or higher] in memory; rules from the other level were instructed later). If that were the case, this would be indeed surprising, because it would require that rule representations change locations between different task phases, and not just representational format (as in Warden & Miller, 2007; Sigala et al., 2008; Hebart et al., 2018). Further work is necessary to test this possibility (see Section 4.6.2).

4.5 Discussion of methodological study

The methodological Study 4 (Görgen et al., 2018; 3.4) of this thesis is part of the recent endeavour to "mature" MVPA (similar to the maturation of classical statistics in the beginning of last century, e.g. by the work of Fisher, 1935). The study has been motivated by observations I made among others during data analyses of the empirical studies of this thesis (Studies 1–3; Reverberi, Görgen et al., 2012a; Reverberi*, Görgen* et al., 2012b; Pischedda*, Görgen* et al., 2017; 3.1, 3.3) as well as observations of similar problems

* Equal contribution

[28] As discussed in Section 3.3, it might be interesting to harvest the differences between both studies to test by systematic experimental manipulation which of their design features (if any) are critical to elicit gradients.

reported by others (Allefeld, Görgen et al., 2016; Etzel et al., 2013; Etzel & Braver, 2013; Haufe, Meinecke, Görgen et al., 2014; Mumford et al., 2012, 2015; Noirhomme et al., 2014; Schreiber & Krekelberg, 2013; Todd et al., 2013; Woolgar et al., 2014; Hebart & Baker, 2018; for review, see e.g. Haynes, 2015). As MVPA is a rather young and rapidly developing methodological discipline in neuroimaging, our studies aimed to analyse experimental designs of MVPA to detect, avoid, and eliminate confounds and other pitfalls.

The "Same Analysis Approach (SAA)" (Study 4, Görgen et al., 2018; 3.4) deals with a question that lies at the foundation of any experimental study: How to set up the experimental design to ensure that only the experimental variable(s) of interest will influence the final statistical assessment? The key insight that has motivated this study is that different analysis methods require different experimental designs. The insight which is not novel at all but apparently has been ignored in the MVPA community so far; Fisher (1935) has used it to develop his today classical design principles. An important implication is that the current wide-spread practice to design experiments using design principles (such as counterbalancing) that have been developed for conventional analysis methods (such as *t*-tests or ANOVAs) do not necessarily safeguard against the influence of potential confounds if novel analysis methods (such as MVPA) are employed. As a practical, systematic way to test whether a given design controls a desired analysis method, we suggest to apply this same analysis to a variety of control datasets constructed from the design, and examine the result pattern that these analyses yield. In short: If only the experimental variable(s), but no confounder variable(s), systematically influence the final statistical result, this provides empirical evidence for the validity of the design. An important additional point we make is that the same principle – to test which result the same analysis method would yield if a different variable would influence the analysed data – is also necessary when performing control analyses on control data (such as reaction times). Indeed, a common practice in cognitive neuroimaging is to employ different tests for control analyses and main analyses, such as univariate *t*-tests, *F*-tests, or ANOVAs on control data (e.g. reaction times, age, or IQ), and MVPA on the main experimental data (e.g. fMRI or EEG). The reason behind this seems to be that control data is univariate and has always been analysed using *t*-test, *F*-tests, or ANOVAs. However, because different types of tests are not sensitive to the same properties of data (as we demonstrate in the paper using examples), inferences between tests are invalid. Instead, we show that structurally equivalent tests ("same" analyses) are necessary for valid control analyses, too.

Some ideas underlying SAA have been already incorporated in the analyses of the empirical Studies 2 and 3 (Reverberi*, Görgen* et al., 2012b; Pischedda*, Görgen* et al., 2017; 3.2, 3.3). In Study 2 (Reverberi*, Görgen* et al., 2012b; 3.2), we correlated absolute differences in error rate and reaction time between experimental conditions (different rules that had to be applied) with fMRI decoding accuracies across participants to test the hypothesis that a potential confound, namely differences in difficulty of rules within each participant, could explain our fMRI decoding results. A year later, the same hypothesis has been put forward by Todd and colleagues as a methodological criticism of our and others work (Todd et al., 2013). In contrast to Todd et al., but in agreement with others (Woolgar et al., 2014), we did not find any evidence for this hypothesis. Repeating the analysis in Study 3 (Pischedda*, Görgen* et al., 2017; 3.3), i.e. using the same analysis used for imaging data (i.e. decoding) for reaction times yielded the same result, again speaking against Todd et al.'s hypothesis that reaction times would confound results.

Using its potential, the "Same Analysis Approach" (Study 4; Görgen et al., 2018; 3.4) could, I believe, have a high impact on quality, reproducibility, interpretability, and effectiveness for conducting future studies.

* Equal contribution

4.6 Open issues and future directions

4.6.1 Parietal "how" versus temporal "what" response

One unresolved issue is that in Study 2 (Reverberi*, Görgen* et al., 2012b; 3.2) we unexpectedly found S-R rule representations only in temporal cortex, but no sign for representations in the commonly observed parietal cortex, which we found – as most other studies – in Studies 1 and 3 (Reverberi, Görgen et al., 2012a; Pischedda*, Görgen* et al., 2017; 3.1, 3.3).

As stated above (Section 4.3.2), we hypothesised that this discrepancy might be caused by differences between rules that involve action "where/how" responses ("If X, press the left button") versus symbolic "what" responses, in our experiments instructions to locate objects ("If X, find the letter A, and then press the button on the side where that letter occurred"). This hypothesis agrees with the dual-stream hypothesis (Mishkin et al., 1983; Goodale & Milner, 1992) from vision neuroscience that postulate a parietal "where/how" versus a temporal "what" pathway, but extends its scope to content-specific representations of rule consequences. It thus corroborates the idea that rule information is stored in regions that also process this information (as hypothesised in Study 2; Reverberi*, Görgen* et al., 2012b; 3.2) and at the same time restricts the role of parietal cortex from a general task sets processing region to the more specific task to process "where/how" response rules. Further work would be required to test this hypothesis.

4.6.2 Gradient in PFC?

A second open question concerns our findings in Studies 2 and 3 (Reverberi*, Görgen* et al., 2012b; Pischedda*, Görgen* et al., 2017; 3.2, 3.3), which contradict the idea of a functional gradient within PFC, especially because a similar study (Nee & Brown, 2012) reports evidence for these theories.[29] A promising way to resolve the issue would be a systematic stepwise transition between experiments that do and that do not find topographical differences. Study 3 (Pischedda*, Görgen* et al., 2017; 3.3) and the study by Nee & Brown (Nee & Brown, 2012) seem ideal candidates as a starting point, because both used similar paradigms and analysis methods. MEG-fMRI fusion (e.g. Cichy et al., 2014) based on representational similarity analysis (Kriegeskorte et al., 2008) as recently employed to trace task-set components in space and time (Hebart et al., 2018) could be a promising way to investigate how task-set representations of hierarchical rule sets build up during application.

4.6.3 Making SAA run

With SAA (Study 4; Görgen et al., 2018; 3.4), we made some suggestions that might help researchers doing better experiments, thereby saving time, money, and avoiding frustration. Yet, an important step to make researchers actually use SAA would be to provide a concrete implementation to make it easy for researchers to apply SAA to their experiments. An implementation in beta state is available for "TDT – The Decoding Toolbox" (Görgen et al., 2012; Hebart*, Görgen* et al., 2015) and has been successfully tested by a handful of users; some documentation on how to create SAA implementations for developers exist as well (both available upon request). Further work would be required to bring software and documentation to release state and to make it publicly available for download. I am convinced that this effort would be worth the while: Not only because it would boost the adoption of SAA; but also because of the good it would bring to the world of science.

[29] Note however that substantially different gradient theories exist, all of which mainly rest on indirect evidence. Despite the wealth of secondary literature, only very few studies directly tested the hypothesis in humans (Koechlin et al., 2003; Kouneiher et al., 2009; Badre & D'Esposito, 2007; Nee & Brown, 2012; Reynolds et al., 2012; Bahlmann et al., 2015). None of them has been replicated by independent groups, and the only published replication attempt (Reynolds et al., 2012) that tested the two most common theories failed to replicate both. Given all this, a critical review of the literature (all major reviews come from the same two groups) and maybe reanalyses of the data would be highly desirable.

* Equal contribution

5 References

Allefeld, C., Görgen, K., & Haynes, J.-D. (2016). Valid population inference for information-based imaging: From the second-level t-test to prevalence inference. *NeuroImage*, *141*, 378–392. doi:10.1016/j.neuroimage.2016.07.040

Allefeld, C., & Haynes, J.-D. (2014). Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. *NeuroImage*, *89*, 345–357. doi:10.1016/j.neuroimage.2013.11.043

Allport, D. 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 15: Conscious and nonconscious information processing* (pp. 421–452). Cambridge, MA, US: The MIT Press.

Alvarez, J. A., & Emory, E. (2006). Executive Function and the Frontal Lobes: A Meta-Analytic Review. *Neuropsychology Review*, *16*(1), 17–42. doi:10.1007/s11065-006-9002-x

Anderson, J. R. (1976). *Language, memory, and thought*. Oxford, England: Lawrence Erlbaum.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *22*(3), 261–295. doi:10.1016/S0022-5371(83)90201-3

Anderson, J. R. (1993). *Rules of the Mind*. Psychology Press.

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. doi:10.1152/jn.2000.84.1.451

Averbeck, B. B., Crowe, D. A., Chafee, M. V., & Georgopoulos, A. P. (2003). Neural activity in prefrontal cortex during copying geometrical shapes. Part II. Decoding shape segments from neural ensembles. *Experimental Brain Research*, *150*(2), 142–153. doi:10.1007/s00221-003-1417-5

Averbeck, B. B., Sohn, J. W., & Lee, D. (2006). Activity in prefrontal cortex during dynamic selection of action sequences. *Nature Neuroscience*, *9*(2), 276–282.

