Neural mechanisms for flexible behaviour in humans and artificial neural networks

Emilia Paulina Piwek

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Experimental Psychology at the University of Oxford.



New College, Oxford

Hilary Term, 2023

Neural mechanisms for flexible behaviour in humans and artificial neural networks

Emilia Paulina Piwek

New College, Oxford

Hilary Term, 2023

Thesis abstract

The natural environment is non-stationary – whilst our daily life often follows a routine, circumstances can change, and no two experiences are the same. To thrive, intelligent agents must therefore be able to flexibly adjust the way they process information to support the pursuit of their goals. In this thesis, I explored different aspects of such flexible, intelligent behaviour. To this end, I used a combination of artificial neural network modelling and human behavioural methods, comparing network activity patterns and behaviour against their biological counterparts. In the first set of experiments, I examined how artificial agents manipulate information in their short-term memory to match changing task demands. I demonstrated that neural networks used two distinct representational formats to prevent cross-interference and support the generalisation of information across contexts. Crucially, these representations were a striking match to those previously observed in biological brains. I then built on this result by analysing aspects of the neural network model not readily accessible to biological researchers, revealing the mechanistic and normative reasons for the observed results. In the second experiment, I explored how information can be prioritised in short term memory in situations where it can prove to be irrelevant for guiding behaviour at a later stage. I showed that artificial agents achieved this goal by amplifying the proportion of memory resources dedicated to prioritised information maintenance, without transforming it into a format that could guide subsequent behaviour. In the third experiment, I examined how logical categories can be learnt from an algorithmic perspective. I showed that a behavioural marker previously thought to index rule-based processes involving symbolic representations in humans was equally predictive of the behaviour of artificial agents not equipped with such features. In the last experiment, I investigated how humans learn simple problems and showed that they might have an innate predisposition to search for abstract rules, allowing them to generalise their knowledge to new contexts. Taken together, it is hoped that the work presented in this thesis furthers our understanding of the principles behind flexible, intelligent behaviour and sheds new light on their mechanistic implementation in neural circuits.

This thesis was generously funded by an ESRC DTP Studentship

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisors, Mark Stokes and Christopher Summerfield for their endless inspiration, encouragement, and support throughout my DPhil journey. Their kindness and patience were invaluable in helping me to develop as a researcher.

I am very grateful to all the members of my research groups, both past and present, for their support and helpful discussions. I would also like to thank the Economics and Social Sciences Research Council (ESRC) for their financial support.

I want to give a special thanks to my friend and office-mate Michał Wójcik for his advice, help and support.

Finally, I would also like to thank my friends for their love, support and encouragement.

Contributions

All sections of this thesis constitute independent work.

Prof. Christopher Summerfield provided input about the conceptualisation, design, analysis and writing for Chapters 2 and 3. Prof. Christopher Summerfield and Dr Sophie Arana provided further input regarding the conceptualisation, design and analysis for Chapter 4.

Assoc. Prof. Mark Stokes provided input about the conceptualisation, design and analysis for Chapters 2, 3 and 5.

Human participant data from Chapter 4 has previously been published in Feldman (2000).

This work was funded by an ESRC DTP Studentship and the COVID-19 Scholarship Extensions Fund (CSEF). It has also received indirect funding from European Research Council grant (ERC Consolidator Award: 725937, https://erc.europa.eu) and the Human Brain Project (Special Grant Agreement No: 945539, https://www.humanbrainproject.eu/en/) awarded to Christopher Summerfield, as well as the James S. Mc-Donnell Foundation (award 220020405) awarded to Mark Stokes.

Table of contents

Table of contents 5

Chapter 1. General Introduction 10

1.1 Introductory remarks 10

1.2 Cognitive control 10

1.2.1 Definitions, theories, and key features 10

1.2.2 Control in the brain – the prefrontal cortex 15

1.3 Working memory 20

1.3.1 Persistent activity as the neural basis of working memory 20

1.3.2 Dynamic codes and activity-silent mechanisms 21

1.4 Artificial neural networks 22

1.4.1 History 23

1.4.2 Multi-layer perceptron 24

1.4.3 Recurrent neural networks 25

1.4.4 Learning rules and objectives 27

1.4.5 Applications in cognitive neuroscience 28

1.5 Synopsis 31

Chapter 2. Selection of information from working memory in a recurrent neural network 32

2.1 Abstract 32

2.2 Introduction 33

2.3 Methods 35

2.3.1 Model architecture 35

2.3.2 Experiment 1 37

2.3.3 Experiment 2 46

2.3.4 Experiment 3 47

2.3.5 Software 47

2.3.6 Statistics abbreviations 48

2.4 Results 48

2.4.1 Experiment 1 48

2.5 Discussion 77

Chapter 3. Exploring the neural basis of retro-cueing benefit in recurrent neural networks 84

3.1 Abstract 84

3.2 Introduction 85

3.3 Methods 88

3.3.1 Task structure 88

3.3.2 Model architecture and training 88

3.3.3 Hidden activity analysis 89

3.3.4 Multiple linear regression analysis 89

3.3.5 Behavioural analysis 89

3.3.6 Software 91

3.3.7 Statistics abbreviations 91

3.4 Results 91

3.4.1 Networks trained with informative retrocues show the retrocueing effect 91

3.4.2 Parallel Cued item geometry is only observed following deterministic retrocues 96

3.4.3 Neural geometries formed in the post-probe delay period are behaviourally relevant 97

3.4.4 Retrocueing effect is linked to a gain-modulation mechanism of cued item representations 100

3.5 Discussion 105

Chapter 4. Contrasting the symbolic and connectionist accounts of rule learning 110

4.1 Abstract 110

4.2 Introduction 111

4.3 Methods 114

4.3.1 Task structure 114

4.3.2 Model architecture 116

4.3.3 Model training 116

4.3.4 Learning speed measure 117

4.3.5 Sensitivity to noise metrics 117

4.3.6 Regression methods 118

4.3.7 Prototype and exemplar models of category learning 118

4.3.8 Human participant task and data 119

4.4 Results 120

4.4.1 Learning speed of NNs is sensitive to Boolean complexity 120

4.4.2 Noise sensitivity in networks and Boolean complexity 121

4.4.3 Alternative models of categorisation 124

4.4.4 Comparison of learning speed in networks and humans 125

4.5 Discussion 128

Chapter 5. Unsupervised learning and generalisation of abstract task rules 132

5.1 Abstract 132

5.2 Introduction 133

5.3 Materials and Methods 139

5.3.1 Subjects 139

5.3.2 Stimuli 139

5.3.3 Experiment protocol 140

5.3.4 Analysis 143

5.4 Results of Study 1 143

5.4.1 RT signatures of hierarchical learning 146

5.4.2 Performance benefits during transfer 149

5.4.3 Transfer choice patterns 151

5.4.4 Association between the hierarchical strategy use during Baseline problem and magnitude of transfer benefit 153

5.5 Results of the replication study (Study 2) 155

5.5.1 Signatures of hierarchical learning - RT cost analysis 155

5.5.2 Performance benefits during transfer 155

5.5.3 Transfer choice probability 156

5.5.4 Association between the hierarchical strategy use during Baseline problem and magnitude of transfer benefit 157

5.6 Discussion 159

Chapter 6. General Discussion 165

5.1 Chapter summaries and contributions 165

5.2 Limitations, outstanding questions and future directions 169

5.4 Conclusion 173

References 174

Supplementary Information for Chapter 2 193

Supplementary Information for Chapter 4 195

Supplementary Information for Chapter 5 200

Chapter 1. General Introduction

1.1 Introductory remarks

The natural environment is non-stationary – whilst our daily life often follows a certain routine, circumstances can change, and no two experiences are the same. To thrive, intelligent agents must therefore be able to flexibly adjust their behaviour to support the pursuit of their goals. In this thesis, I explore different aspects underlying such flexible, intelligent behaviour. The results chapters of this thesis are relatively self-contained and contain their own introductions and discussions. Therefore, the purpose of the current chapter is to describe the unifying concepts behind the described work. I start by introducing the construct of cognitive control, which has been used to describe the collection of mechanisms that allow intelligent systems to modulate their behaviour in a context-dependent way. I then discuss how cognitive control is implemented in the brain, with a particular emphasis on the prefrontal cortex. Subsequently, I introduce the concept of working memory – a cognitive function required for the maintenance and manipulation of information no longer present in the environment, alongside its neural mechanisms. Finally, I introduce artificial neural network models, which are used as a tool for investigating questions from cognitive neuroscience throughout the chapters of this thesis.

1.2 Cognitive control

1.2.1 Definitions, theories, and key features

In psychology and cognitive science, mechanisms that allow intelligent systems to modulate their behaviour in a context-dependent way have collectively been studied under the construct of cognitive control (Cohen, 2017). Early definitions thereof drew on a distinction between automatic (habitual) and controlled processes, the latter of which was characterised

by slower execution speed, being subject to interference from the former, and capacity constraints (Posner and Snyder, 1975; Schneider and Shiffrin, 1977; Shiffrin and Schneider, 1977). This classification was based on behavioural findings from classic paradigms such as the Stroop task (Stroop, 1935) and dual-task interference paradigms (Telford, 1931; Shaffer, 1975).

In the Stroop task, participants are asked to either read the word presented on the screen, on name the colour in which it is written. Firstly, subjects are almost universally faster at word reading than colour naming. Furthermore, there is a characteristic performance difference on incongruent trials (where the font colour does not match the semantic content of the word, such as the word 'GREEN' written in red). Under such conditions, word reading performance is unaffected, whilst colour naming is subject to severe interference, which lead to their classification as automatic and 'controlled' processes, respectively (MacLeod, 1991). However, it was quickly recognised that a given task (such as colour naming) could exhibit the putative characteristics of either an automatic or controlled process, depending on factors like context (e.g., competing task) or the degree of practice (Schneider and Shiffrin, 1977; MacLeod and Dunbar, 1988), indicating the existence of a spectrum of automaticity. More recent definitions of cognitive control thus focus on the flexibility aspect, stressing its role in regulating the execution of tasks in a goal-directed and context-dependent way (e.g., Braver, 2012; Cohen, 2017; Musslick and Cohen, 2021). This capacity for flexibility is arguably the most important aspect of intelligent behaviour.

Dual-task interference paradigms generally require subjects to perform two tasks within the same trial (e.g. Telford, 1931). Results from such studies show that performance on control-demanding tasks suffers when subjects are asked to perform another task concurrently, even

when the two tasks are unrelated (i.e., use different stimuli and responses; e.g., Levy and Pashler, 2001). This highlights one of the key features of control-dependent processing, namely - capacity constraints (Cohen, 2017). Many theories of cognitive control assume that such constraints are a feature of the control mechanism itself (e.g., Posner and Snyder, 1975; Shiffrin and Schneider, 1977). This includes accounts collectively known as *structural bottleneck theories*, which describe the control mechanism in terms of a single, centralised, serial-execution process, akin to the central processing unit of a modern computer (Broadbent, 1957; Welford, 1967; Pashler, 1994). The assumption of a central control resource is shared by *unitary resource theories*, which however allow for its sharing across tasks to support parallel execution, albeit at the cost of performance degradation (Kahneman, 1973; Norman and Bobrow, 1975; Tombu and Jolicoeur, 2003).

In contrast, *multiple-resource theories* argue for the existence of multiple, specialised, local resources (Wickens et al., 1980; Meyer and Kieras, 1997a, 1997b; Salvucci and Taatgen, 2008). Tasks can demand one or more of such resources (e.g. representations or processes), with some theories arguing that a resource can only be used by one task at a time (e.g., Salvucci and Taatgen, 2008) and other allowing for resource sharing between tasks (e.g., Wickens et al., 1980). Experimental evidence consistent with the *multiple-resource theories* includes the findings of increased dual-task interference for tasks characterised by a structural overlap, such as stimulus or response modalities (Koch, 2009; Göthe et al., 2016).

A particularly compelling recent theoretical perspective casts the capacity-limited nature of cognitive control as *a feature, not a bug* (Musslick and Cohen, 2021). The authors argue that, rather than reflecting an evolutionary fiasco (after all, there are around 86 billion neurons in the human brain (Herculano-Houzel, 2012), it receives over 20% of the total

energy budget, and its parallel processing capabilities in other domains such as vision are formidable, begging the question of why higher cognitive function should be so neglected in comparison), the limits on the multitasking capacity are in fact adaptive in nature. If it is the overlap in task resources that leads to cross-interference and the need for control (as posited by the multiple-resource theories), why does the brain rely on shared task representations at all? Based in part on the results of their numerical and analytical work involving *artificial neural networks* (a class of models introduced in more detail in section **1.4**; Musslick et al., 2017, 2020; Sagiv et al., 2020; Ravi et al., 2021), the authors argue that representation sharing confers a rapid task learning and generalisation advantage. Thus, intelligent systems might optimise their neural representations for learning speed, at the (initial) cost of multitasking capability. Under this account, cognitive control reflects an adaptation that allows the system to resolve interference in the short-term, with repeated multitasking training leading to the separation of task representations and their subsequent automatization (Musslick and Cohen, 2021). This framework elegantly illustrates the utility of artificial network models as a tool for investigating questions from cognitive neuroscience. The themes of shared and separate task representations, as well as their neural implementation, are explored further in this thesis (Chapters 2, 3 and 5).

When discussing the capacity-limited nature of cognitive control, it is important to note that such limits pertain not only to the number of tasks that can be executed at the same time, but also the total length of time that a given control-demanding task can be performed for without lapses (such limits should be intimately known to anyone who has ever attempted to read a particularly convoluted scientific paper in a single attempt). The temporal constraints of cognitive control have also been assumed to reflect a limited resource. One theory, known as the '*ego depletion hypothesis*' (Baumeister et al., 1998) argues that the

resource is metabolic in nature. Whilst this notion fits well with the subjective feeling of effort that often accompanies the allocation of control (imagine sitting down in front of your computer in the morning, prior to having your coffee, and reflecting on the sheer amount of raw intellectual work that is ahead of you), both the empirical evidence for it (Carter et al., 2015) and its physiological basis (Kurzban, 2010) have since been questioned. A different account suggests that the temporal limits reflect motivational factors Whilst a full discussion of the theories addressing the topic of *allocation* of cognitive control is outside the scope of this thesis (refer to Botvinick and Braver, 2015 and Kool et al., 2017 for excellent reviews), one related theoretical account of interest is the *dual mechanisms of control* (DMC) framework (Braver, 2012). The central tenet of this proposal is that there exist two fundamentally different ways in which control can be exerted. The first one, termed *proactive* control, is conceptualised as a conflict prevention mechanism whereby goal information is selected and actively maintained in memory from the outset of the task. Conversely, *reactive* cognitive control is a corrective mechanism that relies on the detection of conflict between context-dependent and habitual responses. The two mechanisms are associated with distinctive computational trade-offs. Whilst proactive control might be a more reliable strategy, this comes at the cost of flexibility and capacity for the maintenance and processing of other sources of information. In turn, reactive control is theorised to be less resource-consuming but more prone to failure, jointly due to its reliance on salient triggers as well as the higher potential for attentional capture by irrelevant information. The two mechanisms further differ in terms of their spatial and temporal profiles of brain activity – in particular, proactive and reactive control are associated with the sustained and anticipatory versus transient activity in the lateral prefrontal cortex (the area of the brain proposed to be the locus of cognitive control execution by Miller and Cohen (2001), as elaborated on in the next section). Lastly, the extent to which either mechanism is recruited

has been shown to depend on a range of factors, including the expected amount of interference (Speer et al., 2003; Burgess and Braver, 2010) as well as individual differences in cognitive and affective domains (Burgess and Braver, 2010; Jimura et al., 2010).