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. (2013). Hierarchical Cognitive Control and the Functional Organization of the Frontal Cortex. In *The Oxford Handbook of Cognitive Neuroscience, Volume 2: The Cutting Edges* (Vol. 2, p. 300). Retrieved from http://books.google.de/books?hl=de&lr=&id=UdhBAgAAQBAJ&oi=fnd&pg=PA300&dq=Hierarchical+Cognitive+Control+and+the+Functional+Organization+of+the+Frontal+Cortex&ots=0FpKe8sPCE&sig=yVUbFwX836fLLagDum1Msvlh2nM

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., & Frank, M. J. (2011). Mechanisms of Hierarchical Reinforcement Learning in Cortico-Striatal Circuits 2: Evidence from fMRI. *Cerebral Cortex*. doi:10.1093/cercor/bhr117

Badre, D., Hoffman, J., Cooney, J. W., & D'Esposito, M. (2009). Hierarchical cognitive control deficits following damage to the human frontal lobe. *Nature Neuroscience*, *12*(4), 515–522. doi:10.1038/nn.2277

Badre, D., Kayser, A. S., & D'Esposito, M. (2010). Frontal cortex and the discovery of abstract action rules. *Neuron*, *66*(2), 315–326. doi:10.1016/j.neuron.2010.03.025

Badre, D., & Nee, D. E. (2018). Frontal Cortex and the Hierarchical Control of Behavior. *Trends in Cognitive Sciences*, *22*(2), 170–188. doi:10.1016/j.tics.2017.11.005

Bahlmann, J., Blumenfeld, R. S., & D'Esposito, M. (2015). The Rostro-Caudal Axis of Frontal Cortex Is Sensitive to the Domain of Stimulus Information. *Cerebral Cortex*, *25*(7), 1815–1826. doi:10.1093/cercor/bht419

Banich, M. T., Milham, M. P., Atchley, R. A., Cohen, N. J., Webb, A., Wszalek, T., … Brown, C. (2000). Prefrontal regions play a predominant role in imposing an attentional "set": evidence from fMRI. *Brain Research. Cognitive Brain Research*, *10*(1–2), 1–9.

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

Baron, S. G., Thompson-Schill, S. L., Weber, M., & Osherson, D. (2010). An early stage of conceptual combination: Superimposition of constituent concepts in left anterolateral temporal lobe. *Cognitive Neuroscience*, *1*(1), 44–51. doi:10.1080/17588920903548751

Barone, P., & Joseph, J. P. (1989). Prefrontal cortex and spatial sequencing in macaque monkey. *Experimental Brain Research*, *78*(3), 447–464.

Beckett, A., Peirce, J. W., Sanchez-Panchuelo, R.-M., Francis, S., & Schluppeck, D. (2012). Contribution of large scale biases in decoding of direction-of-motion from high-resolution fMRI data in human early visual cortex. *NeuroImage*, *63*(3), 1623–1632. doi:10.1016/j.neuroimage.2012.07.066

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

Bianchi, L. (1922). *The Mechanism of the brain and the function of the frontal lobes*. New York : William Wood ; Edinburgh : Livingstone. Retrieved from http://archive.org/details/mechanismofbrain00bianrich

Bießmann, F., Meinecke, F. C., Gretton, A., Rauch, A., Rainer, G., Logothetis, N. K., & Müller, K.-R. (2010). Temporal kernel CCA and its application in multimodal neuronal data analysis. *Machine Learning*, *79*(1–2), 5–27. doi:10.1007/s10994-009-5153-3

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

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

Bortz, J. (2005). *Statistik*. Springer London, Limited.

Bortz, J., & Döring, N. (2002). *Forschungsmethoden und Evaluation: für Human- und Sozialwissenschaftler ; mit ... 70 Tabellen*. Springer DE.

Botvinick, M. M. (2007). Multilevel structure in behaviour and in the brain: a model of Fuster's hierarchy. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *362*(1485), 1615–1626. doi:10.1098/rstb.2007.2056

Botvinick, M. M. (2008). Hierarchical models of behavior and prefrontal function. *Trends in Cognitive Sciences*, *12*(5), 201–208.

Botvinick, M. M., & Bylsma, L. M. (2005). Distraction and action slips in an everyday task: Evidence for a dynamic representation of task context. *Psychonomic Bulletin & Review*, *12*(6), 1011–1017.

Botvinick, M., & Plaut, D. C. (2004). Doing Without Schema Hierarchies: A Recurrent Connectionist Approach to Normal and Impaired Routine Sequential Action. *Psychological Review*, *111*(2), 395–429. doi:10.1037/0033-295X.111.2.395

Boynton, G. M. (2011). Spikes, BOLD, Attention, and Awareness: A comparison of electrophysiological and fMRI signals in V1. *Journal of Vision*, *11*(5), 12. doi:10.1167/11.5.12

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

Brass, M., Cramon, V., & Yves, D. (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., Ullsperger, M., Knoesche, T. R., Cramon, D. Y. von, & 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

Brodmann, K. (1909). *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. Leipzig : Barth. Retrieved from http://archive.org/details/b28062449

Brown, E. N., & Behrmann, M. (2017). Controversy in statistical analysis of functional magnetic resonance imaging data. *Proceedings of the National Academy of Sciences*, *114*(17), E3368–E3369. doi:10.1073/pnas.1705513114

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

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., Hazeltine, E., Scanlon, M. D., Rosen, A. C., & Gabrieli, J. D. E. (2002). Dissociable Contributions of Prefrontal and Parietal Cortices to Response Selection. *NeuroImage*, *17*(3), 1562–1571. doi:10.1006/nimg.2002.1252

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. (Eds.). (2008). *Neuroscience of Rule-Guided Behavior*. Oxford University Press, USA.

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., Gray, J. R., Conway, A. R. A., & Braver, T. S. (2011). Neural Mechanisms of Interference Control Underlie the Relationship Between Fluid Intelligence and Working Memory Span. *Journal of Experimental Psychology. General*, *140*(4), 674–692. doi:10.1037/a0024695

Burgess, P. W., Veitch, E., de Lacy Costello, A., & Shallice, T. (2000). The cognitive and neuroanatomical correlates of multitasking. *Neuropsychologia*, *38*(6), 848–863. doi:10.1016/S0028-3932(99)00134-7

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

Button, K. S. (2019). Double-dipping revisited. *Nature Neuroscience*, *22*(5), 688. doi:10.1038/s41593-019-0398-z

Canavan, A. G. M., Nixon, P. D., & Passingham, R. E. (1989). Motor learning in monkeys (Macaca fascicularis) with lesions in motor thalamus. *Experimental Brain Research*, *77*(1), 113–126. doi:10.1007/BF00250573

Carlson, T. A., Schrater, P., & He, S. (2003). Patterns of activity in the categorical representations of objects. *Journal of Cognitive Neuroscience*, *15*(5), 704–717. doi:10.1162/089892903322307429

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: The Official Journal of the Society for Neuroscience*, *26*(10), 2704–2713. doi:10.1523/JNEUROSCI.3176-05.2006

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. doi:10.3758/BF03331976

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

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, *17*(3), 455–462. doi:10.1038/nn.3635

Cole, M. W., Bagic, A., Kass, R., & Schneider, W. (2010). Prefrontal Dynamics Underlying Rapid Instructed Task Learning Reverse with Practice. *Journal of Neuroscience*, *30*(42), 14245–14254. doi:10.1523/JNEUROSCI.1662-10.2010

Cole, M. W., Etzel, J. A., Zacks, J. M., Schneider, W., & Braver, T. S. (2011). Rapid Transfer of Abstract Rules to Novel Contexts in Human Lateral Prefrontal Cortex. *Frontiers in Human Neuroscience*, *5*. doi:10.3389/fnhum.2011.00142

Cole, M. W., Ito, T., & Braver, T. S. (2016). The Behavioral Relevance of Task Information in Human Prefrontal Cortex. *Cerebral Cortex*, *26*(6), 2497–2505. doi:10.1093/cercor/bhv072

Coolican, H. (2009). *Research Methods and Statistics in Psychology*. Routledge.

Cooper, R. P., & Shallice, T. (2006). Hierarchical schemas and goals in the control of sequential behavior. *Psychological Review*, *113*(4), 887–916. doi:10.1037/0033-295X.113.4.887

Cooper, R., & Shallice, T. (2000). Contention Scheduling and the Control of Routine Activities. *Cognitive Neuropsychology*, *17*(4), 297–338. doi:10.1080/026432900380427

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. doi:10.1007/BF00994018

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. doi:10.1016/S1053-8119(03)00049-1

Cox, D. R., & Reid, N. (2000). *The Theory of the Design of Experiments*. CRC Press.

Crittenden, B. M., & Duncan, J. (2014). Task Difficulty Manipulation Reveals Multiple Demand Activity but no Frontal Lobe Hierarchy. *Cerebral Cortex*, *24*(2), 532–540. doi:10.1093/cercor/bhs333

Crone, E. A., Wendelken, C., Donohue, S. E., & Bunge, S. A. (2006). Neural Evidence for Dissociable Components of Task-switching. *Cerebral Cortex*, *16*(4), 475–486. doi:10.1093/cercor/bhi127

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, *79*(1), 1–37. doi:10.1016/S0010-0277(00)00123-2

D'Esposito, M., Deouell, L. Y., & Gazzaley, A. (2003). Alterations in the BOLD fMRI signal with ageing and disease: a challenge for neuroimaging. *Nature Reviews Neuroscience*, *4*(11), 863–872. doi:10.1038/nrn1246

D'Esposito, M., Postle, B. R., Jonides, J., & Smith, E. E. (1999a). The neural substrate and temporal dynamics of interference effects in working memory as revealed by event-related functional MRI. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(13), 7514–7519.

D'Esposito, M., Zarahn, E., Aguirre, G. K., & Rypma, B. (1999b). The Effect of Normal Aging on the Coupling of Neural Activity to the Bold Hemodynamic Response. *NeuroImage*, *10*(1), 6–14. doi:10.1006/nimg.1999.0444

Diamond, A. (1990). The Development and Neural Bases of Memory Functions as Indexed by the A-not-B and Delayed Response Tasks in Human Infants and Infant Monkeys. *Annals of the New York Academy of Sciences*, *608*(1), 267–317. doi:10.1111/j.1749-6632.1990.tb48900.x

Diamond, A., & Goldman-Rakic, P. S. (1989). Comparison of human infants and rhesus monkeys on Piaget's AB task: evidence for dependence on dorsolateral prefrontal cortex. *Experimental Brain Research*, *74*(1), 24–40. doi:10.1007/BF00248277

Dias, R., Robbins, T. W., & Roberts, A. C. (1997). Dissociable Forms of Inhibitory Control within Prefrontal Cortex with an Analog of the Wisconsin Card Sort Test: Restriction to Novel Situations and Independence from "On-Line" Processing. *The Journal of Neuroscience*, *17*(23), 9285–9297.

Dragoi, G., & Buzsáki, G. (2006). Temporal encoding of place sequences by hippocampal cell assemblies. *Neuron*, *50*(1), 145–157. doi:10.1016/j.neuron.2006.02.023

Duncan, J. (1993). Selection of input and goal in the control of behaviour. In *Attention: Selection, awareness, and control: A tribute to Donald Broadbent* (pp. 53–71). New York, NY, US: Clarendon Press/Oxford University Press.

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

Duncan, J. (2005). Frontal Lobe Function and General Intelligence: Why it Matters. *Cortex*, *41*(2), 215–217. doi:10.1016/S0010-9452(08)70896-7

Duncan, J. (2006). EPS Mid-Career Award 2004: brain mechanisms of attention. *Quarterly Journal of Experimental Psychology*, *59*(1), 2–27.

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, *14*(4), 172–179.

Duncan, J. (2013). The structure of cognition: attentional episodes in mind and brain. *Neuron*, *80*(1), 35–50. doi:10.1016/j.neuron.2013.09.015

Duncan, J., Seitz, R. J., Kolodny, J., Bor, D., Herzog, H., Ahmed, A., … Emslie, H. (2000). A Neural Basis for General Intelligence. *Science*, *289*(5478), 457–460. doi:10.1126/science.289.5478.457

Düzel, E., Cabeza, R., Picton, T. W., Yonelinas, A. P., Scheich, H., Heinze, H. J., & Tulving, E. (1999). Task-related and item-related brain processes of memory retrieval. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(4), 1794–1799.

Edelman, S. (1995). Representation, similarity, and the chorus of prototypes. *Minds and Machines*, *5*(1), 45–68. doi:10.1007/BF00974189

Edelman, S., Grill-Spector, K., Kushnir, T., & Malach, R. (1998). Toward direct visualization of the internal shape representation space by fMRI. *Psychobiology*, *26*(4), 309–321.

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.

Efron, B., & Tibshirani, R. (1995). *Cross-validation and the bootstrap: Estimating the error rate of a prediction rule*. Division of Biostatistics, Stanford University. Retrieved from https://statistics.stanford.edu/sites/default/files/BIO%20176.pdf

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 201602413. doi:10.1073/pnas.1602413113

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., … Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science (New York, N.Y.)*, *338*(6111), 1202–1205. doi:10.1126/science.1225266

Etzel, J. A., & Braver, T. S. (2013). MVPA Permutation Schemes: Permutation Testing in the Land of Cross-Validation. In *2013 International Workshop on Pattern Recognition in Neuroimaging (PRNI)* (pp. 140–143). doi:10.1109/PRNI.2013.44

Etzel, J. A., Zacks, J. M., & Braver, T. S. (2013). Searchlight analysis: Promise, pitfalls, and potential. *NeuroImage*, *78*, 261–269. doi:10.1016/j.neuroimage.2013.03.041

Fahrmeir, L., Kneib, T., & Lang, S. (2009). *Regression: Modelle, Methoden und Anwendungen* (2nd ed.). Berlin Heidelberg: Springer-Verlag. Retrieved from //www.springer.com/de/book/9783642018367

Fedorenko, E., Duncan, J., & Kanwisher, N. (2013). Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, *110*(41), 16616–16621. doi:10.1073/pnas.1315235110

Fedoroff, S., & Richardson, A. (2001). *Protocols for Neural Cell Culture*. Springer Science & Business Media.

Ferrier, D. (1876). *The Functions of the brain*. London : Smith, Elder & Co. Retrieved from http://archive.org/details/functionsofbrain1876ferr

Fisher, R. A. (1935). The design of experiments. Retrieved from http://psycnet.apa.org/psycinfo/1939-04964-000

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x

Flandin, G., & Friston, K. J. (2017). Analysis of family-wise error rates in statistical parametric mapping using random field theory. *Human Brain Mapping*, *40*(7). doi:10.1002/hbm.23839

Frank, M. J., & Badre, D. (2011). Mechanisms of Hierarchical Reinforcement Learning in Corticostriatal Circuits 1: Computational Analysis. *Cerebral Cortex*. doi:10.1093/cercor/bhr114

Freeman, J., Brouwer, G. J., Heeger, D. J., & Merriam, E. P. (2011). Orientation Decoding Depends on Maps, Not Columns. *Journal of Neuroscience*, *31*(13), 4792–4804. doi:10.1523/JNEUROSCI.5160-10.2011

Friston, K. J. (2009). Modalities, Modes, and Models in Functional Neuroimaging. *Science*, *326*(5951), 399–403. doi:10.1126/science.1174521

Friston, K. J., Frith, C. D., Liddle, P. F., & Frackowiak, R. S. J. (1991). Comparing Functional (PET) Images: The Assessment of Significant Change. *Journal of Cerebral Blood Flow & Metabolism*, *11*(4), 690–699. doi:10.1038/jcbfm.1991.122

Friston, K. J., Holmes, A. P., Poline, J. B., Grasby, P. J., Williams, S. C. R., Frackowiak, R. S., & Turner, R. (1995). Analysis of fMRI time-series revisited. *Neuroimage*, *2*(1), 45–53.

Friston, K. J., Poline, J.-B., Holmes, A. P., Frith, C. D., & Frackowiak, R. S. J. (1996). A multivariate analysis of PET activation studies. *Human Brain Mapping*, *4*(2), 140–151. doi:10.1002/(SICI)1097-0193(1996)4:2<140::AID-HBM5>3.0.CO;2-3

Frith, C. D., Friston, K., Liddle, P. F., & Frackowiak, R. S. (1991). Willed action and the prefrontal cortex in man: a study with PET. *Proceedings. Biological Sciences*, *244*(1311), 241–246. doi:10.1098/rspb.1991.0077

Fujita, I., Tanaka, K., Ito, M., & Cheng, K. (1992). Columns for visual features of objects in monkey inferotemporal cortex. *Nature*, *360*(6402), 343. doi:10.1038/360343a0

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. *Current Opinion in Neurobiology*, *37*, 66–74. doi:10.1016/j.conb.2016.01.010

Fuster, J. M. (1973). Unit activity in prefrontal cortex during delayed-response performance: neuronal correlates of transient memory. *Journal of Neurophysiology*, *36*(1), 61–78. doi:10.1152/jn.1973.36.1.61

Fuster, J. M. (1984). Behavioral electrophysiology of the prefrontal cortex. *Trends in Neurosciences*, *7*(11), 408–414. doi:10.1016/S0166-2236(84)80144-7

Fuster, J. M. (1989). *The Prefrontal Cortex: Anatomy, Physiology, and Neuropsychology of the Frontal Lobe* (2nd edition.). New York: Raven Press.

Fuster, J. M. (2001). The prefrontal cortex--an update: time is of the essence. *Neuron*, *30*(2), 319–333.

Fuster, J. M. (2015). *The Prefrontal Cortex* (5th edition.). Elsevier. doi:10.1016/C2012-0-06164-9

Fuster, J. M., & Alexander, G. E. (1970). Delayed response deficit by cryogenic depression of frontal cortex. *Brain Research*, *20*(1), 85–90. doi:10.1016/0006-8993(70)90156-3

Gaffan, D., Easton, A., & Parker, A. (2002). Interaction of Inferior Temporal Cortex with Frontal Cortex and Basal Forebrain: Double Dissociation in Strategy Implementation and Associative Learning. *Journal of Neuroscience*, *22*(16), 7288–7296. doi:10.1523/JNEUROSCI.22-16-07288.2002

Gaffan, D., & Harrison, S. (1988). Inferotemporal-frontal disconnection and fornix transection in visuomotor conditional learning by monkeys. *Behavioural Brain Research*, *31*(2), 149–163. doi:10.1016/0166-4328(88)90018-6

Gaffan, D., & Harrison, S. (1989). A comparison of the effects of fornix transection and sulcus principalis ablation upon spatial learning by monkeys. *Behavioural Brain Research*, *31*(3), 207–220. doi:10.1016/0166-4328(89)90003-X

Gail, A., & Andersen, R. A. (2006). Neural dynamics in monkey parietal reach region reflect context-specific sensorimotor transformations. *Journal of Neuroscience*, *26*(37), 9376–9384. doi:10.1523/JNEUROSCI.1570-06.2006

Garrett, D. D., Lindenberger, U., Hoge, R. D., & Gauthier, C. J. (2017). Age differences in brain signal variability are robust to multiple vascular controls. *Scientific Reports*, *7*(1), 10149. doi:10.1038/s41598-017-09752-7

Genovesio, A., Brasted, P. J., Mitz, A. R., & Wise, S. P. (2005). Prefrontal Cortex Activity Related to Abstract Response Strategies. *Neuron*, *47*(2), 307–320. doi:10.1016/j.neuron.2005.06.006

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, *62*(1), 451–482. doi:10.1146/annurev-psych-120709-145346

Gilbert, S. J., & Shallice, T. (2002). Task switching: A PDP model. *Cognitive Psychology*, *44*(3), 297–337.

Giurfa, M., Zhang, S., Jenett, A., Menzel, R., & Srinivasan, M. V. (2001). The concepts of 'sameness' and 'difference' in an insect. *Nature*, *410*(6831), 930–933. doi:10.1038/35073582

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(4), 585–595.