1.2.2 Control in the brain – the prefrontal cortex

Several theoretical accounts attribute the role of guiding behaviour according to the current goals to the prefrontal cortex (PFC; Cohen and Servan-Schreiber, 1992; Passingham, 1993; Miller and Cohen, 2001; Grafman, 2002) - a neocortical area most developed in primates, which are well known for the behavioural flexibility (Miller and Cohen, 2001). One of them goes further to specify that this is critically dependent on the ability to actively maintain goal representations in mind (Miller and Cohen, 2001), thus linking control to another cognitive function known as working memory (Baddeley and Hitch, 1974). I first review the multiple lines of evidence that indicate that the execution of control processes in the brain is carried out by the PFC, before describing the neurophysiological mechanisms of working memory (in section **1.3.1**).

1.2.2.1 Patient and lesion studies

Patients with damage to the PFC show striking deficits in terms of behavioural flexibility, whilst often being virtually unaffected in other domains (including perception and motor skills, as well as the execution of habitual behaviours; Shallice and Burgess, 1991). In a laboratory setting, this is evident from the impairments frontal patients show on the previously introduced Stroop task (Perret, 1974; Dunbar and Sussman, 1995; Vendrell et al., 1995), particularly on incongruent trials (Perret, 1974; Vendrell et al., 1995) or when the instructions vary frequently (Dunbar and Sussman, 1995).

Deficits can also be observed on the Wisconsin Card Sorting Task (WCST), which requires subjects to sort a deck of cards based on one of their visual features (shape, colour, or number of symbols). Each card thus affords different actions, depending on the currently valid sorting rule. The rule changes periodically without instructions, requiring the switches to be inferred based on the received feedback. Frontal patients have little difficulty learning the initial stimulus-response mappings but show impaired performance following rule switches (i.e., perseveration; Milner, 1963; Barceló and Knight, 2002). Similar deficits have been observed in monkeys with PFC lesions on analogous tasks (Dias et al., 1996; Buckley et al., 2009; Moore et al., 2009). Furthermore, impairments can also be seen in other tasks when patients are cued about the rule switches and they occur at predictable time intervals (Aron et al., 2004).

Finally, patients also show impairments in situations that require them to chain sequences of actions together, whereby they often skip steps, insert incorrect ones, or fail to organise them in an appropriate order (Luria, 1980; Duncan, 1986; Grafman, 1989; Duncan et al., 1997), despite being able to perform the individual steps correctly (Shallice and Burgess, 1991). Collectively, these findings point towards the role of the PFC in guiding behaviour according to specific task rules (including overriding other, more habitual responses) as well as structuring it in time.

1.2.2.2 Neuroimaging evidence in humans

In line with the deficits described in the section above, neuroimaging studies of healthy human subjects show prefrontal engagement on the Stroop task (Carter et al., 1995; Kerns et al., 2004) and WCST (Monchi et al., 2001; Nagahama et al., 2001; Nakahara et al., 2002).

Furthermore, modulation of prefrontal activity is observed when participants are asked to respond to stimuli according to a set of abstract rules (Bunge, 2004; Crescentini et al., 2011; Mansouri et al., 2020). This includes match and non-match rules (Bunge et al., 2003), rules involving numerical concepts (Nieder, 2005; Nieder and Dehaene, 2009) and rules involving sequences of stimuli (Sakai and Passingham, 2003).

1.2.2.3 PFC as a multimodal hub

As pointed out by Miller and Cohen (2001), one of the requirements for a cognitive control system is that it have access to all the relevant sources of information necessary to guide behaviour. This includes both information about the external environment, as well as the internal states of the organism.

The PFC can be subdivided into orbital, medial, lateral, and mid-dorsal regions (Miller and Cohen, 2001). It enjoys convergent and reciprocal connections with all sensory systems (visual, auditory, somatosensory, which mostly project to the lateral and mid-dorsal subdivisions), receives inputs from the limbic and midbrain regions associated with memory, affect and reward processing (hippocampus and associated cortex, amygdala, hypothalamus; these projections target the orbital and medial PFC) and projects to cortical and subcortical motor structures (such as supplementary and pre-supplementary motor cortices, cerebellum, superior colliculus and basal ganglia). Furthermore, it is also densely interconnected with itself, both between and within the subdivisions (for a more comprehensive review, including details of regional specialisation, see Miller and Cohen, 2001). Thus, anatomically, the PFC is a site of neural convergence, well positioned not only to integrate information from multiple sources, but also affect its processing and influence the outputs in a top-down way to guide behaviour.

1.2.2.4 Evidence from single cell recordings

Electrophysiological recordings in non-human primates show a large proportion of prefrontal neurons (in particular, those located in the dorsolateral PFC, dlPFC) modulates their firing rates depending on virtually all kinds of task-relevant variables, and this is especially true when the task is demanding (Duncan and Miller, 2002). Examples include features of visual stimuli (e.g., location: Funahashi et al., 1989; Wilson et al., 1993, shape: Meyer et al., 2007, colour: Buschman et al., 2011), but also extend to other sensory modalities such as auditory tones (Fuster et al., 2000) and proprioception (Romo et al., 1999). Furthermore, individual cells have also been shown to selectively respond to motor (e.g., the direction of the required saccade: di Pellegrino and Wise, 1993) and reward variables (including identity and magnitude: Watanabe, 1996; Cavanagh et al., 2018). Collectively, these results demonstrate that PFC neurons represent a wide range of task variables.

Furthermore, PFC neurons have been shown to encode abstract task rules (Asaad et al., 1998; Wallis et al., 2001). For example, when a monkey is required to compare the identity of two visual stimuli, activity of individual PFC neurons encodes the 'match' and 'mismatch' rules (Wallis et al., 2001). Importantly, prefrontal neurons have been shown to encode task rules in an abstract format - the same profile of activation is observed irrespective of the type of cue (e.g., auditory tone or drop of juice) used to signal a given rule (Wallis et al., 2001). Interestingly, neurons selective for a particular object category from a pair (such as cats vs dogs) can show the same selectivity profile to an entirely different category (e.g., type of car), when the overall task (binary classification) remains the same (Cromer et al., 2010). Collectively, these findings demonstrate that PFC neurons can represent task information in an abstract format that uncouples it from lower-level (e.g.,

sensory) information, and such coding can be readily reapplied in a new setting with the same task structure.

Lasty, some neurons also show different firing patterns to visual stimuli depending on the currently valid rule (Asaad et al., 2000). Others dynamically modulate their selectivity for different visual features (e.g., object identity versus location) in the course of the single trial, depending on the changing task demands (Rao et al., 1997; Spaak et al., 2017). Such conditional, conjunctive coding of task information has been termed 'mixed selectivity' (Rigotti et al., 2013). Crucially, it makes it challenging to investigate the function of PFC using classical neuroscience methods, which require the researcher to characterise the computations carried out by single neurons before explaining how they are combined at the circuit-level. Instead, it has been argued that to understand the function of many higher order brain areas with heterogenous response profiles of single neurons (including PFC), one must simultaneously consider the activity of the entire neuronal population (Quian Quiroga and Panzeri, 2009; Panzeri et al., 2015; Richards et al., 2019). This is not only because information is assumed to be encoded in a distributed code rather than in the firing patterns of individual neurons (Hebb, 1949), but also because complex systems made up from multiple interacting simple units (such as neuronal circuits) are known to exhibit emergent properties which cannot be explained in terms of the operations of their constituent elements (McClelland, 2010). Fortunately, recent years have seen a rise in the development of theoretical modelling and analysis methods required to understand how information is encoded and processed through the simultaneous activity of a population of neurons. One promising approach has been the use of artificial neural networks (Richards et al., 2019), which are introduced in section **1.4** of this chapter.

1.3 Working memory

As previously articulated by Miller and Cohen (2001), working memory is one of the key requirements for a system capable of implementing cognitive control. Working memory (WM) refers to the ability to maintain and manipulate information in mind over short timescales on the order of seconds (Baddeley and Hitch, 1974). It allows agents to decouple their actions from the immediate perceptual input, supporting flexible, context-dependent behaviour in situations where the relevant information is no longer readily available from the environment. Neurally, WM is implemented by a wide network of brain areas, including sensory regions as well as areas constituting the frontoparietal circuit – especially the PFC (Christophel et al., 2017). Early evidence for its critical dependence on the PFC came from studies in non-human primates, where it was demonstrated that lesions to this region selectively impaired the animal's ability to maintain information over time, without affecting behaviour that was not reliant on memorisation (Jacobsen, 1935; Blum, 1952; Mishkin, 1957). Similar observations have been made in frontal patients on tasks requiring the manipulation of information in memory, despite the lack of differences in their short-term maintenance capacity relative to control subjects (D'Esposito and Postle, 1999).

1.3.1 Persistent activity as the neural basis of working memory

In the 1970s, electrophysiological recordings in non-human primates demonstrated persistent activity in the dlPFC that spanned the memory delay between stimulus presentation and response (Fuster and Alexander, 1971; Kubota and Niki, 1971) Similar observations were made in later studies which proposed that such persistent, stimulus-specific activity is the neural substrate of working memory (Goldman-Rakic et al., 1992; Goldman-Rakic, 1995a). Accordingly, persistent spiking has been shown to correlate with

behavioural performance (Constantinidis et al., 2001; Brody et al., 2003; Zhou et al., 2013). Significant maintenance related activity in the PFC has also been observed in several fMRI studies with healthy human subjects (Courtney et al., 1998; Zarahn et al., 1999; Sakai et al., 2002), with its magnitude predicting task performance (Pessoa et al., 2002; Sakai et al., 2002). In a further parallel to the non-human primate findings, such activity was observed to be specific to the task-relevant variables such as visual and auditory stimuli (Ester et al., 2015; Kumar et al., 2016). Convergent evidence has recently been obtained in human intracranial recording studies, where persistent activity centered on the PFC was shown to encode abstract task variables and predict the subject's response (Haller et al., 2018). On the basis of such diverse findings, the assertion that persistent sustained activity is the neural basis for working memory has become somewhat of a central dogma in working memory research (Constantinidis et al., 2018; Wang, 2021a).

1.3.2 Dynamic codes and activity-silent mechanisms

Interestingly, persistently firing cells can make up only a small proportion of the neuronal population (Brody et al., 2003) and patterns of PFC population activity (as well as the activity of individual neurons, as described in the previous section) often vary across time (Rao et al., 1997; Spaak et al., 2017). This is especially true for more complex tasks, both in biological (Meyers et al., 2008; Stokes et al., 2013) and artificial neural networks (Orhan and Ma, 2019), suggesting that such dynamic activity patterns might play a role in the manipulation of information in working memory. However, could such dynamic activity patterns also support the stable maintenance of information over time? One possibility is that both coding schemes (stable and dynamic) can coexist within the same population, but maintenance relies only on a stable subpopulation of persistently active neurons, whilst the time-varying dynamics are not task-related (Murray et al., 2017). However, activity patterns

of PFC neurons with time-varying dynamics do contain information about task variables (Baeg et al., 2003; Brody et al., 2003; Zaksas and Pasternak, 2006), and its amount (as measured by the test accuracy of a classification algorithm trained to discriminate activation patterns to different task conditions) does not differ between the persistent and 'dynamic' cells (Baeg et al., 2003). Most surprisingly, both monkey and human neuroimaging studies show that, even when the task-relevant neuronal activity falls back to baseline during the memory delay, it can be reinstantiated and guide subsequent behaviour (LaRocque et al., 2013; Watanabe and Funahashi, 2014; Wolff et al., 2017). These findings suggest that persistent activity is not always critical for working memory maintenance, and another, 'activity-silent' mechanism might be at play. Accordingly, it has been proposed that memories could be maintained through short-term synaptic changes to the prefrontal circuitry (Mongillo et al., 2008; Barak and Tsodyks, 2014; Stokes, 2015). Empirical evidence in favour of such accounts comes from human neuroimaging studies, which demonstrate that non-specific stimulation (either in the form of a blank visual stimulus, or a transcranial magnetic stimulation pulse) can reinstate the pattern of activation observed to a previously encoded stimulus after its active representation has decayed back to baseline (Rose et al., 2016; Wolff et al., 2017). Importantly, such synaptic changes have been proposed to lead to the creation of a temporary task-relevant circuit which can affect the processing of subsequent inputs in a context-dependent way (Stokes, 2015).

1.4 Artificial neural networks

Artificial neural networks (NNs) are a class of powerful machine learning models, mathematically proven to be capable of approximating any function given an arbitrary network size and a nonpolynomial activation function (Leshno et al., 1993). In addition to a

wide range of real-world applications, they also have a long history of use in modelling cognitive and neurobiological phenomena (Zipser and Andersen, 1988; Cohen et al., 1990).

1.4.1 History

The origins of artificial neural networks can be traced back to the work of McCulloch & Pitts who proposed that biological neurons could express the truth values of Boolean expressions through their all-or-none firing pattern (McCulloch and Pitts, 1943). Per their proposal, a neuron (or a circuit thereof) could implement propositional logic through two elementary operations: the summation of its inputs and thresholding of its output. A decade later, their model was extended by Rosenblatt (1958), giving rise to the *perceptron*. This computational unit forms the basic building block of many modern artificial neural networks. The perceptron, like the McCulloch-Pitts neuron, receives N numerical inputs x and multiplies them by their respective connection weights w. The weighted inputs are summated with the bias value b (which sets the base level of the activation of the neuron). Finally, a nonlinear activation function f is applied to produce the output y (see also **Fig. 1.1A**):

$$y = f(\sum_{i}^{N} x_i w_i + b)$$

In the original formulation this was the heavyside step function which maps positive values to 1 and returns 0 otherwise (shown in **Fig. 1.1A**). Importantly, the power of the perceptron lies in its adaptability - it can learn to solve a problem by adjusting its parameters (the weights and bias) through experience, via an error-based learning mechanism.

The invention of the perceptron – a flexible algorithm that could learn problem solutions from experience – caused a wave of excitement in the AI community. However, it was quickly recognised that its use was heavily limited to the linear domain whilst more complex, non-linear problems (such as the classic exclusive-or explored in the chapters of this thesis) required a more complex approach (Minsky and Papert, 1969). Such problems could be solved by a network composed of multiple perceptron units arranged into successive layers, but a major obstacle to progress proved to be the training of such networks. Whilst the outmost layer of the network could be trained by the same principle as the perceptron – that is, by updating the parameters proportionally to the difference between the network prediction and the required output, it was unclear how to calculate an equivalent error signal for the preceding layers. Thus, the field came to a halt until the invention of the backpropagation algorithm (Rumelhart et al., 1986) which allowed the error-driven parameter updates to be calculated efficiently for multiple layers by applying the chain rule of calculus.

1.4.2 Multi-layer perceptron

Networks composed of multiple layers of perceptron units are classed under the multi-layer perceptron (MLP) architecture (**Fig. 1.1B**). An MLP includes an input layer (which receives the external data), an output layer (producing the response such as a binary classification value), and at least one intermediate (*hidden*) layer. In common terminology, layer *width* corresponds to the number of comprising units, whereas network *depth* describes the number of layers. Units are referred to as *nodes* (or neurons), and the connections between them as *edges*. Edges have weights w_{ij}, where the indices correspond the edge's input and output nodes. Typically, each node connects to every node of the successive layer, giving rise to a fully-connected network. Individual nodes in the network follow the basic order of

operations of the perceptron, though other activation functions f are allowed. These include the sigmoid function (resembling the relationship between summated input strength and firing rate in biological neurons), or the rectified linear unit (ReLU) function (which returns the same value for positive inputs, and 0 otherwise). The activity of a node j in layer $n+1$ is given by:

$$a_{n+1,j} = f(\sum_{i}^{width} a_{n,i} w_{n+1,ij} + b_{n+1,j})$$

To compute the output, activation in the network is propagated forwards through the successive layers – hence why the MLP is classed as a type of a *feedforward* network.