Goldman, P. S., & Galkin, T. W. (1978). Prenatal removal of frontal association cortex in the fetal rhesus monkey: Anatomical and functional consequences in postnatal life. *Brain Research*, *152*(3), 451–485. doi:10.1016/0006-8993(78)91103-4

Goldman, P. S., Rosvold, H. E., & Mishkin, M. (1970). Selective sparing of function following prefrontal lobectomy in infant monkeys. *Experimental Neurology*, *29*(2), 221–226. doi:10.1016/0014-4886(70)90053-1

Goldman-Rakic, P. S. (1987). Circuitry of Primate Prefrontal Cortex and Regulation of Behavior by Representational Memory. In R. Terjung (Ed.), *Comprehensive Physiology*. Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/cphy.cp010509

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–25.

Gopher, D., Armony, L., & Greenshpan, Y. (2000). Switching tasks and attention policies. *Journal of Experimental Psychology. General*, *129*(3), 308–339.

Görgen, K. (2010, June). *Neural Representations of Conditional Rules and Their Hierarchy: A Decoding fMRI Study* (Master's Thesis). BCCN/TU Berlin, Berlin.

Görgen, K., Hebart, M. N., Allefeld, C., & Haynes, J.-D. (2018). The Same Analysis Approach: Practical protection against the pitfalls of novel neuroimaging analysis methods. *NeuroImage*, *180*, 19–30. doi:10.1016/j.neuroimage.2017.12.083

Görgen, K., Hebart, M. N., & Haynes, J.-D. (2012). The Decoding Toolbox (TDT): a new fMRI analysis package for SPM and MATLAB. In *Proceedings of the Human Brain Mapping Organization OHBM 2012*. Beijing, China. Retrieved from http://f1000.com/posters/browse/summary/1092032

Grafman, J. (1995). Similarities and Distinctions among Current Models of Prefrontal Cortical Functions. *Annals of the New York Academy of Sciences*, *769*(1), 337–368. doi:10.1111/j.1749-6632.1995.tb38149.x

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, *87*, 96–110. doi:10.1016/j.neuroimage.2013.10.067

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annual Review of Neuroscience*, *37*(1), 435–456. doi:10.1146/annurev-neuro-062012-170325

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. doi:10.1126/science.1063736

Haynes, J.-D. (2015). A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. *Neuron*, *87*(2), 257–270. doi:10.1016/j.neuron.2015.05.025

Haynes, J.-D., & Rees, G. (2005a). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience*, *8*(5), 686–691. doi:10.1038/nn1445

Haynes, J.-D., & Rees, G. (2005b). Predicting the Stream of Consciousness from Activity in Human Visual Cortex. *Current Biology*, *15*(14), 1301–1307. doi:10.1016/j.cub.2005.06.026

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, *7*(7), 523–534. doi:10.1038/nrn1931

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. doi:10.1016/j.cub.2006.11.072

Hebart, M. N., & Baker, C. I. (2018). Deconstructing multivariate decoding for the study of brain function. *NeuroImage*, *180*, 4–18. doi:10.1016/j.neuroimage.2017.08.005

Hebart, M. N., Bankson, B. B., Harel, A., Baker, C. I., & Cichy, R. M. (2018). The representational dynamics of task and object processing in humans. *ELife*, *7*, e32816. doi:10.7554/eLife.32816

Hebart, M. N. [*], Görgen, K.[*], & Haynes, J.-D. (2015). The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. *Frontiers in Neuroinformatics*, *8*, 88. doi:10.3389/fninf.2014.00088

Heeger, D. J., Huk, A. C., Geisler, W. S., & Albrecht, D. G. (2000). Spikes versus BOLD: what does neuroimaging tell us about neuronal activity? *Nat Neurosci*, *3*(7), 631–633. doi:10.1038/76572

[*] Equal contribution

Heinzle, J., Wenzel, M. A., & Haynes, J.-D. (2012). Visuomotor Functional Network Topology Predicts Upcoming Tasks. *The Journal of Neuroscience*, *32*(29), 9960–9968. doi:10.1523/JNEUROSCI.1604-12.2012

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83; discussion 83-135. doi:10.1017/S0140525X0999152X

Henson, R. N. A. (2004). Analysis of fMRI time series: Linear time-invariant models, event-related fMRI and optimal experimental design. In *Human Brain Function 2nd ed.* (pp. 793–822). Elsevier, London.

Hesselmann, V., Zaro Weber, O., Wedekind, C., Krings, T., Schulte, O., Kugel, H., … Lackner, K. J. (2001). Age related signal decrease in functional magnetic resonance imaging during motor stimulation in humans. *Neuroscience Letters*, *308*(3), 141–144. doi:10.1016/S0304-3940(01)01920-6

Hinton, G. E. (1984). Distributed representations.

Honig, W. K., & Dodd, P. W. D. (1983). Delayed discriminations in the pigeon: The role of within-trial location of conditional cues. *Animal Learning & Behavior*, *11*(1), 1–9. doi:10.3758/BF03212300

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. doi:10.1152/jn.1998.80.6.3392

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. doi:10.1152/jn.2000.83.4.2355

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. *Neuron*, *76*(6), 1210–1224. doi:10.1016/j.neuron.2012.10.014

Jamalabadi, H., Alizadeh, S., Schönauer, M., Leibold, C., & Gais, S. (2016). Classification based hypothesis testing in neuroscience: Below-chance level classification rates and overlooked statistical properties of linear parametric classifiers. *Human Brain Mapping*, *37*(5), 1842–1855. doi:10.1002/hbm.23140

Jersild, A. T. (1927). *Mental set and shift*. New York: Archives of Psychology, Vol. 14, No. 89. Retrieved from http://archive.org/details/mentalsetshift00jers

Johnson, P. D., & Besselsen, D. G. (2002). Practical aspects of experimental design in animal research. *ILAR Journal*, *43*(4), 202–206. doi:10.1093/ilar.43.4.202

Kahnt, T., Grueschow, M., Speck, O., & Haynes, J.-D. (2011). Perceptual Learning and Decision-Making in Human Medial Frontal Cortex. *Neuron*, *70*(3), 549–559. doi:10.1016/j.neuron.2011.02.054

Kahnt, T., Heinzle, J., Park, S. Q., & Haynes, J.-D. (2010). The neural code of reward anticipation in human orbitofrontal cortex. *Proceedings of the National Academy of Sciences*, *107*(13), 6010–6015. doi:10.1073/pnas.0912838107

Kamitani, Y., & Sawahata, Y. (2010). Spatial smoothing hurts localization but not information: Pitfalls for brain mappers. *NeuroImage*, *49*(3), 1949–1952. doi:10.1016/j.neuroimage.2009.06.040

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679–685. doi:10.1038/nn1444

Kaplan, J. T., Man, K., & Greening, S. G. (2015). Multivariate cross-classification: applying machine learning techniques to characterize abstraction in neural representations. *Frontiers in Human Neuroscience*, *9*. doi:10.3389/fnhum.2015.00151

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355. doi:10.1038/nature06713

Kessler, D., Angstadt, M., & Sripada, C. S. (2017). Reevaluating "cluster failure" in fMRI using nonparametric control of the false discovery rate. *Proceedings of the National Academy of Sciences*, *114*(17), E3372–E3373. doi:10.1073/pnas.1614502114

Kim, C., Johnson, N. F., Cilles, S. E., & Gold, B. T. (2011). Common and Distinct Mechanisms of Cognitive Flexibility in Prefrontal Cortex. *The Journal of Neuroscience*, *31*(13), 4771–4779. doi:10.1523/JNEUROSCI.5923-10.2011

Kleinsorge, T. (2012). Task switching with a 2:1 cue-to-task mapping: separating cue disambiguation from task-rule retrieval. *Psychological Research*, *76*(3), 329–335. doi:10.1007/s00426-011-0344-5

Koechlin, E., & Jubault, T. (2006). Broca's Area and the Hierarchical Organization of Human Behavior. *Neuron*, *50*(6), 963–974. doi:10.1016/j.neuron.2006.05.017

Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science (New York, N.Y.)*, *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

Kok, P., Brouwer, G. J., Gerven, M. A. J. van, & Lange, F. P. de. (2013). Prior Expectations Bias Sensory Representations in Visual Cortex. *The Journal of Neuroscience*, *33*(41), 16275–16284. doi:10.1523/JNEUROSCI.0742-13.2013

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., Cusack, R., & Bandettini, P. (2010). How does an fMRI voxel sample the neuronal activity pattern: Compact-kernel or complex spatiotemporal filter? *NeuroImage*, *49*(3), 1965–1976. doi:10.1016/j.neuroimage.2009.09.059

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(10), 3863–3868. doi:10.1073/pnas.0600244103

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/pmc2605405/

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, *12*(5), 535–540.

Kubota, K., & Niki, H. (1971). Prefrontal cortical unit activity and delayed alternation performance in monkeys. *Journal of Neurophysiology*, *34*(3), 337–347. doi:10.1152/jn.1971.34.3.337

Lashley, K. S. (1951). The problem of serial order in behavior. In *Cerebral mechanisms in behavior; the Hixon Symposium* (pp. 112–146). Oxford, England: Wiley.

Lebiere, C., & Anderson, J. R. (1993). A connectionist implementation of the ACT-R production system. In *Proceedings of the fifteenth annual conference of the Cognitive Science Society* (pp. 635–640).

Li, S., Ostwald, D., Giese, M., & Kourtzi, Z. (2007). Flexible Coding for Categorical Decisions in the Human Brain. *The Journal of Neuroscience*, *27*(45), 12321–12330. doi:10.1523/JNEUROSCI.3795-07.2007

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*(6843), 150–157. doi:10.1038/35084005

Luria, A. R. (1966). *Higher cortical functions in man*. New York: Basic Books (Russian orig. 1962).

Luria, A. R. (1973). The frontal lobes and the regulation of behavior. In A. R. Luria & K. H. Pribram (Eds.), *Psychophysiology of the frontal lobes* (pp. 3–26). Elsevier. Retrieved from doi.org/10.1016/C2013-0-07500-7

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

Mahalanobis, P. C. (1936). On the generalized distance in statistics. National Institute of Science of India.