1.4.3 Recurrent neural networks

Whilst feedforward neural networks enjoy a wide range of applications, they are not suitable for tasks characterised by a temporal element, such as working memory paradigms used in cognitive neuroscience. This is because the activation in such networks is influenced only by the current input, thus rendering feedforward models unable to remember information supplied at previous points in time. A different class of models - recurrent neural networks (RNNs) - has been developed specifically to allow for processing of sequences of inputs structured in time. Their simplest representative is the vanilla RNN (illustrated in **Fig. 1.1C-D**), which is an extension of the MLP architecture described in the section above. Specifically, in addition to the between-layer projections, hidden nodes of the network are also connected to one another with an additional set of weights, allowing them to integrate the memory of the previous network state into their activation and thus letting it influence the current output. The activity of a hidden node n at a timepoint t is given by:

$$h_{n,t} = f(\sum_{i}^{N_{inp}} x_{j,t} w_{ij} + \sum_{j}^{N_h} h_{n,t-1} w_{jn} \ + b_n)$$

where N_{inp} and N_h correspond to the widths of the input and recurrent (also called hidden) layers, respectively. The activity of hidden nodes can be readout by the output layer (usually composed of perceptron-like units) at every timestep (as shown in **Fig. 1.1C**), or only at the end of the stimulus sequence, depending on the demands of the task. At the beginning of a stimulus episode, the hidden layer unit activations are initialised to a state h_0. Typically, this is a zero-vector, which ensures that the activity of the network elicited in response to the previous stimulus does not influence the activity on the current trial.

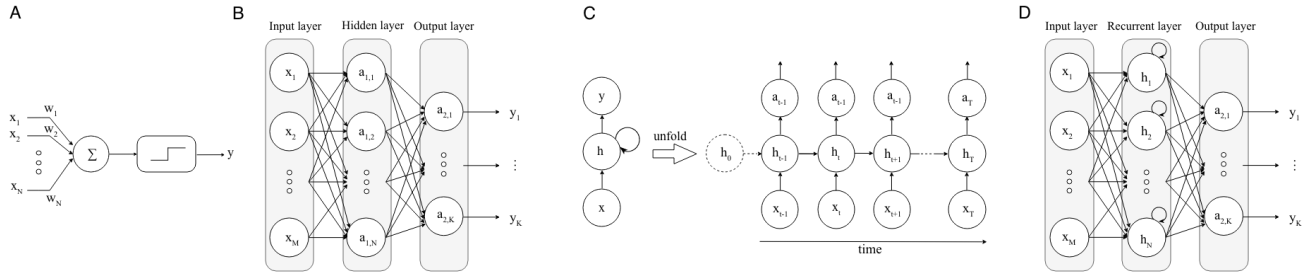


Fig. 1.1 Basic feedforward and recurrent neural network architectures. A. Diagram illustrating the order of processing steps in a perceptron unit. Inputs $(x_1, x_2 \ldots, x_N)$ are multiplied by their respective weights $(w_1, w_2 \ldots, w_N)$ and summated. The result is then passed through the heavy-side function to produce the output y. Only three inputs and weights are shown for simplicity. **B.** Network diagram for an MLP with a single hidden layer. Nodes represented by circles, edges by arrows. Layers of nodes depicted by shaded boxes. Only 3 input and hidden nodes, and 2 output nodes shown for simplicity. **C.** Diagram illustrating the operation of a vanilla RNN. Circles correspond to the input, hidden and output layer vectors (x, h, and y, respectively), arrows denote the direction of information flow. The network can be 'unfolded' in time – at each timepoint t, hidden layer activity vector from the previous timepoint h_{t-1} is combined with the current input vector x_t to produce the new hidden state h_t, which influences the activation vector of the output layer units a_t. This computation is repeated for the total duration of the stimulus sequence T. **D**.

Network diagram for a vanilla RNN. All conventions as in panel B. Circular arrows denote recurrent connections.

1.4.4 Learning rules and objectives

Machine learning approaches can be broadly divided into 3 categories depending on the type of information supplied during training, namely: supervised, unsupervised, and reinforcement learning. In a supervised paradigm, the model is given an explicit teaching signal informing it of the correct response (such as a classification label) for each training item. Models trained in reinforcement learning set-ups also receive external feedback (in the form of a scalar reward signal), though it is less explicit in nature. Finally, unsupervised training proceeds without any feedback from the environment. As only the supervised training paradigm is relevant for the chapters of this thesis (and is perhaps the most common technique used for NN training), it is introduced in more detail below.

Neural networks such as the example architectures described in the sections above can be trained to solve an impressive range of complex real-world problems, including image recognition, text recognition and completion. This remarkable flexibility stems from the principle of function optimisation, whereby network parameters are gradually adjusted to satisfy some functional objective, such as the minimisation of classification error. Training proceeds through the repeated presentation of a training dataset to the network. On each iteration, the discrepancy between the network and desired behaviour is calculated according to a mathematical function termed the *loss function* and minimised by adjusting the network parameters according to another equation termed the *learning rule*. One of the most common optimisation techniques is gradient descent, in which the first-order derivative of the loss function with respect to model parameters is calculated. This can be done using the entire

training dataset, small subsets thereof (in which case it is termed mini-batch gradient descent), or individual items (corresponding to stochastic gradient descent, SGD). Model parameters are subsequently updated in the opposite direction of the calculated gradient until a local minimum is reached, with the update magnitude controlled by a hyperparameter called the *learning rate*. For simple problems characterised by convex solutions, this procedure is guaranteed to find the optimal solution (i.e., the local minimum is equivalent to the global minimum). For more complex problems, small learning rate magnitudes are preferred to minimise the chances of training converging to a local minimum.

1.4.5 Applications in cognitive neuroscience

The use of NNs to model human brain and behaviour is generally referred to as connectionism (Fodor and Pylyshyn, 1988; Pinker and Prince, 1988; Bechtel and Abrahamsen, 1991). The practice quickly rose to popularity in the 1980s, partly in thanks to the Parallel Distributed Processing group (Rumelhart et al., 1987). Early applications of this modelling framework focused on the use of small networks to explain cognitive-level phenomena and included investigations of semantic memory and language development in childhood (Rumelhart and McClelland, 1987; McClelland and Rogers, 2003) With the advent of new recording and analysis techniques, the scope for comparative analysis grew and activity of network units begun to be compared against neurophysiological data. Subsequently, NNs were found to develop similar coding properties of biological neurons (Mante et al., 2013; Yamins and DiCarlo, 2016; Chaisangmongkon et al., 2017; Banino et al., 2018). For example, RNNs have been found to replicate both the selectivity profiles and population dynamics observed in mammalian brains during complex decision-making tasks (Mante et al., 2013; Chaisangmongkon et al., 2017). Nowadays, NN research into cognition often combines different levels of analysis, allowing researchers to ask both mechanistic and

normative questions (cf. Chapters 2 and 3 of this thesis) and search for the answers through the analysis of different aspects of the model, from unit- and population-level activity to behavioural measures such as learning dynamics, error patterns and even reaction times (Rajalingham et al., 2018).

In the case of investigating higher cognitive functions, perceptual processes are often not of interest and thus perceptual difficulty of the experimental paradigm is kept at a minimum. For that reason, perception is usually bypassed altogether in the modelling process, including NN modelling. To this end, assumptions about the input representations must be made. Categorical input dimensions can be represented by one-hot vectors (i.e., vectors filled with zero-values par one position with a unit value), which ensures that all possible feature values are uniformly distinct from one another (this coding scheme is used in Chapter 4). For tasks with continuous dimensions such as those relying on colour discrimination, input representations can take the form of tuning curves (as used in Chapters 2 and 3). However, other possibilities exist - networks can also be trained on raw perceptual input of the kind presented to participants in a laboratory setting, though this approach might require the use of more advanced architectures for more efficient training.

Whilst choices made about the input representations are an important source of constraints for the modelling process, it is the distributed and parallel nature of representations and computations that form the core assumptions of the connectionist approach. This information processing style is often contrasted with the discrete, symbolic representations and sequential operations of modern computers (a dichotomy explored with regards to human categorisation behaviour in Chapter 4). Connectionist networks are composed of simple neuron-like units performing stereotypical local computations and as such, they

abstract away a great deal of biological detail. The continuous unit activity can be seen as corresponding to the mean firing rate of a biological neuron, or a population thereof. Complex computations emerge from the interactions between the simple units, which are in turn governed by the network connectivity patterns. Importantly, connectivity is usually only loosely specified at the onset of the investigation (through the chosen network architecture such as the MDL and vanilla RNN examples introduced before) and is fine-tuned during network training (by applying general-purpose learning rules). The relative lack of biological detail of network units is often put forward as an argument against the biological plausibility of connectionist models (Crick, 1989). However, as previously mentioned, neurons in simplified artificial NNs have been shown to develop representations similar to those seen in biological brains (Mante et al., 2013; Yamins and DiCarlo, 2016; Chaisangmongkon et al., 2017; Banino et al., 2018), suggesting that at least in some cases, adding more biological detail might not be essential in the pursuit of functional understanding at the chosen level of analysis. Furthermore, biological constraints can be added as needed when differences between network and biological function emerge, in a hypothesis-driven fashion. Similar arguments against biological plausibility have been previously made based on the use of the backpropagation algorithm for supervised network training (Crick, 1989). However, there is a growing body of literature investigating how the algorithm may be approximated in the brain (O'Reilly, 1996; Körding and König, 2001; Roelfsema and van Ooyen, 2005; Lillicrap et al., 2016; Bengio et al., 2017; Guerguiev et al., 2017; Scellier and Bengio, 2017; Whittington and Bogacz, 2017, 2019; Sacramento et al., 2018; Pozzi et al., 2019), with other learning rules often leading to the formation of highly similar internal representations.

1.5 Synopsis

In the following chapters, I present the results from four lines of research. Chapters 2 and 3 are concerned with the neural mechanisms by which information is manipulated in short-term memory according to task rules. In Chapter 4 and 5, I investigated learning of abstractions from experience, from an algorithmic and mechanistic perspective, respectively.

Chapter 2. Selection of information from working memory in a recurrent neural network

2.1 Abstract

When multiple items are held in short-term memory, cues that retrospectively prioritise one item over another (retro-cues) can facilitate subsequent recall. However, the neural and computational underpinnings of this effect are poorly understood. One recent study recorded neural signals in the macaque lateral prefrontal cortex (LPFC) during a retro-cueing task, contrasting delay-period activity before (pre-cue) and after (post-cue) retrocue onset. They reported that in the pre-cue delay, the individual stimuli were maintained in independent subspaces of neural population activity, whereas in the post-cue delay, the prioritised items were rotated into a common subspace, potentially allowing a common readout mechanism. To understand how such representational transitions can be learnt through error minimisation, we trained recurrent neural networks (RNNs) with supervision to perform an equivalent cued-recall task. RNNs were presented with two inputs denoting conjunctive colour-location stimuli, followed by a pre-cue memory delay, a location retrocue, and a post-cue delay. We found that the orthogonal-to-parallel geometry transformation observed in the macaque LPFC emerged naturally in RNNs trained to perform the task. Interestingly, the parallel geometry only developed when the cued information was required to be maintained in short-term memory for several cycles before readout, suggesting that it might confer robustness during maintenance. We extend these findings by analysing the learning dynamics and connectivity patterns of the RNNs, allowing us to make predictions for future studies. Overall, our findings are consistent with recent theoretical accounts which propose

that retrocues transform the prioritised memory items into a prospective, action-oriented format.

2.2 Introduction

Decisions are often based on information that occurred in the recent past. To bridge perception and action across an interval, items need to be maintained in short-term memory (Goldman-Rakic, 1995b; Fuster, 2001). In the mammalian neocortex, this memory process may rely on neurons that mutually excite each other in a recurrent circuit (Durstewitz et al., 2000). Simulations of this process using recurrent neural networks (RNNs) show how maintenance can rely on self-sustaining attractor states in neural state space. These models can account for empirical phenomena such as persistent delay-period activity and the resistance of memories to external perturbation (Wang, 2021b). When RNNs are trained from random weights with gradient descent (rather than being endowed with coding properties by hand) they can recreate neural motifs that are observed in primates in short-term memory studies. Tools from machine learning can thus offer a normative account of how dynamic memory processes are learned through error minimisation (Orhan and Ma, 2017; Masse et al., 2019). In this Chapter, we describe an example in which the activity dynamics in an RNN bear a striking resemblance to those reported in the primate brain.

When humans encode multiple stimuli into short-term memory, retrieval performance is enhanced by cues that reveal which item is more likely to be probed for recall. Surprisingly, this benefit occurs even if cues are presented during the delay period, retrospectively prioritising an item that was encoded in the past (Schmidt et al., 2002; Griffin and Nobre, 2003a). At first glance, this seems a puzzling phenomenon because these *retrocues* do not provide additional information about the probed stimulus, so that the benefit appears “out

of thin air". One possibility is that retrocues allow stored memories to be reformatted into a new frame of reference that renders them more robust (Myers et al., 2017a).

One recent study exploited multi-neuronal recordings from macaque lateral prefrontal cortex (LPFC) to understand the neural mechanisms of attention to memory (Panichello and Buschman, 2021). On each trial, monkeys were shown two coloured squares. During the *pre-cue delay*, they maintained both colours in short term memory, before viewing a retrocue that prioritised one of the two items. Following an additional *post-cue delay*, they reported the cued colour by making a saccade to the corresponding position on a colour wheel (**Fig 2.1A**). Using multivariate analyses, the authors studied the neural geometry (that is, patterns of similarity in neural states elicited by probed and unprobed items as a function of their colour) during both delay periods. Dissimilarities among neural patterns evoked by each colour scaled with distance around the colour wheel, such that each item occupied a position in neural state space that lay on a ring embedded on a 2D plane. This circular geometry was also seen in human EEG signals for an equivalent task (Hajonides et al., 2021).

However, the critical finding relates to the orientation of these 2D planes for each delay period. In the pre-cue delay, the two planes lay perpendicular to each other. This is consistent with a growing literature suggesting that the brain minimises interference by encoding information in orthogonal neural subspaces (Flesch et al., 2021; Libby and Buschman, 2021; Nieh et al., 2021; Xie et al., 2022). However, in the post-cue delay, the planes rotated so that they lay both in parallel and in phase, forming a common 2D plane whose geometry was isomorphic to the organisation of the colour wheel used for responding (Bernardi et al., 2020; Sheahan et al., 2021). This is consistent with the view that retrocues benefit

performance by transforming memories into a prospective, action-oriented format, which may render them robust to interference (Myers et al., 2017a).

In this Chapter, we modelled the emergence of this neural geometry using a recurrent neural network (RNN). We trained randomly initialised RNN models with gradient descent to perform an equivalent retrocue task, predicting that this orthogonal-to-parallel coding geometry would emerge naturally as a consequence of training the network to minimise recall error. This hypothesis was inspired in part by recent findings that neural networks project information into orthogonal subspaces to minimise interference, and parallel subspaces to permit generalisation (Remington et al., 2018; Luyckx et al., 2019; Bernardi et al., 2020; Flesch et al., 2021; Sheahan et al., 2021). We find that this is indeed the case. We analyse several aspects of the model, including its learning dynamics and sensitivity to the variability in the delay period, allowing us to make predictions for future neural recording studies.

2.3 Methods

2.3.1 Model architecture

Each recurrent neural network (RNN) consisted of an input, recurrent and output layer. The input layer consisted of 36 units in total. There were 2 binary units corresponding to the location retro-cues (and probes, if used), 17 colour units encoding the feature value of the stimulus shown at location 1 (L1) and another set of 17 units encoding the feature value at location 2 (L2). Here, we assume that feature values are hues drawn from a circular colour space, but our simulations hold for any circular feature space (e.g., orientation).

Each of the feature-sensitive units had a circular normal tuning function centred on one of 17 equally spaced values. On stimulation, each unit i produced activity u_i which depended on its preferred colour ϕ and the current stimulus colour α, and was given by the von Mises probability density function (1.1):

$$u_i = \frac{e^{\kappa \cos(\alpha - \phi)}}{2\pi \, I_0(\kappa)} \tag{1.1}$$

where $I_o(\kappa)$ is the modified Bessel function of order 0, and κ corresponds to the tuning width (chosen to be 5.0 for the reported simulations). The distribution of input activity was rescaled to be in the range [0,1]. The retrocue inputs (which occurred later) were one-hot signals denoting whether the top or bottom cue was presented.

The recurrent layer of the network consisted of $N_{rec} = 200$ units. The activity of the hidden unit vector h_t at time-step t was given by:

$$h_t = f\left(W_{rec} h_{t-1} + W_{inp} u_t + b_{rec} + \eta\right) \tag{1.2}$$

where W_{rec} and W_{inp} correspond to the recurrent and input weight matrices, u_t is the vector of input unit activity on timestep t, b_{rec} is a bias term and η is a noise term drawn from $N(0, \sigma^2)$. At the beginning of each trial, the hidden state was initialised at 0. The activation function f used was the rectified linear unit (ReLU) function. On the final timestep, hidden activity was read out by a single linear layer of N_{out} = 17 units, each associated with a response probability $\hat{r}_i$:

$$y = W_{out} h + b_{out} \tag{1.3}$$

$$\hat{r}_i = \frac{e^{y_i}}{\sum_j^{N_{out}} e^{y_j}} \tag{1.4}$$

where $\mathrm{W_{out}}$ is the output weight matrix and $\mathrm{b_{out}}$ is a bias term. The resulting output y is transformed to vector of responses probabilities (of making each of the $\mathrm{N_{out}} = 17$ colour choices) by a softmax function (1.4). For the reported behavioural analyses, we used a softmax policy to generate network colour choices from the vector of response probabilities $\hat{r}_i$.

2.3.2 Experiment 1

2.3.2.1 Task structure and training procedures

Each trial started with the presentation of two colour-location stimuli for one cycle, followed by 7 cycles of (*pre-cue*) delay, a single-cycle retro-cue input, and another 7-cycle (*post-cue*) delay. Networks (N=30) were trained on a dataset consisting of 512 unique trials, corresponding to all the possible combinations of 16 colours at the two stimulus locations and 2 retro-cues. We initialised the recurrent weights as an orthogonal matrix (Saxe et al., 2014) and used the standard Xavier initialisation for all other learnable network weights and biases. Each trial was associated with scalar feature values (α_{cued} and α_{uncued}), corresponding to the value of the tuning curve peak in colour space for the cued and uncued stimuli. Network parameters were tuned with stochastic gradient descent (batch size = 1 trial) using the RMSprop optimiser. We defined the ground truth response r as a one-hot vector over output units, with a 1 at α_{cued} and zeros elsewhere. The loss term was given by

the mean squared product of the (i) distance between r and the network output $\hat{r}$ and (ii) the circular distance between the vector of output unit tuning centres ϕ and scalar α_{cued}:

$$\mathcal{L} = \sum \left((\boldsymbol{r} - \hat{\boldsymbol{r}})(\phi - \alpha_{cued}) \right)^2 \quad (2.1)$$

The latter term ensures that the penalty the network receives is proportional to the circular distance from the correct response, and thus constrains the output values in $\hat{r}$ to have a smooth distribution with a peak at the correct answer location. Note that using a different loss function, whereby a scalar choice was sampled based on the output layer activity if the network and the angular error from the ground truth angle was calculated, would not encourage the network to learn a smooth distribution (in the case of a hardmax choice policy) or preclude network convergence (in the case of a softmax policy - due its stochastic nature, i.e. the one-to-many mapping from network output layer activity to choice).

We conducted a hyperparameter search over a range of different recurrent layer sizes, learning rate values and noise levels. The data reported in this paper comes from networks with 200 recurrent units and trained with learning rate set to 0.0001. We chose $\sigma = 0.07$, as it resulted in model performance that was qualitatively similar to that reported for monkeys (measured by the absolute mean of the error distribution). The pattern of results did not differ between networks trained under different hyperparameter choices. We stopped model training as soon as the loss function reached its final plateau. This moment was defined by two criteria: the (negative) slope of the loss function over the last 15 epochs falling below a threshold value of 2e-05, and the mean loss value over the most recent epoch falling below a threshold of 0.0036 (a value that corresponded to the performance level, as measured by absolute mean angular error, reported in monkeys). Note that imposing a threshold on the

absolute loss value was necessary to avoid stopping training prematurely (i.e., at the mid-training learning plateau, see also **Fig 2.4**).

2.3.2.2 Data analysis

We evaluated the trained models on the complete training dataset after freezing the weights (i.e., disallowing further gradient updates). The testing dataset contained 100 repetitions for each of the 512 unique trials (note that this was not redundant because the noise term introduced variability). To generate network responses, we sampled colour choices proportionally to their probabilities as given by the output layer activations (see equation 1.3).

2.3.2.3 Plane fitting, angles and phase alignment

To obtain planes of best fit for each stimulus location, we adapted the method described in Panichello and Buschman (2021). In short, for each trained RNN, we sorted the trials into B = 4 colour category bins and L = 2 locations, giving a total of B x L = 8 conditions. For each timestep of interest (the end points of the memory delays), we then calculated the mean hidden unit activity h by averaging within these bins, giving rise to a population activity matrix $\mathbf{X}^{full}$ **of size** $8 \times N_{rec}$. We then applied principal component analysis (PCA) to the data in two steps: first, reducing the dimensionality of X^{full} to the reduced-rank matrix X^{3D} **of** size 8×3; and second, reducing each location L separately (corresponding to a 4×3 input subset) to a X_L^{2D} matrix of size 4×2. The two principal components (PCs) calculated during the second step correspond to the directions of maximum variance in the data, and thus define the axes of the best-fit planes for the given location in the input (3D) space.

To calculate the angle θ between the two planes, we first applied a correction to the plane-defining vectors for both locations that put them in a common frame of reference. This was necessary because the vectors obtained from PCA correspond to the directions of maximum spread in the data rather than, for example, specific sides of the quadrilateral defined by the datapoints. If the directions of maximum variance differ between the two locations, this could influence the value of the angle between planes calculated. For example, imagine the colour-category datapoints form rectangles for both stimulus locations considered. However, the longer side for the rectangle in L1 corresponds to the line joining the 'yellow' and 'red' datapoints, whereas for L2, this is the shorter side of the rectangle. In this case, despite the data rectangles being parallel (as well as phase-aligned), the sets of plane-defining vectors obtained from PCA for both locations would differ in the ordering of the calculated vectors. As the angle between planes is based on calculating the plane using the cross-product (see equation 5.1), which is sensitive to the vector order, this would then result in the angle taking on a value of 180° rather than 0°. Thus, to make the metric insensitive to any distortions to the square geometry, we applied a correction to the set of vectors obtained from PCA. The first vector from the set was rotated to be collinear to a side of the data quadrilateral (giving rise to $\mathrm{PC1}_{\mathrm{L}_{\mathrm{corrected}}}$), and the second vector was rotated by the shortest angular distance to be orthogonal to the former ($\mathrm{PC2}_{\mathrm{L}_{\mathrm{corrected}}}$). After applying this correction, we calculated the normal n_{L} for each location L:

$$\mathrm{n}_{\mathrm{L}} = \mathrm{PC1}_{\mathrm{L}_{\mathrm{corrected}}} \times \mathrm{PC2}_{\mathrm{L}_{\mathrm{corrected}}} \tag{5.1}$$

The angle between the two planes was then be defined as:

$$\cos\theta = n_1 \cdot n_2 \tag{5.2}$$

$$\theta = \cos^{-1}(\theta) \tag{5.3}$$

Note that following the correction, an angle of 0° is only possible when the two planes are parallel and are rigid rotations and/or translations of one another. In contrast, parallel planes that are mirror reflections of one another and bowtie geometries would result in an angle of 180°.

To test for significant clustering of the plane angles, we used Rayleigh, v- and circular one-sample mean tests. All tests assume a null uniform distribution around the entire circle, but the range of $\cos^{-1}$ is restricted to [0,180°]. Therefore, we set the sign of each angle as the sign of the z-coordinate of n_1 in the 3-dimensional coordinate system obtained from the PCA analysis, thus extending the range to the full [-180°,180°] interval. Furthermore, all tests also assume a unimodal distribution as the alternate hypothesis. Therefore, in cases where the data distribution was bimodal (i.e. pre-cue angles, as well as the pre-post angle difference) or the angle estimates were first calculated for subsets of trials and then averaged across them (i.e., the cued/uncued angles for trials where location 1 was cued and trials where location 2 was cued, as well as the rotated/unrotated plane analyses), we first rectified the plane angles (to force the distribution to be unimodal) and then multiplied the individual values by a factor of 2 (to extend the possible range to [0,360°]).

Phase alignment was defined as the rotation angle ψ that minimised the distances between corresponding points from the two locations. Planes with absolute θ values greater than 90° (which correspond to mirror reflections) were excluded from the analysis, as were bowtie geometries. For all other geometries, we projected the entries from X^{3D} onto their

corresponding planes, before forcing the planes to be co-planar by rotating plane 2 by -θ°. We then applied the orthogonal Procrustes analysis. This method finds an orthogonal matrix R which most closely maps one matrix onto another (in our case, the sets of points corresponding to locations 1 and 2). Phase alignment angle ψ was defined as:

$$\psi = -\tan^{-1}\left(\frac{R_{1,0}}{R_{0,0}}\right) \tag{5.4}$$

To assess for significant phase alignment between planes, we performed a v-test with the hypothesised mean set to 0°.

2.3.2.4 Subspace alignment index (AI)

The alignment index between a pair of subspaces was calculated as described in (Elsayed et al., 2016). In short, we defined two matrices, $S_1$and S_2 (each of size $4 \times N_{rec}$), containing the binned activation patterns corresponding to the subspaces of interest. For example, when calculating the AI between the L1 and L2 subspaces during the pre-cue delay, the two matrices would correspond to the first and last B rows of $\mathbf{X}^{full}$**.** Next, we calculated the covariance matrices for S_1 and S_2 and performed eigenvalue decomposition. The alignment index between S_1 and S_2 is then given by:

$$\mathrm{AI}_{S_1,S_2} = \frac{\mathrm{Tr}\left(Q_{S_1}^{T}\,\mathrm{cov}_{S_2}\,Q_{S_1}\right)}{\Sigma_1^{n} Q_{S_2}(i)} \tag{6.1}$$

$$\mathrm{AI}_{S_2,S_1} = \frac{\mathrm{Tr}\left(Q_{S_2}^{T}\,\mathrm{cov}_{S_1}\,Q_{S_2}\right)}{\Sigma_1^{n} Q_{S_1}(i)} \tag{6.2}$$

$$\mathrm{AI} = \frac{\mathrm{AI}_{S_1,S_2} + \mathrm{AI}_{S_2,S_1}}{2} \tag{6.3}$$

where cov_{s_1} corresponds to the covariance matrix of S_1, and Q_{s_1} is the matrix containing its eigenvectors. AI was calculated for 2 and 3-dimensional subspaces.

For the cross-validated AI analysis of *Cued Geometry,* we used half of the data to calculate the AI values for the L1 subspaces between the two memory delays, repeated this step for the L2 subspaces, and compared the two values to define planes as *unrotated* and *rotated* according to their AI values. Using the other half of the dataset, we then once again calculated the AI, this time comparing the *unrotated* and *rotated* rings across time for all trained networks.

2.3.2.5 Uncued colour decoding

For each RNN model, we trained a set of 6 x 2 = 12 linear discriminant analysis (LDA) classifiers (one for each colour category pair and retrocue location) to perform pairwise colour-category discriminations. We first extracted the population hidden activity vectors at the last timestep of the post-cue delay for all stimuli and labelled the data according to the uncued stimulus on each trial. We then binned the labels into B colour categories, as described before. We trained the LDA classifiers in two-fold cross-validation – that is, for each of the 12 conditions, we split the corresponding data into two subsets. We trained and tested a given LDA classifier on the first and second subset, respectively. Then, we reinitialised it and used the reverse subset order for training and testing. The two test decoding accuracy scores for the classifier were then averaged to give a final score for the

condition. The test scores were further averaged across the 12 trained classifiers for each model prior to statistical analysis at the group level.

2.3.2.6 Colour discriminability

Hidden activity data from each memory delay was sorted into 4 conditions, according to the cued location: L1-cued, L2-uncued, L1-uncued and L2-cued. For each trial type (defined by the location of the cue), we first performed PCA to obtain the 3-dimensional coordinates corresponding to the colour category representations for each condition (i.e., the pre-cue, cued and uncued items). Colour discriminability index (CDI) was then defined as:

$$\mathrm{CDI} = \frac{|d_1 \times d_2|}{2} \tag{7}$$

where d_1 and d_2 are vectors corresponding the two diagonals of the data coordinate-defined quadrilateral. The surface area data was averaged across both locations. The pre-cue cued and uncued conditions were also averaged to create a single 'pre-cue' category, as the coding scheme was observed to be location- (and not cue-) based at that timepoint. We assessed for differences between the pre-cue, cued and uncued conditions with contrasts based on the results from Panichello and Buschman (2021).

2.3.2.7 Learning_dynamics

The plane fitting procedure and AI calculation was identical to that described above. We compared the training loss curves between models and observed that they mostly followed the same shape, characterised by a pronounced plateau period occurring at approximately half the initial loss value (see **Fig 2.4A**)**.** Therefore, we chose to compare AI values across

three training stages: pre-training, at the mid-learning plateau, and after the training had concluded. To find the epoch corresponding to the plateau for each model, we first smoothed the loss curve by applying a Gaussian filter (with standard deviation set to 4) and subsequently calculated its derivative with respect to training time. The epoch designated as the plateau timepoint was defined as the first local maximum in the derivative curve, constrained to exclude the early epochs where the loss value was within a 5% margin of the initial loss value. This constraint was necessary because all models showed a plateau period at the beginning of the training before the loss started to decrease exponentially.

2.3.2.8 Regression of representational geometry versus training latency

We trained N=100 models using the same hyperparameter settings as described above (in the section *Task structure and training procedures*) but using a different training stop procedure. To equate performance across the trained networks, we stopped training after the loss value averaged over the last epoch fell below the threshold of 0.0005 and noted the total number of training epochs needed to reach that timepoint by each network (time to convergence). Models were then evaluated on the test dataset. We calculated the subspace AI values for the three geometries of interest (*Cued*, *Uncued*, *Cued/Uncued*) using the hidden activity patterns from the last timepoint of the post-cue delay. The AI for the *Cued* geometry was computed by substituting $\mathbf{X}^{full}$ entries corresponding to the cued-L1 and cued-L2 activity patterns for $S_1$and S_2. The AI for the *Uncued* geometry was calculated using an analogous matrix where the data had been sorted and binned by the uncued stimulus colour. Finally, to compute the AI for the *Cued/Uncued* geometry, we first split the test dataset into two subsets according to the location of the retrocue. For each subset, we sorted and binned the trials into B = 4 bins by the cued colour, and then repeated this step using the uncued colour labels. We then averaged within these bins to calculate the mean hidden

activity patterns for the cued and uncued colours, which served as S_1 and S_2 for the AI analysis. The two AI scores (corresponding to the cued-L1/uncued-L2 and cued-L2/uncued-L1 subspaces) were averaged to obtain a final *Cued/Uncued* score for each network. The *Cued*, *Uncued* and *Cued/Uncued* AI values were entered as the predictor variables in a multiple linear regression model, predicting the time to convergence. The dependent variable was log-transformed prior to being used in the regression model.

2.3.3 Experiment 2

2.3.3.1 Task structure and training procedures

The overall task structure was the same as for Experiment 1, but we manipulated the lengths of the two delay intervals during training and test. The previously used training set was further expanded to include all combinations of delay lengths (2, 6, 7 and 9 cycles for both pre- and post-cue delay, drawn independently), resulting in a total of 8192 unique trials. The models (N = 30) were not subject to any noise ($\sigma = 0$), to investigate the potential effects of the variable delay length manipulation on the learnt task solutions in isolation. We trained the individual networks until the mean loss value from the last epoch crossed the threshold of 0.0005.