McIntosh, A. R., Bookstein, F. L., Haxby, J. V., & Grady, C. L. (1996). Spatial pattern analysis of functional brain images using partial least squares. *NeuroImage*, *3*(3 Pt 1), 143–157. doi:10.1006/nimg.1996.0016

Meiran, N. (1996). Reconfiguration of processing mode prior to task performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1423.

Meiran, N. (2000). Reconfiguration of stimulus task sets and response task sets during task switching. In S. Monsell & J. Driver (Eds.), *Control of cognitive processes: Attention and performance XVIII* (pp. 377–399). Cambridge, Massachusetts; London, England: MIT Press. Retrieved from https://books.google.com/books?hl=de&lr=&id=kO_baYlSVbwC&oi=fnd&pg=PA377&dq=Reconfiguration+of+Stimulus+Task+Sets+and&ots=pr3DPKdpM0&sig=S7N8ra1iibnvT-2WWEKrqFScsIY

Miller, E. K., & Cohen, J. D. (2001). An Integrative Theory of Prefrontal Cortex Funtion. *Annual Review of Neuroscience*, *24*(1), 167–202. doi:10.1146/annurev.neuro.24.1.167

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior*. New York: Holt, Rinehart and Winston, Inc. Retrieved from http://archive.org/details/plansstructureof00mill

Milner, A. D., & Goodale, M. A. (2008). Two visual systems re-viewed. *Neuropsychologia*, *46*(3), 774–785.

Milner, B. (1963). Effects of different brain lesions on card sorting: The role of the frontal lobes. *Archives of Neurology*, *9*(1), 90.

Mishkin, M., & Ungerleider, L. G. (1982). Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. *Behavioural Brain Research*, *6*(1), 57–77.

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends in Neurosciences*, *6*, 414–417.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science*, *320*(5880), 1191–1195. doi:10.1126/science.1152876

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

Momennejad, I., & Haynes, J.-D. (2013). Encoding of Prospective Tasks in the Human Prefrontal Cortex under Varying Task Loads. *The Journal of Neuroscience*, *33*(44), 17342–17349. doi:10.1523/JNEUROSCI.0492-13.2013

Monsell, S. (1996). Control of mental processes. In V. Bruce (Ed.), *Unsolved mysteries of the mind: Tutorial essays in cognition* (pp. 93–148). Psychology Press.

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140.

Monsell, S., & Driver, J. (Eds.). (2000). *Control of Cognitive Processes: Attention and Performance XVIII* (Vol. 18). Cambridge, Massachusetts; London, England: MIT Press.

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.

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

Mumford, J. A., Poline, J.-B., & Poldrack, R. A. (2015). Orthogonalization of Regressors in fMRI Models. *PLoS ONE*, *10*(4), e0126255. doi:10.1371/journal.pone.0126255

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, *59*(3), 2636–2643. doi:10.1016/j.neuroimage.2011.08.076

Mushiake, H., Saito, N., Sakamoto, K., Itoyama, Y., & Tanji, J. (2006). Activity in the lateral prefrontal cortex reflects multiple steps of future events in action plans. *Neuron*, *50*(4), 631–641.

Myers, G. J., Sandler, C., & Badgett, T. (2011). *The Art of Software Testing*. John Wiley & Sons.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *Neuroimage*, *56*(2), 400–410.

Nee, D. E., & Brown, J. W. (2012). 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

Nee, D. E., & Brown, J. W. (2013). 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., & D'Esposito, M. (2016). The hierarchical organization of the lateral prefrontal cortex. *ELife*, *5*, e12112. doi:10.7554/eLife.12112

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, *135*(3), 370–384. doi:10.2307/2344614

Nichols, T. E. (2016, July 6). Bibliometrics of Cluster Inference. Retrieved from http://blogs.warwick.ac.uk/nichols/entry/bibliometrics_of_cluster/

Niki, H., & Watanabe, M. (1979). Prefrontal and cingulate unit activity during timing behavior in the monkey. *Brain Research*, *171*(2), 213–224.

Ninokura, Y., Mushiake, H., & Tanji, J. (2003). Representation of the temporal order of visual objects in the primate lateral prefrontal cortex. *Journal of Neurophysiology*, *89*(5), 2868–2873. doi:10.1152/jn.00647.2002

Noirhomme, Q., Lesenfants, D., Gomez, F., Soddu, A., Schrouff, J., Garraux, G., … Laureys, S. (2014). Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions. *NeuroImage: Clinical*, *4*, 687–694.

Norman, D. A. (1981). Categorization of action slips. *Psychological Review*, *88*(1), 1–15. doi:10.1037/0033-295X.88.1.1

Norman, D. A., & Shallice, T. (1980). Attention to Action: Willed and Automatic Control of Behavior. Technical Report No. 8006. Univ. California, Cent. Hum. Inform. Process. Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a094713.pdf

Norman, D. A., & Shallice, T. (1986). Attention to action. In *Consciousness and self-regulation* (pp. 1–18). Springer.

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. doi:10.1016/j.tics.2006.07.005

Nyhus, E., & Barceló, F. (2009). The Wisconsin Card Sorting Test and the cognitive assessment of prefrontal executive functions: A critical update. *Brain and Cognition*, *71*(3), 437–451. doi:10.1016/j.bandc.2009.03.005

Op de Beeck, H. P. (2010a). Against hyperacuity in brain reading: Spatial smoothing does not hurt multivariate fMRI analyses? *NeuroImage*, *49*(3), 1943–1948. doi:10.1016/j.neuroimage.2009.02.047

Op de Beeck, H. P. (2010b). Probing the mysterious underpinnings of multi-voxel fMRI analyses. *NeuroImage*, *50*(2), 567–571. doi:10.1016/j.neuroimage.2009.12.072

O'Reilly, R. C. (2010). The What and How of prefrontal cortical organization. *Trends in Neurosciences*, *33*(8), 355–361. doi:10.1016/j.tins.2010.05.002

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

Owen, A. M., Evans, A. C., & Petrides, M. (1996). Evidence for a Two-Stage Model of Spatial Working Memory Processing within the Lateral Frontal Cortex: A Positron Emission Tomography Study. *Cerebral Cortex*, *6*(1), 31–38. doi:10.1093/cercor/6.1.31

Parker, A., & Gaffan, D. (1998). Memory after frontal/temporal disconnection in monkeys: conditional and non-conditional tasks, unilateral and bilateral frontal lesions. *Neuropsychologia*, *36*(3), 259–271. doi:10.1016/S0028-3932(97)00112-7

Passingham, R. E. (1993). *The frontal lobes and voluntary action*. New York, NY, US: Oxford University Press.

Pavlov, P. I. (1927). Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex, Lecture I [eng. transl.]. *Annals of Neurosciences, Annals of Neurosciences, July 2010*, *17, 17*(3, 3), 136, 136–141. doi:10.5214/ans.0972-7531.1017309, 10.5214/ans.0972-7531.1017309

Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., & Nichols, T. E. (2011). *Statistical Parametric Mapping: The Analysis of Functional Brain Images: The Analysis of Functional Brain Images*. Academic Press.

Petrides, M. (1982). Motor conditional associative-learning after selective prefrontal lesions in the monkey. *Behavioural Brain Research*, *5*(4), 407–413.

Petrides, M. (1985a). Deficits in non-spatial conditional associative learning after periarcuate lesions in the monkey. *Behavioural Brain Research*, *16*(2–3), 95–101.

Petrides, M. (1985b). Deficits on conditional associative-learning tasks after frontal-and temporal-lobe lesions in man. *Neuropsychologia*, *23*(5), 601–614.

Petrides, M. (1990). Nonspatial conditional learning impaired in patients with unilateral frontal but not unilateral temporal lobe excisions. *Neuropsychologia*, *28*(2), 137–149.

Petrides, M. (1991a). Functional specialization within the dorsolateral frontal cortex for serial order memory. *Proceedings of the Royal Society of London B: Biological Sciences*, *246*(1317), 299–306.

Petrides, M. (1991b). Monitoring of selections of visual stimuli and the primate frontal cortex. *Proceedings of the Royal Society of London B: Biological Sciences*, *246*(1317), 293–298.

Petrides, M. (1996). Specialized Systems for the Processing of Mnemonic Information within the Primate Frontal Cortex [including a short discussion with A. Baddeley]. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *351*(1346), 1455–1462. doi:10.1098/rstb.1996.0130

Petrides, M. (1997). Visuo-motor conditional associative learning after frontal and temporal lesions in the human brain. *Neuropsychologia*, *35*(7), 989–997.

Petrides, M. (2000). Dissociable Roles of Mid-Dorsolateral Prefrontal and Anterior Inferotemporal Cortex in Visual Working Memory. *Journal of Neuroscience*, *20*(19), 7496–7503. doi:10.1523/JNEUROSCI.20-19-07496.2000

Petrides, M. (2005). Lateral prefrontal cortex: architectonic and functional organization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 781–795. doi:10.1098/rstb.2005.1631

Petrides, M., & Milner, B. (1982). Deficits on subject-ordered tasks after frontal- and temporal-lobe lesions in man. *Neuropsychologia*, *20*(3), 249–262.

Petrides, M., & Pandya, D. N. (1984). Projections to the frontal cortex from the posterior parietal region in the rhesus monkey. *Journal of Comparative Neurology*, *228*(1), 105–116.

Petrides, M., & Pandya, D. N. (1999). Dorsolateral prefrontal cortex: comparative cytoarchitectonic analysis in the human and the macaque brain and corticocortical connection patterns. *The European Journal of Neuroscience*, *11*(3), 1011–1036.

Petrides, M., & Pandya, D. N. (2002). Comparative cytoarchitectonic analysis of the human and the macaque ventrolateral prefrontal cortex and corticocortical connection patterns in the monkey. *European Journal of Neuroscience*, *16*(2), 291–310.