2.3.3.2 Behavioural analysis

We created 3 datasets, which differed in the length of the delay interval used. These corresponded to a delay length experienced during training and directly comparable with the first set of trained networks (7 cycles), as well as two novel delay lengths, chosen from within and outside the training range (4 and 10 cycles, respectively).

2.3.3.3 Cross-temporal decoding

For each trial timepoint, we trained a set of 12 LDA classifiers to perform pairwise discriminations between colour categories using the population hidden activity vectors. Data was labelled according to the cued stimulus (labels corresponded to 4 colour categories). Each set of classifiers was subsequently tested on an unseen subset of the data which included all trial timepoints. To assess the stability of memory maintenance dynamics for a given condition (fixed or variable delay), we computed the average off-diagonal delay cross-temporal decoding accuracy scores for all models and tested them against the chance decoding level (50%) with a one-sample t-test. To compare the stability of the delay code between models trained with fixed and variable delay lengths, we calculated the ratio between the off- and on-diagonal cross-temporal decoding accuracy scores from the two delays and compared this metric between the groups with an independent samples t-test.

2.3.4 Experiment 3

We trained models (N = 30 per task variant) on variants of the original task from Experiment 1, parametrically varying the length of the post-cue delay interval, whilst holding the overall trial length constant. The models were not subject to any noise, and we stopped training once the loss (summed over all trials in the last epoch) crossed the threshold of 0.0005.

2.3.5 Software

All statistical analyses described were implemented in python using the scipy.stats package (version 1.7.1), with the exception of the ANOVA, multiple linear regression and Bayesian

statistics analyses, which were performed in JASP (version 0.12.2). All networks were trained using the pytorch library (version 1.10.2).

2.3.6 Statistics abbreviations

M: mean

SEM: standard error of the mean

CI: confidence interval

2.4 Results

We report results from 3 simulations (denoted experiments). In Experiment 1, we explore the ability of recurrent neural networks to recreate the data described in (Panichello and Buschman, 2021). In Experiment 2, we adapt the retrocue paradigm to a new case with variable delay periods and explore how this affects the neural geometry in RNNs. Finally, in Experiment 3, we study why and when the geometry emerges.

2.4.1 Experiment 1

We trained recurrent neural network models (RNNs, N = 30 models with different initialisation and trial order) to perform a retro-cueing task comparable to that previously described in Panichello and Buschman (2021) (see **Fig 2.1A** and **B**). Each model was a vanilla RNN with a single recurrent layer, incorporating rectified linear unit (ReLU) nonlinearities, and a single decoding layer (see **Fig 2.1C** for a schematic of the network architecture). Each trial was composed of four successive inputs. Firstly, networks received two concatenated circular normal inputs (corresponding to distributed representations of the colour stimuli at two different locations, which we denote *colour-L1, colour-L2*), followed

by seven cycles of delay (*pre-cue delay*). This was followed by a one-hot input (the *retro-cue*) denoting which of the two stimulus locations was relevant on the current trial and then finally, a second memory delay (*post-cue delay*). An example trial is shown in **Fig 2.1B**.

The networks were trained using stochastic gradient descent to minimise the mean squared error (MSE) between their output and the stimulus colour shown at the cued location. Thus, we do not engineer the representations into the networks but allow them to form autonomously during optimisation. To bound performance, we added zero-mean Gaussian noise to the hidden activations of the models and trained until a plateau in the loss curve had been detected (see *Methods*). After approximately 220 ± 105 epochs the networks converged to a mean absolute angular error of 16.40 ± 0.94° (*M*±*SEM*). This was somewhat smaller than the value reported for monkeys (51.8°). This discrepancy could be seen to reflect other sources of noise (e.g., motor error during saccadic responding) that are present in the animals but not in our models but is otherwise not consequential for our findings (increasing the noise added to network units precluded convergence).

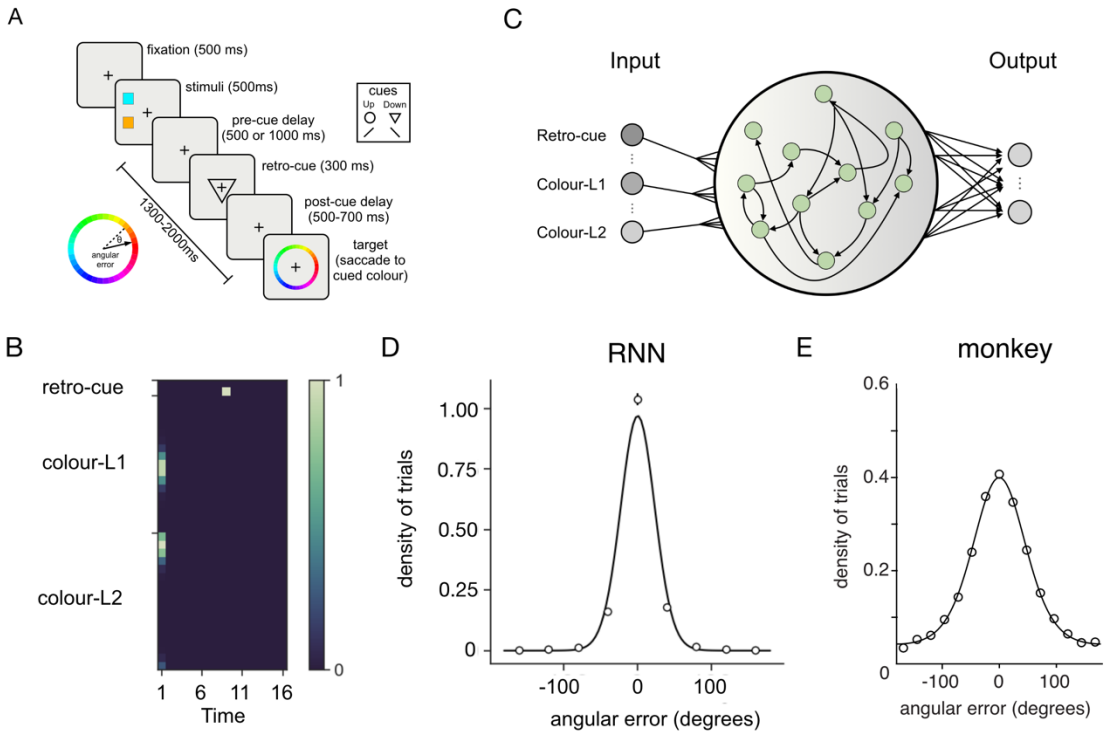


Fig 2.1. Experimental set-up. A. Retro-cueing task structure, taken from Panichello and Buschman (2021). **B.** Input patterns corresponding to an example trial. Colour denotes the activation value of an input unit. Note that inputs near the edges would wrap around because the space is circular **C.** Schematic RNN architecture. Green dots denote recurrent units. Only 3 input units and 2 output are shown (there were in 36 and 17 respectively). **D.** Distribution of errors made by the models at test (note errors were binned into 40 degree-wide bins for plotting; circles and vertical bars correspond to *M*±*SEM*) shown with a best fit von Mises (circular normal) density function (line). **E.** Behavioural performance of the monkeys from Panichello and Buschman (2021), retro-cueing task shown in blue. Lines correspond to mixture-model fits. Panels A and E are redrawn from the data from Panichello and Buschman (2021).

In Experiment 1, we first explored the representational geometry that related both cued and uncued items in the recurrent layer of the RNN, and compared it to that reported in macaque

LPFC (Panichello and Buschman, 2021). Noting a correspondence between the two, we then studied how this geometry emerged in the RNN over the course of learning, what connectivity patterns gave rise to it, and how it related to network behaviour, allowing us to form predictions for future studies.

2.4.1.1 Representation of cued items

To study the geometry of the cued memory representations formed in the pre-cue and post-cue delay periods, we focused on the "selection" task described in Panichello and Buschman (2021). The geometry was computed by applying dimensionality reduction to the features (neural activity over units or neurons) × conditions (each item binned by colour and cue) matrix and plotting the reduced reconstruction in 3D. We begin by alerting the reader to two frames of reference for this analysis. Firstly, we can ask about the relationship between *colour-L1* and *colour-L2* items when they are cued (*Cued Geometry* below) or uncued (*Uncued Geometry* below). Alternatively, we can ask about the relationship between the two items when one of them is cued and the other is not (*Cued/Uncued Geometry*).

We focused first on the *Cued Geometry*. In monkey LPFC, in both the pre- and post-cue periods, the neural representations arranged themselves in two rings, representing similarity in the circular colour space for each of the two respective locations (L1 and L2; Panichello and Buschman, 2021). During the pre-cue delay, the angle between the two planes that lay parallel to the interior of each ring was close to 90°. However, during the post-cue delay the two rings became aligned and phase matched, i.e., corresponding points in colour space were aligned on the two rings. A visualisation of this orthogonal-to-parallel transformation from the paper is shown in **Fig 2.2A**.

To study the memory representations created by the trained RNNs, we followed the same approach. We extracted hidden unit activity patterns corresponding to each trial type (defined by the unique colour-L1, colour-L2 and retro-cue combination) at the final timepoint of each delay period. For visualisation shown in **Fig 2.2B**, we binned the activation patterns corresponding to different stimulus colours into 4 equally spaced colour categories and collapsed the data across the uncued stimuli, resulting in a 200 x 8 (unit activity by conditions) matrix. Subsequently, we performed principal components analysis (PCA) to project the data into a 3-dimensional space.

As reported in Panichello and Buschman (2021), we observed that the input representations were structured by their similarity, with the "colour" representations for each stimulus location largely confined to 2-dimensional subspaces. Critically however, visual inspection suggested that the orientation of the location-specific planes in the 3-dimensional space differed between the two trial periods in a very similar way to that seen in monkey LPFC. During the pre-cue delay, the location-specific planes were orthogonal to one another, whereas during the post-cue delay, the two planes had a roughly parallel configuration. Moreover, the corresponding colour category representations were also phase-aligned, potentially permitting a common readout mechanism (see **Fig 2.2B**).

To formally check if the same linear readout could be used to discriminate between colours irrespective of the cued location, we trained binary LDA classifiers on hidden activity data from the last timepoint of the trial (in 2-fold cross-validation) and tested them on (1) the withheld trials on which the same location was cued (to check the decoding test accuracy) and (2) all the trials on which the other location was cued (to see if the classifier would cross-generalise to the other plane). Mean test decoding accuracy (1) was 99.35% (averaged

across individual classifiers, cross-validation folds, datasets (defined by the cued location) and models). It was significantly higher than chance (50%) across all models, as evidenced by the results of a one-tailed Wilcoxon signed-rank test: $W(29) = 465$, $p < .001$ (Cohen's d = 63.18). Mean cross-generalisation decoding accuracy (2) was 83.08% and was also significantly higher than expected by chance (one-sample one-tailed t-test: $t(29) = 19.59$, $p < .001$; Cohen's d = 3.58). The results of this analysis confirmed that a single linear readout could be used for both Cued colour planes.

To quantify the described geometry, we reduced the data to three dimensions for each network, estimated best-fitting planes for the datapoints corresponding to each cued location and calculated the angle θ between them in this low-rank space. Across all trained models, θ was clustered at the +90° and -90° marks in the pre-cue delay (**Fig 2.2C**). To test for the statistical significance of this effect, we rectified the data and performed a v-test with a hypothesised angular mean of +90° (see *Methods* for more details). This returned a significant result ($v(29) = 22.18$, $p < .001$), confirming that the pre-cue planes were indeed orthogonal. In contrast, in the post-cue delay, θ was significantly clustered around 0° (circular mean = -5.66°; v-test results: $v(29) = 24.09$, $p < .001$, see also **Fig 2.2D**). The mean absolute difference between pre- and post-cue θ was also significantly different from 0° (one-sample test for mean angle, $M = 93.2°$, 95% $CI = 69.66°, 116.74°$).

In their study, Panichello and Buschman (2021) report that the planes were phase aligned during the post-cue period. To quantify the phase alignment between the L1 and L2 datapoints in the RNN, we projected them onto their respective planes before rotating the planes to be coplanar. The phase alignment angle ψ was then obtained from a rotation matrix that most closely mapped L1 datapoints onto L2 datapoints. Importantly, ψ was only defined

for geometries that could not be classified as either mirror reflections or bowtie shapes. Given that the pre-cue planes were significantly orthogonal, we only report ψ for the post-cue delay, where it was defined for all models and significantly clustered around 0° (circular mean = 1.67°, v-test results: $z(29) = 29.57$, $p < .001$, $N = 30$). Thus, our results indicate that in the RNN, like in the monkeys, the post-cue planes were both parallel and phase-aligned.

We confirmed the above results with an analogous analysis using the subspace alignment index (AI) introduced in (Elsayed et al., 2016; see also *Methods*). This method projects the data from condition i into the subspace of condition j and calculates the amount of variance of i explained by the subspace of j. The AI metric takes values between 0 and 1, which correspond respectively to fully orthogonal and fully aligned subspaces. We calculated the AI for 2 and 3 dimensional subspaces and found that in both cases, the mean pre-cue AI was low (AI = 0.26 and 0.3, for 2 and 3 dimensions, respectively), indicating nearly orthogonal subspaces for the two locations. In contrast, the mean post-cue AI was significantly higher (more parallel) in both target dimensionalities (2D: AI = 0.63, paired t-test: $t(29) = 6.13$, $p < .001$, one-tailed, Cohen's d = 1.12; 3D: AI = 0.64; $t(29) = 5.98$, $p < .001$, one-tailed, Cohen's d = 1.09), indicating a significant increase in subspace alignment across the two delays (see also **Fig 2.2G**).

Lastly, we investigated how a given colour ring was oriented with respect to itself before and after the presentation of the retro-cue. This allowed us to study the nature of the transformation that the network had learned to perform. To parallelise the subspaces, the networks could either learn to rotate both pre-cue subspaces, or to rotate only one, keeping the other static. To establish which solution was utilised by the models, we performed the AI analysis in 2-fold cross-validation. This was necessary because the frame of reference of

location (e.g., left vs. right side) is arbitrary, and so if different networks learned to rotate left or right stimuli, averaging the data would obscure the results. Thus, we used half of the data to identify the *unrotated* and *rotated* locations (based on their AI values), and the other half to confirm whether there was a significant difference between them. We observed that the mean AI for the *unrotated* location (AI = 0.60 and 0.61 for 2 and 3D, respectively) was significantly higher than for the *rotated* location (2D: AI = 0.28, Wilcoxon signed-rank test $W(29) = 462$, $p < .001$, matched rank biserial correlation = 0.987; 3D: AI = 0.32; $W(29) = 465$, $p < .001$, matched rank biserial correlation = 1.0; see **Fig 2.2H**) and comparable to the AI for the post-cue subspaces (**Fig 2.2G**). We obtained similar results when we computed rotation of the between-plane angles using a similar cross-validated approach. In short, we used half the data to identify the *unrotated* and *rotated* locations (based on the absolute $\cos(\theta)$ value between the pre-cue and post-cue Cued planes) and the other half to confirm whether there were any significant differences in θ and ψ between them. Focusing first on θ, we rectified the rotated and unrotated angles (**Fig. 2.2I)** and calculated the difference between them. We found evidence for significant clustering in this difference (Rayleigh test: $z(29) = 16.59$, $p < .001$) with a mean of 24.93°, indicating that θ for the rotated plane was on average higher than the one for the unrotated plane. This is in agreement with the AI results from **Fig 2.2H**. With respect to the phase-alignment between the pre-cue and post-cue planes, we found the unrotated, but not the rotated ψ to be significantly clustered (Rayleigh test: $z(27) = 3.29$, $p = .036$ and $z(23) = 0.7$, $p = .503$, respectively), with the angular means of 10.99 ° and 10.54 °, respectively (see also **Fig 2.2J**). Together, these results imply that by convergence, the networks had learnt to rotate only one of the pre-cue subspaces, keeping the other in its pre-cue orientation. These findings diverge from the observations made in monkey LPFC, where the different task periods were associated with independent subspaces.