Pischedda, D. [*], Görgen, K.[*], Haynes, J.-D., & Reverberi, C. (2017). Neural Representations of Hierarchical Rule Sets: The Human Control System Represents Rules Irrespective of the Hierarchical Level to Which They Belong. *The Journal of Neuroscience*, *37*(50), 12281–12296. doi:10.1523/jneurosci.3088-16.2017

Polyn, S. M., Natu, V. S., Cohen, J. D., & Norman, K. A. (2005). Category-Specific Cortical Activity Precedes Retrieval During Memory Search. *Science*, *310*(5756), 1963–1966. doi:10.1126/science.1117645

Postle, B. R., Berger, J. S., & D'Esposito, M. (1999). Functional neuroanatomical double dissociation of mnemonic and executive control processes contributing to working memory performance. *Proceedings of the National Academy of Sciences*, *96*(22), 12959–12964. doi:10.1073/pnas.96.22.12959

Privman, E., Nir, Y., Kramer, U., Kipervasser, S., Andelman, F., Neufeld, M. Y., … Malach, R. (2007). Enhanced Category Tuning Revealed by Intracranial Electroencephalograms in High-Order Human Visual Areas. *J. Neurosci.*, *27*(23), 6234–6242.

Quintana, J., Yajeya, J., & Fuster, J. M. (1988). Prefrontal representation of stimulus attributes during delay tasks. I. Unit activity in cross-temporal integration of sensory and sensory-motor information. *Brain Research*, *474*(2), 211–221. doi:10.1016/0006-8993(88)90436-2

Rao, S. C., Rainer, G., & Miller, E. K. (1997). Integration of What and Where in the Primate Prefrontal Cortex. *Science*, *276*(5313), 821–824. doi:10.1126/science.276.5313.821

Reason, J. (1991). *Human Error*. Cambridge England ; New York: Cambridge University Press.

Reason, J. T. (1979). Actions not as planned: The price of automatization.

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., … 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

[*] Equal contribution

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

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–e30284.

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, *497*(7451), 585–590. doi:10.1038/nature12160

Rigotti, M., Rubin, D. B. D., Wang, X.-J., & Fusi, S. (2010). Internal Representation of Task Rules by Recurrent Dynamics: The Importance of the Diversity of Neural Responses. *Frontiers in Computational Neuroscience*, *4*. doi:10.3389/fncom.2010.00024

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, *124*(2), 207.

Rosenblatt, J., Gilron, R., & Mukamel, R. (2016). Better-Than-Chance Classification for Signal Detection. *ArXiv:1608.08873 [Stat]*. Retrieved from http://arxiv.org/abs/1608.08873

Rosenthal, R. (1963). On the social psychology of the psychological experiment: The experimenter's hypothesis as unintended determinant of experimental results. *American Scientist*, *51*(2), 268–283.

Rosenthal, R. (1966). Experimenter effects in behavioral research. Retrieved from http://doi.apa.org/psycinfo/1967-09647-000

Rosenthal, R. (2009). Interpersonal expectations: Effects of the experimenter's hypothesis. *Artifacts in Behavioral Research*, 138–210.

Ross, M. H., Yurgelun-Todd, D. A., Renshaw, P. F., Maas, L. C., Mendelson, J. H., Mello, N. K., … Levin, J. M. (1997). Age-related reduction in functional MRI response to photic stimulation. *Neurology*, *48*(1), 173–176. doi:10.1212/wnl.48.1.173

Rowe, J. B., Sakai, K., Lund, T. E., Ramsøy, 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

Rowe, J. B., Toni, I., Josephs, O., Frackowiak, R. S. J., & Passingham, R. E. (2000). The Prefrontal Cortex: Response Selection or Maintenance Within Working Memory? *Science*, *288*(5471), 1656–1660. doi:10.1126/science.288.5471.1656

Ruge, H., & Wolfensteller, U. (2010). Rapid Formation of Pragmatic Rule Representations in the Human Brain during Instruction-Based Learning. *Cerebral Cortex*, *20*(7), 1656–1667. doi:10.1093/cercor/bhp228

Rumelhart, D. E., & Norman, D. A. (1982). Simulating a Skilled Typist: A Study of Skilled Cognitive-Motor Performance. *Cognitive Science*, *6*(1), 1–36. doi:10.1207/s15516709cog0601_1

Rushworth, M. F. S., Passingham, R. E., & Nobre, A. C. (2002). Components of Switching Intentional Set. *Journal of Cognitive Neuroscience*, *14*(8), 1139–1150. doi:10.1162/089892902760807159

Rushworth, M. F. S., Passingham, R. E., & Nobre, A. C. (2005). Components of attentional set-switching. *Experimental Psychology*, *52*(2), 83–98. doi:10.1027/1618-3169.52.2.83

Rusznak, C., & Davies, R. J. (1998). Diagnosing allergy. *BMJ : British Medical Journal*, *316*(7132), 686.

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

* Equal contribution

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

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

Schenk, T., Franz, V., & Bruno, N. (2011). Vision-for-perception and vision-for-action: Which model is compatible with the available psychophysical and neuropsychological data? *Vision Research*, *51*(8), 812–818. doi:10.1016/j.visres.2011.02.003

Schreiber, K., & Krekelberg, B. (2013). The Statistical Analysis of Multi-Voxel Patterns in Functional Imaging. *PLoS ONE*, *8*(7), e69328. doi:10.1371/journal.pone.0069328

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

Schumacher, E. H., Elston, P. A., & D'Esposito, M. (2003). Neural evidence for representation-specific response selection. *Journal of Cognitive Neuroscience*, *15*(8), 1111–1121. doi:10.1162/089892903322598085

Schumacher, F. K., Schumacher, L. V., Schelter, B. O., & Kaller, C. P. (2018). Functionally dissociating ventro-dorsal components within the rostro-caudal hierarchical organization of the human prefrontal cortex. *NeuroImage*. doi:10.1016/j.neuroimage.2018.10.048

Shaffer, L. H. (1978). Timing in the Motor Programming of Typing. *Quarterly Journal of Experimental Psychology*, *30*(2), 333–345. doi:10.1080/14640747808400680

Shallice, T. I. M., & Burgess, P. W. (1991). Deficits in strategy application following frontal lobe damage in man. *Brain*, *114*(2), 727–741.

Shepard, R. N. (1980). Multidimensional Scaling, Tree-Fitting, and Clustering. *Science*, *210*(4468), 390–398. doi:10.1126/science.210.4468.390

Sherrington, C. S. (1906). *The integrative action of the nervous system*. New York, C Scribner's sons. Retrieved from http://archive.org/details/integrativeacti02shergoog

Siegel, M., Warden, M. R., & Miller, E. K. (2009). Phase-dependent neuronal coding of objects in short-term memory. *Proceedings of the National Academy of Sciences*, *106*(50), 21341–21346. doi:10.1073/pnas.0908193106

Sigala, N., Kusunoki, M., Nimmo-Smith, I., Gaffan, D., & Duncan, J. (2008). Hierarchical coding for sequential task events in the monkey prefrontal cortex. *Proceedings of the National Academy of Sciences*, *105*(33), 11969–11974. doi:10.1073/pnas.0802569105

Soch, J., Haynes, J.-D., & Allefeld, C. (2016). How to avoid mismodelling in GLM-based fMRI data analysis: cross-validated Bayesian model selection. *NeuroImage*, *141*, 469–489. doi:10.1016/j.neuroimage.2016.07.047

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

Sprague, T. C., Ester, E. F., & Serences, J. T. (2014). Reconstructions of Information in Visual Spatial Working Memory Degrade with Memory Load. *Current Biology*, *24*(18), 2174–2180. doi:10.1016/j.cub.2014.07.066

Stoet, G., & Snyder, L. H. (2003). Executive control and task-switching in monkeys. *Neuropsychologia*, *41*(10), 1357–1364.

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

Stokes, M., Thompson, R., Cusack, R., & Duncan, J. (2009). Top-Down Activation of Shape-Specific Population Codes in Visual Cortex during Mental Imagery. *Journal of Neuroscience*, *29*(5), 1565–1572. doi:10.1523/JNEUROSCI.4657-08.2009

Student. (1908). The probable error of a mean. *Biometrika*, 1–25.

Stuss, D. T., & Benson, D. F. (1986). *The frontal lobes*. Raven Press.

Sudevan, P., & Taylor, D. A. (1987). The cuing and priming of cognitive operations. *Journal of Experimental Psychology. Human Perception and Performance*, *13*(1), 89–103.

Tanaka, K. (1992). Inferotemporal cortex and higher visual functions. *Current Opinion in Neurobiology*, *2*(4), 502–505. doi:10.1016/0959-4388(92)90187-P

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, *19*, 109–139. doi:10.1146/annurev.ne.19.030196.000545

Tanaka, K., Saito, H., Fukada, Y., & Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology*, *66*(1), 170–189. doi:10.1152/jn.1991.66.1.170

Taylor, J. E., & Worsley, K. J. (2007). Detecting Sparse Signals in Random Fields, With an Application to Brain Mapping. *Journal of the American Statistical Association*, *102*(479), 913–928. doi:10.1198/016214507000000815

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J.-B., Lebihan, D., & Dehaene, S. (2006). Inverse retinotopy: Inferring the visual content of images from brain activation patterns. *NeuroImage*, *33*(4), 1104–1116. doi:10.1016/j.neuroimage.2006.06.062

Thorpe, S. J., Rolls, E. T., & Maddison, S. (1983). The orbitofrontal cortex: Neuronal activity in the behaving monkey. *Experimental Brain Research*, *49*(1), 93–115. doi:10.1007/BF00235545

Todd, M. T., Nystrom, L. E., & Cohen, J. D. (2013). Confounds in multivariate pattern analysis: theory and rule representation case study. *Neuroimage*, *77*, 157–165.

Tong, F., & Pratte, M. S. (2012). Decoding Patterns of Human Brain Activity. *Annual Review of Psychology*, *63*(1), 483–509. doi:10.1146/annurev-psych-120710-100412

Toni, I., & Passingham, R. E. (1999). Prefrontal-basal ganglia pathways are involved in the learning of arbitrary visuomotor associations: a PET study. *Experimental Brain Research*, *127*(1), 19–32.

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, *17*(4), 401–419. doi:10.1007/BF02288916

Tusche, A., Bode, S., & Haynes, J.-D. (2010). Neural Responses to Unattended Products Predict Later Consumer Choices. *The Journal of Neuroscience*, *30*(23), 8024–8031. doi:10.1523/JNEUROSCI.0064-10.2010

Tusche, A., Kahnt, T., Wisniewski, D., & Haynes, J.-D. (2013). Automatic processing of political preferences in the human brain. *NeuroImage*, *72*, 174–182. doi:10.1016/j.neuroimage.2013.01.020

von Kries, J. (1895). Über die Natur gewisser mit den Gehirnzustände. *Zeitschrift Für Psychologie Und Physiologie Der Sinnesorgane*, *8*, 1–33.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspectives on Psychological Science*, *4*(3), 274–290. doi:10.1111/j.1745-6924.2009.01125.x

Wallis, J. D. (2008). Single Neuron Activity Underlying Behavior-Guiding Rules. In S. A. Bunge & J. D. Wallis (Eds.), *Neuroscience of Rule-Guided Behavior*. Oxford University Press.

Wallis, J. D., Anderson, K. C., & Miller, E. K. (2001). Single neurons in prefrontal cortex encode abstract rules. *Nature*, *411*(6840), 953–956.

Wallis, J. D., & Miller, E. K. (2003a). From rule to response: neuronal processes in the premotor and prefrontal cortex. *Journal of Neurophysiology*, *90*(3), 1790–1806.

Wallis, J. D., & Miller, E. K. (2003b). Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. *European Journal of Neuroscience*, *18*(7), 2069–2081.

Warden, M. R., & Miller, E. K. (2007). The representation of multiple objects in prefrontal neuronal delay activity. *Cerebral Cortex*, *17*(suppl 1), i41–i50.

Warden, M. R., & Miller, E. K. (2010). Task-dependent changes in short-term memory in the prefrontal cortex. *The Journal of Neuroscience*, *30*(47), 15801–15810.

Weichwald, S., Meyer, T., Özdenizci, O., Schölkopf, B., Ball, T., & Grosse-Wentrup, M. (2015). Causal interpretation rules for encoding and decoding models in neuroimaging. *NeuroImage*. doi:10.1016/j.neuroimage.2015.01.036

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

Wisniewski, D., Reverberi, C., Tusche, A., & Haynes, J.-D. (2014). The Neural Representation of Voluntary Task-Set Selection in Dynamic Environments. *Cerebral Cortex (New York, N.Y.: 1991)*. doi:10.1093/cercor/bhu155

Womelsdorf, T., & Valiante, T. A. (2014). Dynamic circuit motifs underlying rhythmic gain control, gating and integration. *Nature Neuroscience*, *17*(8), 1031–9. doi:10.1038/nn.3764

Wood, J. N., & Grafman, J. (2003). Human prefrontal cortex: processing and representational perspectives. *Nature Reviews Neuroscience*, *4*(2), 139–147. doi:10.1038/nrn1033

Woolgar, A., Golland, P., & Bode, S. (2014). Coping with confounds in multivoxel pattern analysis: What should we do about reaction time differences? A comment on Todd, Nystrom & Cohen 2013. *NeuroImage*, *98*, 506–512. doi:10.1016/j.neuroimage.2014.04.059

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

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

Wylie, G., & Allport, A. (2000). Task switching and the measurement of "switch costs." *Psychological Research*, *63*(3), 212–233. doi:10.1007/s004269900003

Yajeya, J., Quintana, J., & Fuster, J. M. (1988). Prefrontal representation of stimulus attributes during delay tasks. II. The role of behavioral significance. *Brain Research*, *474*(2), 222–230.

Yin, H. H. (2009). The role of the murine motor cortex in action duration and order. *Frontiers in Integrative Neuroscience*, *3*. doi:10.3389/neuro.07.023.2009

Yin, H. H. (2010). The sensorimotor striatum is necessary for serial order learning. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *30*(44), 14719–14723. doi:10.1523/JNEUROSCI.3989-10.2010

Appendix A Selbstständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

_____

Datum, Unterschrift

Appendix B Full Publication Record

Journal Publications
(*=equal contribution)

Görgen K, Hebart M N, Allefeld C, Haynes J-D (2018). The same analysis approach: Practical protection against the pitfalls of novel neuroimaging analysis methods. *NeuroImage*. doi:10.1016/j.neuroimage.2017.12.083

Pischedda D*, Görgen K*, Haynes J-D, Reverberi C (2017). Neural Representations of Hierarchical Rule Sets: The Human Control System Represents Rules Irrespective of the Hierarchical Level to Which They Belong. *Journal of Neuroscience, 37*(50), 12281–12296. doi:10.1523/jneurosci.3088-16.2017

Allefeld C, Görgen K, Haynes J-D (2016). Valid population inference for information-based imaging: From the second-level t-test to prevalence inference. *NeuroImage, 141*, 378–392. doi:10.1016/j.neuroimage.2016.07.040

Schultze-Kraft M, Birman D, Rusconi M, Allefeld C, Dähne S, Görgen K, Blankertz B, Haynes J-D (2015). Controlling Predictive Brain Signals. *PNAS,* doi: 10.1073/pnas.1513569112

Hebart M N*, Görgen K*, & Haynes J-D (2015). The Decoding Toolbox (TDT): A versatile software package for multivariate analyses of functional imaging data. *Frontiers in Neuroinformatics, 8*(88). doi:10.3389/fninf.2014.00088

Haufe S, Meinecke F, Görgen K, Dähne S, Haynes J-D, Blankertz B, & Bießmann F (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage 87*: 96–110.

This paper has been selected as "NeuroImage Editors' Choice Award Winner" 2014

Reverberi C*, Görgen K*, & Haynes J-D (2012). Distributed representations of rule identity and rule order in human frontal cortex and striatum. *Journal of Neuroscience*, *32*(48), 17420–17430. doi:10.1523/JNEUROSCI.2344-12.2012

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

Conference Talks & Invited Talks

Pischedda D*, Görgen K*, Haynes J-D, Reverberi C (2017). Neural Representations of Hierarchical Rule Sets: the Human Control System Represents Rules Irrespective of Their Hierarchical Level. *CNEF conference 2017, 28th-30th September 2017, Padova, Italy*

Pischedda D, Seyed-Allaei S, Görgen K, Haynes J-D, Reverberi C (2017). Who does what? Neural Representation of One's Own Subtask, a Partner's Subtask, and of Subtask Assignment. *CNEF conference 2017, 28th-30th September 2017, Padova, Italy*

Görgen K (2017). The Same Analysis Approach: Detect, Avoid & Eliminate Confounds in Neuroimaging and other Data Analysis. *CIMeC - Centro Interdipartimentale Mente/Cervello, 11 July 2017, Rovereto/Trento, Italy*

Görgen K (2017). The Same Analysis Approach: Detect, Avoid & Eliminate Confounds in Neuroimaging and other Data Analysis. *IKW Colloquium, 31 May 2017, Universität Osnabrück, Germany*

Görgen K (2016). The Same Analysis Approach: A Unit-Testing-Like Framework to Detect, Avoid & Eliminate Design & Analysis Confounds in MVPA and Other Experimental Studies. *Intel/Princeton Meeting, Princeton Neuroscience Institute, Princeton, 28 Oct 2016, Princeton University, NJ, USA*

Görgen K (2016). Detect, Avoid & Eliminate Confounds in MVPA. *NIHM, National Institute for Mental Health, 24 Oct 2016, Bethesda, MD, USA*

Görgen K (2016). How to detect, avoid & eliminate confounds in MVPA, neuroimaging, and other experimental studies using the same analysis approach (SAA), illustrated on X>=7 reasons for so-far ominous below chance accuracies. *CCNB Seminar Series, Center for Cognitive Neuroscience, 13 June 2016, FU Berlin, Germany*

Görgen K (2016). A Coxi's Back from the Future. *Alumni Conference Cognitive Science, 13-14 May 2016, Osnabrück, Germany*

Görgen K (2015). Detect, Avoid & Eliminate Confounds in MVPA. *Workshop: Pattern Recognition in Neuroimaging, 1 Sep 2015, FU Berlin, Germany*

Görgen K (2015). The Decoding Toolbox. *Workshop: Pattern Recognition in Neuroimaging, 1 Sep 2015, FU Berlin, Germany*

Görgen K (2015). Dual-BCI, Hyperscanning, and other Fancy Buzzwords. But: What can Multi-Person Neuroscience Really Tell Us? *Invited Talk at the UKE, Hamburg, Germany*

Görgen K (2015). Dual-BCI, Hyperscanning und andere crazy Buzzwords. Nur: Was können wir durch Multi-Personen Neuroscience wirklich lernen? *Talk at Forschungskolloquium "Neuro-Wissen schaffen", Studienstiftung des Deutschen Volkes, 10-12 July 2015, Hamburg*

Schultze-Kraft M, Birman D, Rusconi M, Allefeld C, Dähne S, Görgen K, Blankertz B, Haynes J-D (2015). Interrupting Movement Intentions with a Closed Loop BCI. *Conference Talk (2358) at the Human Brain Mapping Organization OHBM, Honolulu, Hawaii.* This Talk has been selected as a conference highlight.

Görgen K (2014). Hyperscanning, Brain-to-Brain-Coupling, and Dual-BCI. But: What can Multi-Person Neuroscience Really Tell Us? *Invited Talk at the Donders Discussion Conference 2014, Nijmegen, The Netherlands*

Haufe S, Meinecke F, Görgen K, Dähne S, Haynes J-D, Blankertz B, & Bießmann F (2014). Parameter interpretation, regularization and source localization in multivariate linear models, *Talk at PRNI 2014 Workshop, Tübingen, Germany*

Görgen K (2013). Rainbow Course RC: Hyperscanning, Brain-to-Brain Coupling, and other Fancy Buzzwords. But: What can Multi-Person Neuroscience Really Tell Us? *Selected talk at the IK2013 Meeting, Günne, Germany*

Görgen K (2012). Exploring Brain-to-Brain-Connectivity – How Hyperscanning can Help us Advance Communication Neuroscience, *Invited Talk at the Donders Discussion Conference 2012, Nijmegen, The Netherlands*

Görgen K (2012). Rainbow Course RC1: Multivariate decoding of neural data: Introduction and hands-on, *Selected talk at the IK2012 Meeting, Günne, Germany*

Further Contributions (Conference Posters & arXiv)

Görgen K, Hebart MN, Allefeld C, Haynes J-D (2017). The Same Analysis Approach: Practical protection against the pitfalls of novel neuroimaging analysis methods. arXiv:1703.06670 [Q-Bio, Stat]. http://arxiv.org/abs/1703.06670 *(published in NeuroImage)*

Pischedda D, Seyed-Allaei S, Görgen K, Haynes J-D, Reverberi CF (2016). Who does what? Neural representations of identity and ownership of one's own and a partner's subtasks. *Poster 362 at the Society for Neuroscience 2016, November 12-16, San Diego, CA, USA*

Allefeld C, Görgen K, Haynes J-D (2016). Valid population inference for information-based imaging: Information prevalence inference. http://arxiv.org/abs/1512.00810 (follow-up published in *NeuroImage*).