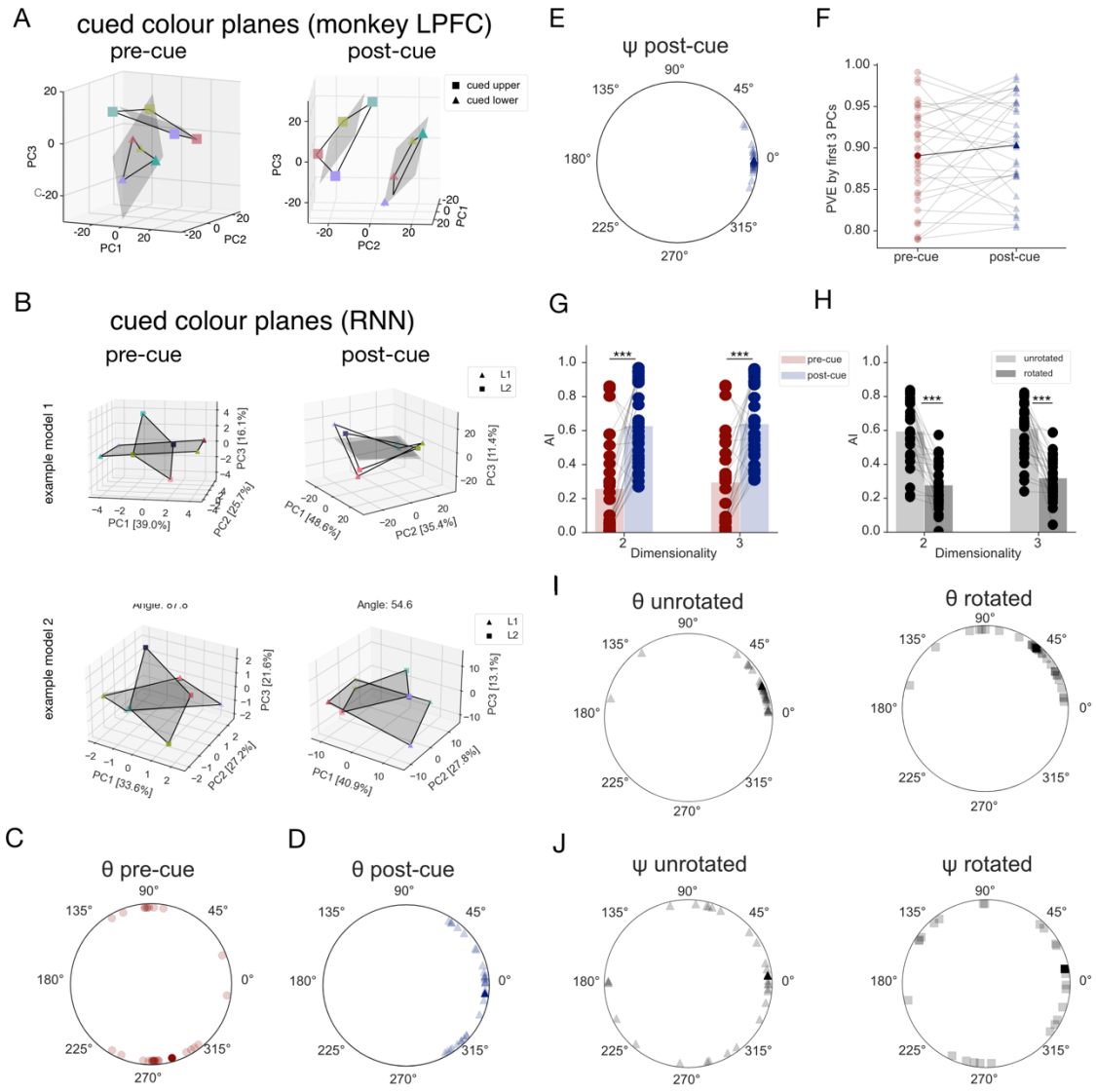


Fig 2.2. The geometry of cued mnemonic representations learnt by the RNNs. A. Visualisation of the pre- (*left*) and post-cue (*right*) geometry of selected (i.e., cued) items reported in Panichello and Buschman (2021). Population responses from LPFC binned into 4 colour categories (denoted by marker colour) for each stimulus location (upper, lower – denoted by marker shape). Data plotted in reduced dimensionality space, defined by the first 3 principal components (PCs) of the 8 location-colour pairs. Planes of best fit for each location shown as grey quadrilaterals. **B.** Analogous visualisation of the hidden activity patterns from two example RNN models for the pre-cue (*left*) and post-cue (*right*) delay. Two locations (L1, L2) correspond to triangles and squares, respectively. The percentage of total variance explained by each PC axis shown in square brackets. Chosen models (in this and subsequent figures) correspond to the ones with geometries qualitatively most and least

similar to the group-level average (in this case, the average post-cue geometry). Note that the principal components were calculated separately for each delay (and model), thus the PC axes shown, e.g. in the upper left (example model 1, pre-cue) panel are not the same as the ones shown in the upper right (example model 1, post-cue) panel. **C.-D.** Between-plane angles θ for the pre- (red dots) and post-cue delay (navy triangles) timepoints, respectively. Lighter and darker colours correspond to the values for individual models and grand averages, respectively. **E.** Phase alignment angles ψ between the cued planes in the post-cue delay. **F.** Proportion of total variance explained (PVE) by the first 3 PCs for the PCA models fit to activation patterns from individual networks. Values for individual models shown in lighter, and grand averages in darker colours. **G.** Subspace alignment index (AI) between location-specific planes during the pre- (red) and post-cue (navy) delay intervals. Individual models shown as points, bars correspond to the grand averages. AI reported for 2- and 3-dimensional subspaces **H.** AI for the *unrotated* (light grey) and *rotated* (dark grey) subspaces – for description, refer to the main text. Statistically significant contrasts denoted by asterisks (*: $p < .05$, **: $p < .01$, ***: $p < .001$). Panel A is reprinted with permission. **I.** Between-plane angles θ (rectified) for the unrotated (triangles) and rotated (squares) planes. **J.** Phase-alignment angles ψ for the unrotated (triangles) and rotated (squares) planes. Individual models and population mean shown as transparent and opaque markers, respectively.

2.4.1.2 Representation of uncued items – *Uncued Geometry*

The RNN thus seems to capture the cued stimulus representations seen in macaque prefrontal cortex. Next, we turned to investigate the *Uncued Geometry*, that is the relationship between items *colour-L1* and *colour-L2* when they were not signalled by the retrocue. First, we asked whether the RNNs retained any information at all about the uncued item in the post-cue period. In monkey LPFC, there was significant colour information about both memory items (quantified as the circular entropy of each neuron's response to colour) in all trial periods (Panichello and Buschman, 2021). This was the case even though the

monkeys were never probed on the uncued memory item and could have, in theory, discarded it after the cue presentation. We found the same to be true in the RNNs. For all networks, linear decoders trained to perform pair-wise colour discriminations of the uncued items in the post-cue delay showed significantly above-chance mean performance when tested on held out data (M = 98.16% decoding accuracy, Wilcoxon signed-rank test, $W(29) = 465$, $p < .001$; matched rank biserial correlation = 1.0).

We then set to find out if the uncued memory items also underwent a transformation from a feature-based representation to a more response-oriented format. In monkey LPFC, the uncued subspaces became slightly, but non-significantly, more parallel in the post-cue delay. Moreover, the two subspaces, unlike their cued counterparts, were not phase-aligned (Panichello and Buschman, 2021; **Fig 2.3A**). To visualise and quantify the geometry of these representations across our models, we followed the same steps as described above (in the section subtitled *representation of cued items*), except the data were averaged across uncued rather than cued locations (the configuration in the post-cue delay shown in **Fig 2.3B**). For each delay period, we calculated and compared the two location subspaces using the AI (**Fig 2.3D**). The mean AI in the post-cue delay was marginally higher than in the pre-cue delay (2D: pre-cue AI = 0.26, post-cue AI = 0.32, paired t-test: $t(29) = 1.36$, $p = .092$, one-tailed, Cohen's d = 0.25; 3D: pre-cue AI = 0.30, post-cue AI = 0.37, paired t-test: $t(29) = 1.77$, $p = .043$, one-tailed, Cohen's d = 0.32). It was also, however, significantly lower than the post-cue index for the cued subspaces (paired t-test: 2D: $t(29) = -6.74$, $p < .001$, Cohen's d = 1.23; 3D: $t(29) = -5.84$, $p < .001$, Cohen's d = 1.07). A complementary analysis of the angles formed by the two planes in the post-cue delay revealed that their orientation did not show any consistent trend across trained models (see **Fig 2.3C**, *left*). This was confirmed by the results of a Rayleigh test which failed to reject the null hypothesis of a uniform distribution

of plane angles around the unit circle ($z(29) = 0.06$, $p = .941$). Similarly, there was no significant phase alignment between the uncued planes (v-test results: $v(15) = 1.81$, $p = .261$, $N = 16$), and nearly 50% of the models developed geometries where the two planes were either mirror reflections of one another, or one of the geometries was twisted into a bowtie (non-planar) shape. Taken all together, these findings show that the two uncued subspaces are generally not transformed into a common frame of reference post-cue and their precise arrangement is variable between the trained models. Whilst one might be tempted to propose that the *Uncued geometry* is irrelevant for the choice, we shall see below that (perhaps surprisingly) it is reliably predictive of behaviour (see Experiment 3 below).

2.4.1.3 Representation of uncued items – *Cued/Uncued Geometry*

To successfully perform the task, the networks need to be able to prevent cross-interference between the two memory items (cued and uncued) across the entire trial duration, including the post-cue delay. In our next analysis, thus, we focussed on the relationship between cued and uncued items. We call this *Cued/Uncued Geometry* and note that this is different from analysing the relationship between each item (say, *colour-L1*) on those trials where it was cued relative to those where it is uncued. Instead, here we examine the relationship between *colour-L1* and *colour-L2* in the frame of reference of the cue.

It seems *a priori* plausible that, in order to minimise interference, the cued and uncued items are orthogonalized, as was observed in monkeys (**Fig 2.3E**). To test this prediction in our models, we calculated the angles between the cued and uncued subspaces and averaged them across the two trial types (defined by the location of the retrocue). In line with our expectation, we found that the distribution of angles across the trained networks was significantly clustered around 90 degrees (v-test: $v(29) = 10.31$, $p < .01$; **Fig 2.3G**, left). In agreement, average AI values between the cued and uncued subspaces were low (2D: M =

0.32, SEM = 0.03; 3D: M = 0.37, SEM = 0.03; see also **Fig 2.3G**, right) and comparable to the values reported for the Cued planes in the pre-cue delay (cf. **Fig 2.2G**, red bars).

Furthermore, as was reported for monkey LPFC, the visualisation of the geometry in **Fig 2.3F** suggested that the uncued colour representations were much more compressed (i.e., less discriminable) in the representational space than their cued counterparts. To quantify this effect, we calculated the area of the polygons formed by the datapoints in the 3D space and compared this colour discriminability index (CDI) across conditions with contrasts based on the results reported for monkey LPFC (Panichello and Buschman, 2021). We found that the mean colour discriminability in the post-cue delay was significantly higher for cued (CDI = 1294.1, arbitrary units) than uncued items (CDI = 45.46 Wilcoxon signed-rank test: $W(29) = 465$, $p < .001$, one-tailed, matched rank biserial correlation = 1.0; see also **Fig 2.3H**). Next, we tested the cued and uncued CDI against the pre-cue baseline. In accordance with previous reports from monkey LPFC (Panichello and Buschman, 2021) and human neuroimaging work (Sprague et al., 2016; Ester et al., 2018), the cued CDI was significantly greater than the pre-cue baseline (CDI = 17.3; Wilcoxon signed-rank test: $W(29) = 465$, $p < .001$, one-tailed, matched rank biserial correlation = 1.0). For the uncued item, our hypothesis stated that there would be no significant increase from the baseline, as was reported in Panichello and Buschman (2021). To test it, we performed a Bayesian non-parametric t-test. This returned a BF_{01} of 0.73, indicating anecdotal evidence for the alternative hypothesis of a difference in medians. For comparison, a two-tailed Wilcoxon signed-rank test was not significant ($W(29) = 149$, $p = .085$,matched rank biserial correlation = -0.36). These results suggest that to prevent interference between the cued and uncued items, the RNNs both increase the representational discriminability of the former, as well as orthogonalize the two items at the end of the post-cue delay.

To verify the results of the CDI analysis reported in **Fig 2.3H** whilst accounting for the noise in the colour representations, we performed an analogous decoding analysis. We trained binary LDA classifiers to discriminate between colour pairs in two-fold cross-validation. To quantify colour decoding in the pre-cue delay, classifiers were trained to discriminate between colours from a single location using the data from the endpoint of the pre-cue delay and tested on withheld trials from the same condition. Similarly, to quantify the cued item decoding accuracy, classifiers were trained to discriminate between cued colours on a subset of trials where e.g. location 1 was cued, and tested on the withheld trials from the same condition. We used the same approach to quantify the cross-validated decoding accuracy for the uncued item. We found that the pre-cue colour test decoding accuracy was near ceiling (M = 99.4% across all colour pairs, locations, cross-validation folds and models). We performed the same contrasts as described in the main text for the CDI analysis. Cued item test decoding accuracy was not significantly higher than the pre-cue baseline (M = 99.4%, Wilcoxon signed-rank test W(29) = 26, p = .917, matched rank biserial correlation r = .205), but it was significantly higher than the uncued item test decoding accuracy (M = 98.2%; Wilcoxon signed-rank test $W(29) = 465$, $p < .001$, matched rank biserial correlation r = -1). To test our hypothesis of no difference between the uncued and pre-cue item CDI, we performed a Bayesian paired samples t-test. We repeated the same analysis with the test decoding scores, which returned a BF_{10} of 922037.11, providing strong evidence for the alternative hypothesis of a significant difference between the pre-cue and uncued colour decoding, Overall, this pattern of results contrasts with those reported for the CDI metric, where we found evidence for a selective increase in the cued item CDI relative to the pre-cue baseline, and no change in the uncued item CDI. In contrast, the decoding results described here suggest no statistically significant increase in the decoding of the cued item

relative to the pre-cue baseline, and a drop in the uncued item decoding. It is however important to note that the colour decoding accuracy in the pre-cue delay was already very close to ceiling, thus limiting the scope for an increase in the decoding of the cued item in the post-cue delay.