Görgen K, Hebart MN, Allefeld C, Haynes J-D (2016). Detecting, Avoiding and Eliminating Confounds in Neuroimaging Data Analysis: Design–Analysis Interactions and the Same Analysis Approach. *Poster W100 at the BCCN Conference 2016, 21-23 September 2016, Berlin, Germany.* doi: 10.12751/nncn.bc2016.0118

Allefeld C, Görgen K, Haynes J-D (2016). Valid population inference for information-based imaging: Information prevalence inference. *Poster W90 at the BCCN Conference 2016, 21-23 September 2016, Berlin, Germany*. doi: 10.12751/nncn.bc2016.0108

Oroz Arigas S, Strang S, Swaboda N, Wiers C, Görgen K, Haynes J-D, Park SQ (2016). Modulation of the automatic approach bias towards high caloric food (preliminary results). *Poster at the Society for Neuroeconomics Annual Meeting, August 28-30, 201, Berlin, Germany*

Paolo C, Baggio G, Pischedda D, Görgen K, Blumenthal A, Haynes J-D, Reverberi C (2015). Concept combination with logical connectives. *Proceedings of the Cognitive Neuroscience Society Annual Meeting 2015, San Francisco, USA*

Pischedda D*, Görgen K*, Haynes J-D, Reverberi C (2014), Neural Representation of Rules at Different Hierarchical Levels. *FENS Conference, July 5-9, Milan, Italy*

Görgen K, Hebart MN, Allefeld C, Haynes J-D (2014), Detecting, Avoiding & Eliminating Confounds in MVPA / Decoding Studies. *Abstract 874, Poster 3436 (WTh), Conference Proceedings of the Human Brain Mapping Organization OHBM, Hamburg, Germany*

Haufe S, Meinecke F, Görgen K, Dähne S, Haynes J-D, Blankertz B, & Bießmann F (2014). Interpreting weight vectors of multivariate linear models in neuroimaging. *Abstract 437, Poster 3502 (WTh), Conference Proceedings of the Human Brain Mapping Organization OHBM, Hamburg, Germany*

Pischedda D*, Görgen K*, Haynes J-D, Reverberi C (2013). Neural representation of rules at different hierarchical levels. Program No. 573.12. 2013 *Neuroscience Meeting Planner. San Diego, CA: Society for Neuroscience*, *2013*.

Görgen K, Schultze-Kraft R, Haynes J-D, Blankertz B (2013). Cooperating Brains: Dual-BCI as a New Paradigm to Investigate Brain-to-Brain Coordination. *Conference Proceedings of the Joint Action Meeting JAM 2013, Berlin, Germany*

Schultze-Kraft R*, Görgen K*, Wenzel M, Haynes J-D, Blankertz B (2013). Cooperating Brains: Joint Control of a Dual-BCI. *Conference Proceedings of the Joint Action Meeting JAM 2013, Berlin, Germany*

Görgen K, Schultze-Kraft R, Haynes J-D, Blankertz B (2013). Cooperating Brains: Dual-BCI as a New Paradigm to Investigate Brain-to-Brain Coordination. *Conference Proceedings of the BCI Meeting 2013, Asilomar, USA*

Görgen K*, Schultze-Kraft R*, Wenzel M, Haynes J-D, Blankertz B (2013). Cooperating Brains: Joint Control of a Dual-BCI. *Conference Proceedings of BCI Meeting 2013, Asilomar, USA*

Görgen K, Schultze-Kraft R, Haynes J-D, Blankertz B. Cooperating Brains: Dual-BCI as a New Paradigm to Investigate Brain-to-Brain Coordination. *Poster presentation at the Workshop "Cooperation: Why, How, and With Whom?", Aarhus, Denmark*

Görgen K, Schultze-Kraft R, Haynes J-D, Blankertz B (2013). Cooperating Brains: Dual-BCI as a New Paradigm to Investigate Brain-to-Brain Coordination. *Conference Proceedings of the IK2013 Meeting, Günne, Germany*

Schultze-Kraft R*, Görgen K*, Haynes J-D, Blankertz B (2013). Cooperating Brains: Joint Control of a Dual-BCI. *Conference Proceedings of the IK2013 Meeting, Günne, Germany*

Görgen K*, Hebart M N*, Haynes J-D (2012). The Decoding Toolbox (TDT): An easy-to-use decoding package for fMRI data. *Conference Proceedings of the Human Brain Mapping Organization OHBM, Beijing, China.* http://f1000.com/posters/browse/summary/1092032

Görgen K*, Hebart M N*, Haynes J-D (2012). The Decoding Toolbox (TDT): An easy-to-use decoding package for fMRI data. *Conference Proceedings of the IK2012 Meeting, Günne, Germany.*

Görgen K, Reverberi C, Haynes J-D (2011). Two Double-Dissociations of Neural Rule Representations: Where- vs. What-Rules & Identity vs. Priority. *Poster Abstract, Bernstein symposium "Bayesian Inference: From Spikes to Behaviour", Tübingen, Germany, Dec 8-10, 2011*

Görgen K, Reverberi C, Haynes J-D (2011). Two Double-Dissociations in How the Brain Encodes Rules – Different Regions Encode (1) Rule Identity vs Rule Order, and (2) Rules Requiring Where- vs What-Responses. *Front. Comput. Neurosci. Conference Abstract: BC11.* doi: 10.3389/conf.fncom.2011.53.00229

Görgen K, Reverberi C, Haynes J-D (2011). Where- vs. What-Rules & Which Rule When: Two Double-Dissociations of Neural Representations of Hierarchical Rules. *Poster presentation at the OCCAM2011 conference, Osnabrück, Germany, June 22-24, 2011.*

Görgen K, Reverberi C, Haynes J-D (2011). Decoding Neural Representations of Rules and Rule Order. *Conference Proceedings of the IK2011 Meeting, Günne, Germany*

Görgen K, Reverberi C, Haynes J-D (2011). Decoding Neural Representations of Rules and Rule Order. *Poster Abstract, ABIM (Alpine Brain Imaging Meeting), Champéry, Switzerland, Jan 9-14, 2011.*

Görgen K, Reverberi C, Haynes J-D (2010). Decoding Neural Representations of Rules and Rule Order. *Poster Abstract, Bernstein Conference on Computational Neuroscience Conference. Berlin, Germany, Sep 27 – Oct 1, 2010.*

Görgen K (2010). Neural Representations of Conditional Rules and Their Hierarchy: A Decoding fMRI Study. Master's Thesis for the Masters Programme Computational Neuroscience, BCCN Berlin, Germany. Supervised by: Reverberi, C. Haynes, J-D.

Görgen K, Reverberi C, Haynes J-D (2010). Decoding Neural Representations of Conditional Rules from fMRI Data. *Poster Presentation, Conference IK 2010, Günne, Germany*

Görgen K, Bosman C, Womelsdorf T, Oostenveld R, Fries P (2009). Improved (I)CA-noise elimination of electrophysiological data using band-pass filtered components. *Frontiers in Computational Neuroscience. Conference Abstract: Computational and systems neuroscience*. doi: 10.3389/conf.neuro.10.2009.03.257

Görgen K (2007). Combining Eyetracking and EEG. Bachelor's Thesis for the BSc. Programme Cognitive Science, University of Osnabrück, Germany. Supervised by: Acik A, König P.

Acik A, Hipp J, Görgen K, König P, Engel AK (2007). Simultaneously EEG Recording and Eyetracking during Active Viewing. *Journal of Eye Movement Research. ECEM2007 Special,* Vol. I, 1, pp. 120.

Appendix C Original Publications

Study 1

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

Only the print version of this thesis contains the manuscript of Study 1. Copyright restrictions by the publisher prohibit to include it in the electronic version.

Study 2

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

*: These authors contributed equally to the manuscript

Only the print version of this thesis contains the manuscript of Study 2. Difficulties with potentially incompatible copyright restrictions prohibit to include it in the electronic version.

Study 3

Pischedda, D.*, Görgen, K.*, Haynes, J.-D., & Reverberi, C. (2017). Neural Representations of Hierarchical Rule Sets: The Human Control System Represents Rules Irrespective of the Hierarchical Level to Which They Belong. *The Journal of Neuroscience*, *37*(50), 12281–12296. doi:10.1523/jneurosci.3088-16.2017

*: These authors contributed equally to the manuscript

Only the print version of this thesis contains the manuscript of Study 3. Difficulties with potentially incompatible copyright restrictions prohibit to include it in the electronic version.

Study 4

Görgen, K., Hebart, M. N., Allefeld, C.*, & Haynes, J.-D.* (2018). The Same Analysis Approach: Practical protection against the pitfalls of novel neuroimaging analysis methods. *NeuroImage*, *180*, 19–30. doi:10.1016/j.neuroimage.2017.12.083

*: These authors contributed equally to the manuscript

Only the print version of this thesis contains the manuscript of Study 4. Difficulties with potentially incompatible copyright restrictions prohibit to include it in the electronic version.

Cite as: Görgen, K. (2019). *On Rules and Methods: Neural Representations of Complex Rule Sets and Related Methodological Contributions* (Doctoral Dissertation, Humboldt-Universität zu Berlin, Berlin, Germany). https://doi.org/10.18452/20711