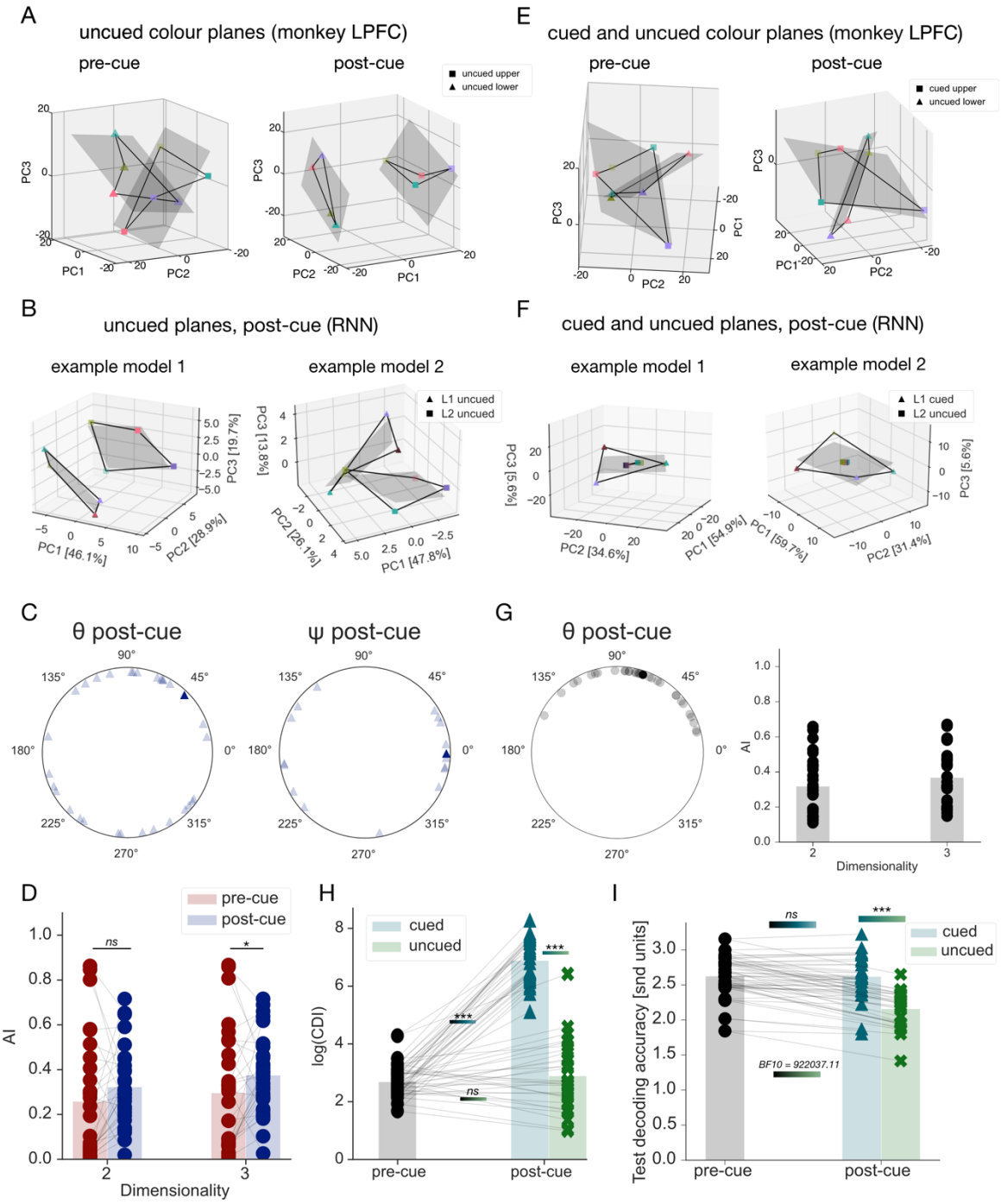


Fig 2.3. The geometry of the uncued item representations. A. Visualisation of the pre- *(left)* and post-cue (right) geometry of the uncued (non-selected) memory items reported in Panichello and

Buschman (2021). Population activity patterns from macaque LPFC were averaged across the cued items and binned into 4 uncued colour categories (denoted by marker colours). Data from each delay period was plotted in reduced dimensionality space, degined by the first 3 PCs of the 8 location-colour pairs. Planes of best fit for each location shown as grey quadrilaterals. All conventions as described in **Fig 2.2A. B**. Hidden activity patterns for the uncued items, from two example models visualised in a 3-dimensional space. All conventions as in Fig **2.2B. C.** Between-plane angles (θ, *left*) and phase alignment angles (ψ, *right*) between the two uncued colour planes in the post-cue delay. Individual models and population mean shown as transparent and opaque markers, respectively. **D.** Alignment index for the uncued subspaces in the pre- and post-cue delays. Note that prior to the presentation of the retro-cue, the coding format is location based. Therefore, the pre-cue data is the same as shown in **Fig 2.2E. E**. Visualisation of the cued and uncued memory items on trials where the upper location was cued reported in Panichello and Buschman (2021). Population activity patterns from macaque LPFC were averaged across the uncued items to calculate the cued item representations, and across cued items for the uncued representations, prior to binning into 4 colour categories each. Data from each delay plotted in the reduced dimensionality space, corresponding to the first 3PCs. **F.** Visualisation of the hidden activity patterns for cued and uncued items in the post-cue delay, on trials where L1 was cued. Data from two example models. Note the plane for L2 is severely compressed and thus hard to see. **G.** *Left*: Angles θ between cued and uncued planes in the post-cue delay, averaged across the two retro-cue locations. Note angles have been rectified. Values for individual models shown in grey, average in black. *Right*: Complementary AI values between the cued and uncued subspaces, averaged across the two retro-cue locations. AI calculated for 2- and 3-dimensional subspaces. Black circles correspond to values for individual models, grey bars denote the means across all networks. **H.** Colour discriminability index (CDI) for the pre-cue subspaces (averaged across both locations, black circles) and cued and uncued subspaces in the post-cue delay (blue triangles and green crosses, respectively). **I.** Results of a decoding analysis complementary to the CDI analysis reported in H. Decoding accuracy (expressed as units of standard normal distribution, snd) for the pre-cue subspaces (averaged across both locations, black circles) and cued and uncued subspaces in the post-cue delay (blue triangles and green crosses,

respectively). Bar height corresponds to the mean across all models. Significance and Bayes factor values correspond to the results of contrasts reported in the main text.

2.4.1.4 Learning dynamics and connectivity

Thus far, we have characterised the representational geometry formed by the RNNs and compared it to that observed in monkey LPFC. Next, we exploited the learning dynamics of the RNN to ask how the observed geometry emerged over the course of training.

Neural networks showed stereotypical learning dynamics characterised by a pronounced plateau period at approximately half of the initial loss value, buttressed by steeper declines both early and late (see **Fig 2.4B** for an example). We therefore decided to track the representational changes in the networks at three timepoints – at initialisation (*untrained*), the mid-learning plateau (*plateau*) and after convergence (*trained*). The timepoint used as the mid-training plateau was designated as the first local minimum in the derivative of the training loss with respect to time (for more details, refer to *Methods*). We quantified the *Cued Geometry* across the different training stages by calculating the angle θ and the AI at the end points of the memory delays. The results of this analysis are shown in **Fig 2.4D**. *Untrained* networks propagated the orthogonal inputs throughout the pre-cue delay without changing their relative arrangement, as evidenced by an AI value of approximately 0 (**Fig 2.4D,** *top right***)** and θ clustering around ±90°. This is expected, given that inputs to the network are orthogonal. After the presentation of the retrocue, both the AI and θ remained close to their pre-cue values. By the mid-training *plateau*, the networks had learnt to parallelise the two planes in the post-cue delay. Interestingly, however, this coding format also extended into the pre-cue delay, as evidenced by an AI value of close to 1 and θ clustering around 0°. As described in the previous section (*Representation of cued items*),

the *trained* networks settled on an orthogonal plane configuration in the pre-cue delay, which was transformed into a parallel arrangement after the presentation of the retrocue. Thus, the network first learned to parallelise the inputs entirely, before gradually learning to keep them separate before the retrocue appeared.

Next, we asked how the retro-cue triggered the representational changes observed in the recurrent layer. To address this, we analysed the connectivity patterns that were acquired over the course of training between the retrocue input units and the recurrent nodes. For each model, we calculated the correlation between the weights from each cue unit and the recurrent units (weights from an example model shown in **Fig 2.4C**). Consistent with other recent observations, we found that the models learnt opposing weight patterns for the two retro-cue inputs (Flesch et al., 2021; Russin et al., 2022). More specifically, the recurrent units that had positive connections weights with the first retro-cue unit had negative connection weights linking them with the other cue unit, and vice versa. Across all neural networks, this manifested as a significantly negative correlation coefficient (one sample t-test: $t(29) = -68.17$, $p < .001$, Cohen's d = -12.45). This negative correlation, in combination with the ReLU nonlinearities, ensures that hidden units that have inputs from one context units have none from the other and vice versa; this effectively creates two subpopulations of units which code for the context provided by the two retrocues.

Lastly, we investigated the relationship between the representational geometry observed in trained networks and their behavioural timecourse. To this end, we asked whether the orientation of the *Cued*, *Uncued* and *Cued/Uncued* subspaces (as quantified with the AI metric) was predictive of the training latency (time to convergence; see *Methods*) across all trained networks. To increase the power of this analysis (which unlike other analyses

described here relies on correlations over the network population) we trained N = 100 models until the training loss fell below a common threshold value (see *Methods*). We then regressed the three AI values against the total number of training epochs. The regression model explained a significant proportion of the variance in training latency (R^2 = .28, F(3,96) = 12.29, p < .001). Out of the three predictor variables, to our surprise only the *Uncued Geometry* showed a significant and positive relationship with the number of training epochs (ß = 0.91*, t(96)* = 4.48, p < .001). This means that the more parallel the *Uncued geometry* in the post-cue period, the longer the network took to train. By contrast, the AI between the (i) *Cued* subspaces and (ii) *Cued/Uncued* subspaces did not contribute significantly to the model (*Cued*: ß = -0.125*, t(96)* = -0.374, p = .374; *Cued/Uncued*: ß = 0.17*, t(96)* = 0.878, p = .382; see also **Fig. 2.4E**). We were not expecting this result, which is further discussed below.

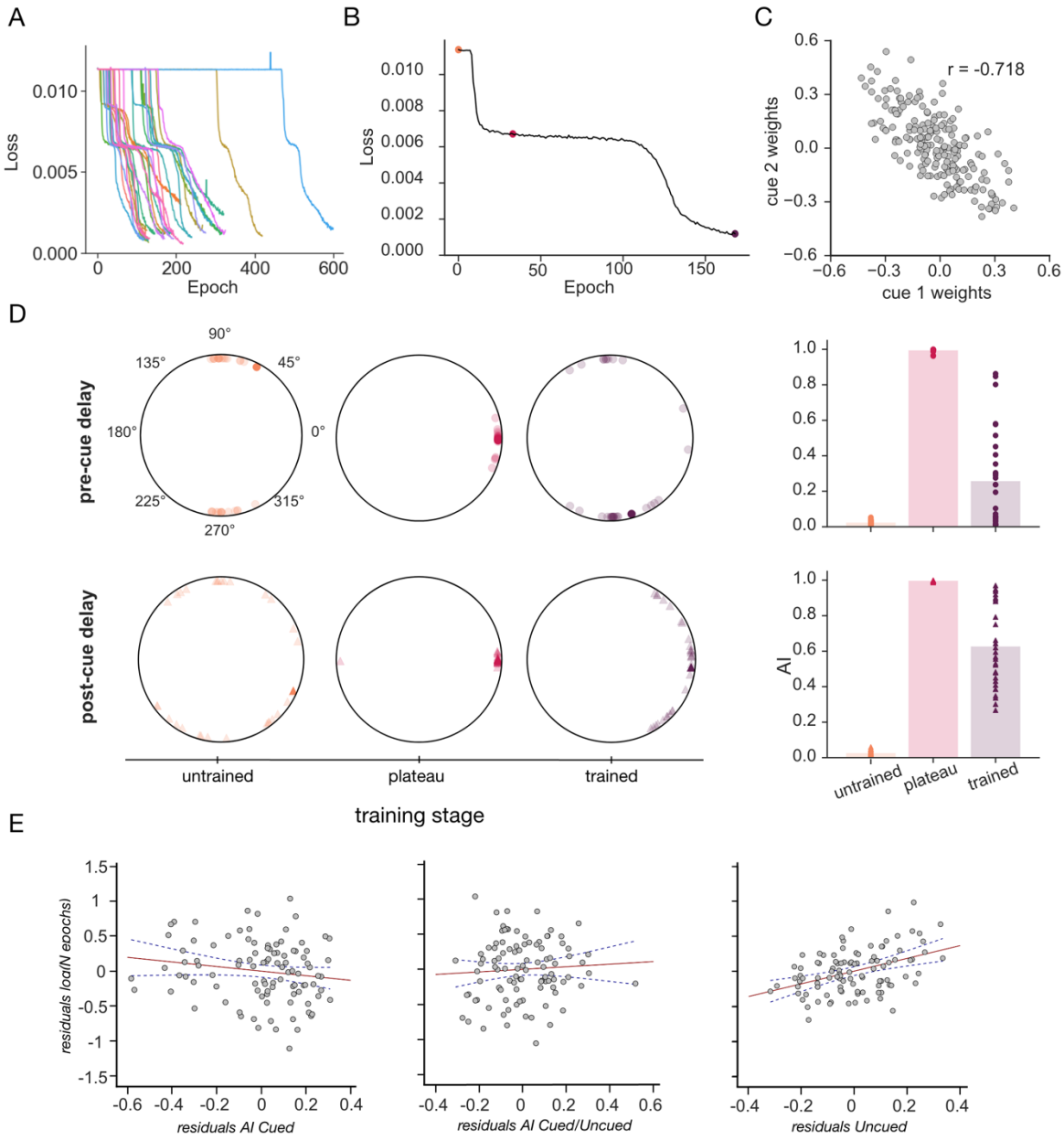


Fig 2.4. Learning dynamics. A. Training dynamics of the networks from Experiment 1. Training loss values across all epochs, colours correspond to individual networks. **B**. Training loss of an example RNN model plotted against training epochs. The three dots correspond to the timepoints chosen for the subsequent analysis: untrained (orange), plateau (magenta) and trained (purple) stages, respectively. **C.** Scatterplot of the weights between the two retro-cue input units and recurrent units for an example model. **D.** Comparison of the plane angles θ and AI between the cued planes across different training stages for the pre- (top row) and post-cue (bottom row) delays. *Left*: Plane angles at the untrained (orange), plateau (magenta) and trained (purple) stages. Values for individual

models show in transparent colours, circular mean in opaque. *Right:* AI at the same training stages. Mean across models depicted as bars with individual datapoints overlaid. **E.** Partial plots showing the log of training epochs (y-axis) versus the three representational geometry metrics used as regression predictors (left to right: AI *Cued*, AI *Cued/Uncued* and AI *Uncued*). Values for individual networks shown as grey dots, regression lines of best fit in red alongside the 95% slope confidence intervals in navy dashed lines.

2.4.2 Experiment 2: Exploring the effects of variable delay periods.

In the simulations described thus far, we trained networks with a fixed number of delay cycles. In biological systems, maintenance processes can be successful even where the delay period length is unknown, and these conditions typically give rise to the formation of more stable dynamics (e.g. attractor states) in RNNs (Orhan and Ma, 2019) and mice (Inagaki et al., 2019). We thus asked whether (i) the networks could perform the task under novel delay period lengths and (ii) whether the representations they formed were dynamic or stable in nature, as well as how these aspects depended on whether the delay period during training was of variable or fixed duration. Lastly, (iii) we examined whether the geometry of cued item representations differed depending on the type of memory maintenance mechanism used by the network.

We sampled delay period durations randomly during training and test, so that networks could not learn to anticipate the precise delay length on any given trial. In this setting, we tested performance under three different conditions: a delay length previously experienced during training, as well as novel in- and out-of-training-range lengths. We found that networks trained with a fixed delay length could not perform the task with novel delay

lengths – their performance, as measured by the mean absolute angular error rose from 16.37±0.94° (*M*±*SEM*) to 87.06±2.82° and 81.30±5.74° when evaluated on the 3 datasets, respectively (see also **Fig 2.5C**). In contrast, task performance in models trained with variable delay durations seemed to be less affected by the precise task timings (*M*±*SEM* = 7.72±0.22°, 36.46±1.94° and 18.60±2.01°, respectively; note that perhaps surprisingly generalisation was better to novel examples outside than inside the training set). Note the overall performance difference between the fixed and variable delay networks is due to a combination of different noise levels and training stop procedures used (for more details, refer to the *Methods* section).

Next, we studied the maintenance mechanisms learned by the networks trained with fixed and variable delay periods, using a cross-temporal generalisation approach. We trained LDA classifiers to discriminate between colour pairs using hidden activity patterns from a single timepoint and tested their performance on all the other trial timepoints (**Fig 2.5A** shows the cross-temporal generalisation scores averaged across models). We observed that neural networks trained with variable- and fixed-length delay periods both tended to maintain information across the memory delays using a temporally stable code. Accordingly, the mean off-diagonal delay cross-temporal generalisation scores were significantly higher than chance (0.5) in both groups (variable-length delay group: $M = 0.62$, one-sample t-test results: $t(29) = 28.88$, $p < .001$, Cohen's d = 5.27; fixed-length delay group: $M = 0.59$, one-sample t-test results on log-transformed data: $t(29) = 4.37$, $p < .001$, Cohen's d = 0.80; **Fig 2.5B**). However, as previously reported (Orhan and Ma, 2019), the temporal stability of the delay-period code, as indexed by the ratio between the off- and on-diagonal delay decoding scores, was higher in the variable- than fixed-length delay group ($M = 0.61$

and 0.59, respectively) although this was only barely significant (independent sample t-test on log-transformed data: t(58) = 1.74, p = .044, one-tailed, Cohen's d = 0.32).

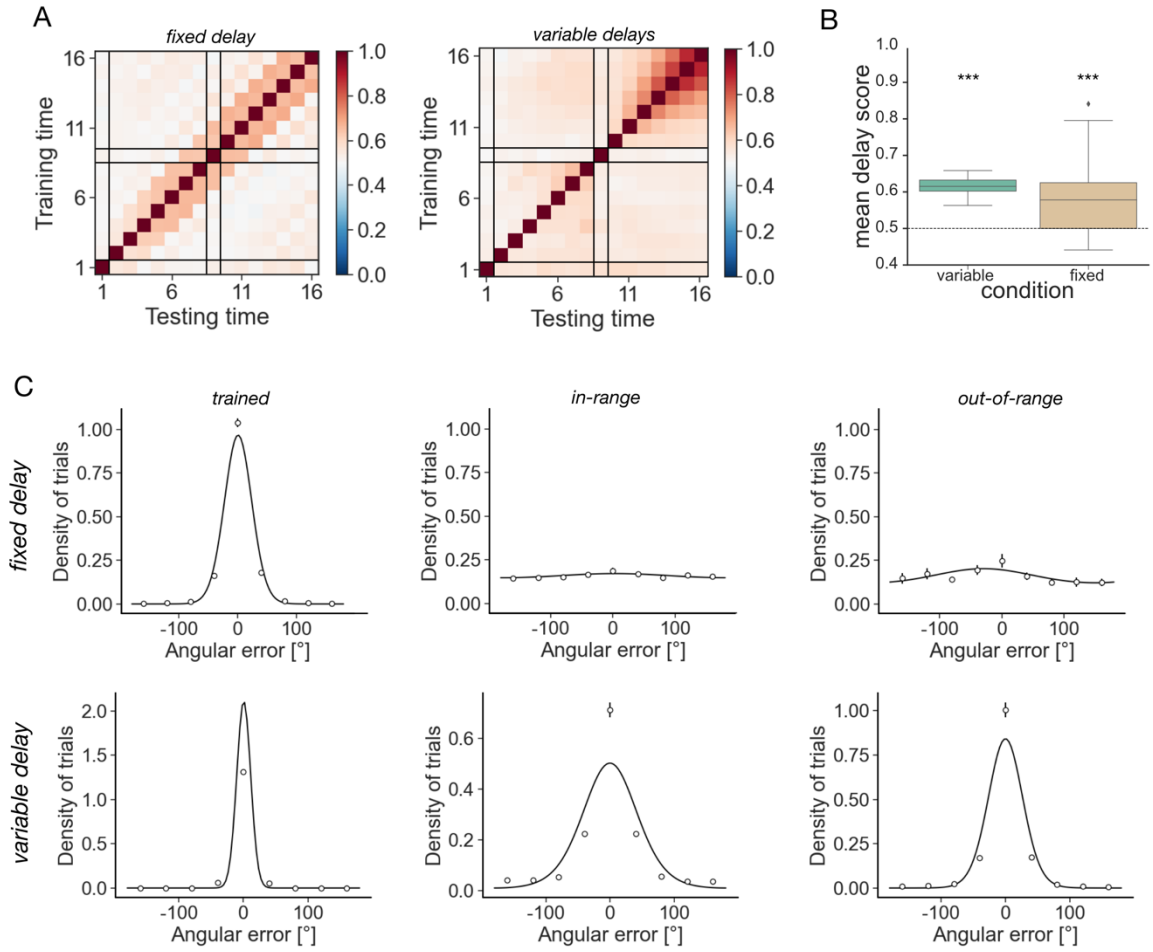


Fig 2.5. Maintenance mechanisms. A. Cross-temporal generalisation scores for decoders trained to discriminate between colour pairs. Data has been averaged across all trained networks. Black lines indicate the junctions between task events – stimulus presentation, pre-cue delay, retro-cue, and post-cue delay. *Left*: Average scores for models trained with a fixed-length delay interval. *Right*: analogous plot for models trained with variable delay lengths. Note that the networks show a stable (cross-generalisable) memory code between the 10 and 16 cycle mark, which covers the range of temporal variability experienced during training (with the post-cue delay starting at the 10 cycle mark and lasting until between the 11 and 18 cycle marks). **B.** Boxplots showing the distribution of mean cross-temporal decoding accuracy scores, averaged across the two memory delays, for all models. Variable delay interval length condition shown on the left in green, fixed delay length

condition on the right in sand. Asterisks denote the results of the one-sided one-sample t-tests against the chance (50%) decoding level, *** corresponds to $p < .001$ **C.** Angular error density plots for the 3 different test conditions: datasets with the delay interval length used during training (*trained*), a previously not experienced delay length that fell within the training range (*in-range*), and a novel delay outside of the training range (*out-of-range*). First and second rows correspond to the data from the networks trained with a fixed delay and variable delay intervals, respectively. Datapoints correspond to *M*±*SEM* across all models (binned into 40 degree wide bins), shown with the best von Mises fit (solid line).

Having confirmed that, compared to the fixed delay length networks, models trained with variable delay lengths learnt a task solution that generalised better across different timescales and showed more temporally stable memory maintenance dynamics, we analysed their hidden activity patterns to verify that our previous findings replicated under the new conditions. We first focused on the *Cued Geometry* and observed that it matched the one previously described for the fixed-delay networks in Experiment 1. Across all trained models, the angles θ between the two cued subspaces were significantly clustered around the 90° mark in the pre-cue delay (v-test results: $v(29) = 27.7$, $p < .001$) and 0° in the post-cue delay (circular mean = 2.73°; $v(29) = 22.27$, $p < .001$, see also **Fig 2.6A**). The mean absolute difference between pre- and post-cue angles was also significantly different from 0° (one-sample test for mean angle, $M = 103.34°$, 95% $CI = 75.39°, 131.28°$). Furthermore, the Cued planes were also significantly phase-aligned in the post-cue delay (mean $\psi = -5.44°$; v-test results: $v(28) = 25.09$, $p < .001$; note one model was excluded from this analysis due to forming a mirror image geometry with $\theta = -105.21°$) . This pattern of results was further confirmed by the AI analysis. The mean pre-cue AI (AI = 0.09 and 0.11 for 2 and 3 dimensions, respectively) was significantly lower than the mean post-cue AI (AI = 0.13 and 0.14, respectively) in both target dimensionalities (one-tailed one-sample t-test: $t(29) = 4.72$,

$p < .001$, Cohen's d = 0.86; $t(29) = 2.94$, $p = .003$, Cohen's d = 0.54). We note that although the differences between pre- and post-cue delays are qualitatively similar, these values are overall lower (i.e., planes are more orthogonal) than for Experiment 1. However, upon further investigation, we found that the results resembled closely those reported for Experinent 1 when networks were tested on trials with the delay interval length set to the maximum value experienced during training (9 cycles, see Supplementary Information for Chapter 2 for more details).

Next, we repeated the *rotated* / *unrotated* Cued plane AI analysis (shown for Experiment 1 in **Fig 2.2H**) found the results to also be in agreement with those reported for Experiment 1 (fixed delay length networks). More specifically, the mean AI for the rotated plane was significantly lower (2D: AI = 0.24, 3D AI = 0.27) than for the unrotated plane (2D: mean AI = 0.49, paired sample t-test: t(29) = -8.03, p < .001, Cohen's d = -1.47, 3D: mean AI = 0.50, paired sample t-test: t(29) = -7.78, p < .001, Cohen's d = -1.42; see also **Fig 2.6B**). The same analysis with plane angles θ revealed that the mean for the rotated plane is ~54° whereas that for the unrotated plane is close to zero. This difference was statistically significant across networks (Rayleigh test, z(29) = 22.48, p < .001). With respect to the phase-alignment between the pre-cue and post-cue planes, we found both the unrotated and rotated ψ to be significantly clustered (Rayleigh test: z(25) = 25.56, p < .001 and z(24) = 23.64, p < .001, respectively), with the angular means of 1.77° and -1.25°, respectively (see also **Fig. 2.6C** below).

Lastly, we characterised the *Uncued* and *Cued/Uncued* geometries (see **Fig 2.6D** and **E**, respectively). We found that the angle θ between the two uncued planes in the post-cue delay was significantly clustered (Rayleigh test: $z(29) = 24.55$, $p < .001$) with a mean

direction of -4.82°. The Uncued planes were also significantly phase-aligned, as evidenced by the results of a v-test on ψ ($v(27) = 22.7$, $p < .001$). These results imply that similarly to the *Cued* planes, the *Uncued* planes also underwent a transformation that placed them in a common frame of reference in the post-cue delay. With regards to the *Cued/Uncued* geometry, the distribution of angles across the trained networks was significantly clustered around 90 degrees (though with some variability across networks, v-test: $v(29) = 24.54$, $p < .001$; **Fig 2.6E**, *left*) and analogous AI values between the cued and uncued subspaces were low (2D: M = 0.13, SEM = 0.01; 3D: M = 0.16, SEM = 0.01; see also **Fig 2.6E**, *right*). Thus, like was observed for networks from Experiment 1, the *Cued* and *Uncued* planes had an orthogonal arrangement. Overall, these results imply that, despite a difference in the temporal stability of the delay-period code, both models trained with fixed as well as variable delay lengths converged to a highly similar solution to the retrocueing task (as evidenced by comparable *Cued, Uncued* and *Cued/ Uncued* geometries).

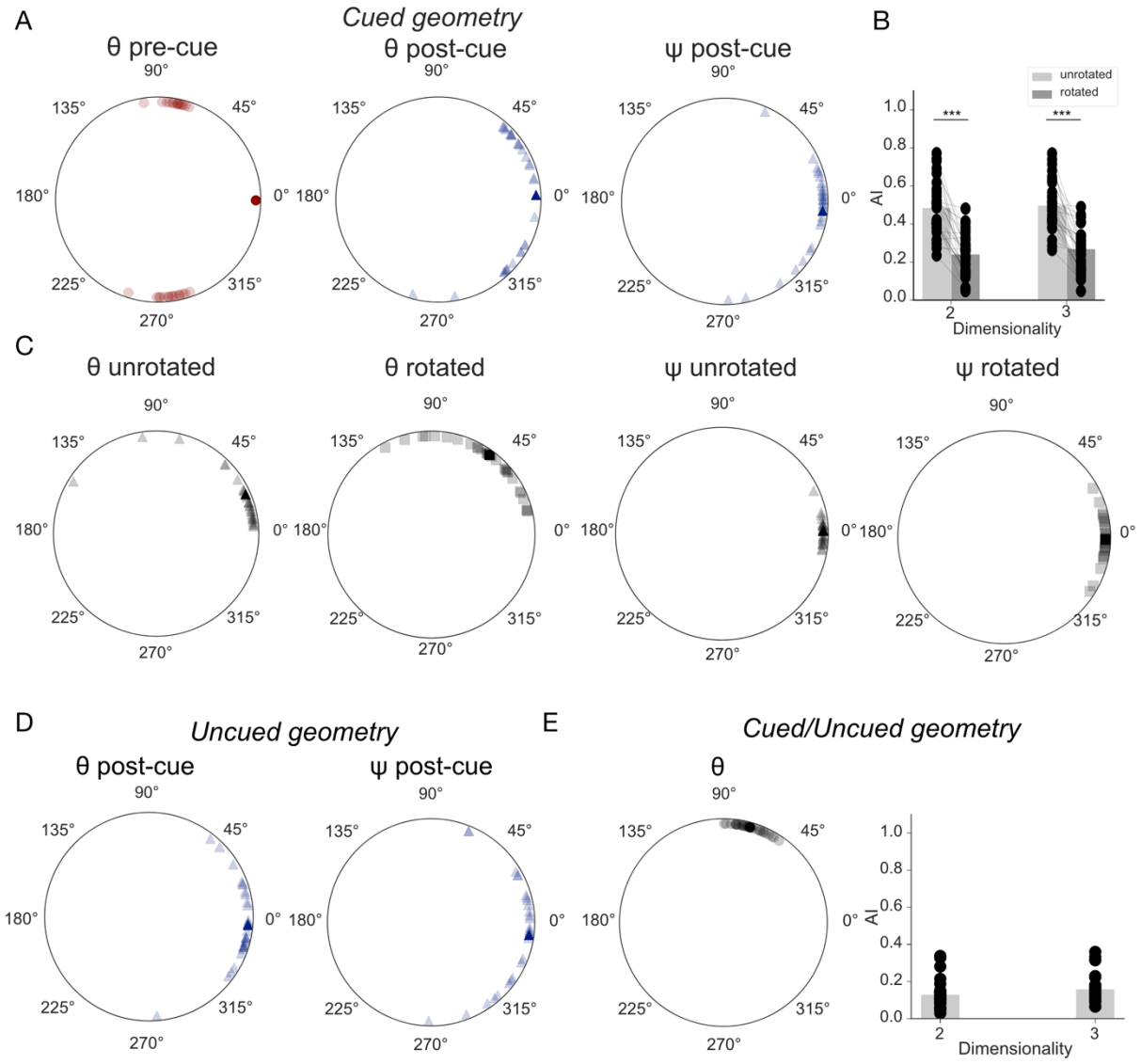


Fig 2.6 Neural geometry formed by networks trained with variable delay lengths. A. Plane angles θ and phase alignment angles ψ between the cued subspaces. From left to right: θ in the pre-cue delay (red dots), θ in the post-cue delay (navy triangles), ψ in the post-cue delay (navy triangles). Values for individual models and grand averages correspond to transparent and opaque markers, respectively. **B.** AI for the *unrotated* (light grey) and *rotated* (dark grey) subspaces. Significance values correspond to the results of a paired t-test. **C**. Analogous plane angles θ and phase alignment angles ψ for the *unrotated* (triangles) and *rotated* (squares) planes. Values for individual models and grand averages shown as transparent and opaque markers, respectively. Note θ values have been rectified. **D.** Angles θ and ψ between the two *Uncued* planes in the post-cue delay interval. All conventions as in panel **A**. **E**. Angles θ (rectified, *left)* and AI values (for 2- and 3-dimensional

subspaces*, right*) for the *Cued/ Uncued* planes in the post-cue delay, averaged across the two cued locations. Plotting conventions as in panels **C** and **B** for angles and AI values, respectively.

2.4.3 Experiment 3: Parallel geometry for more robust maintenance

Thus far, we have observed that RNNs trained on the retro-cueing task learned to transform the cued item information into a common frame of reference after the retrocue. Whilst this representational format affords the re-use of a common readout mechanism across all trials (i.e., irrespective of the retro-cue identity), it is not the only solution to the task. In fact, we observed that in other variants of the task (where the network first receives a probabilistic retrocue input and a deterministic probe - this paradigm is explored in Chapter 3), networks trained with no post-probe delay did not develop this geometry, and instead maintained the cued item information in orthogonal planes throughout the entire trial. Given this observation, we reasoned that the parallel plane format might be beneficial when there is a pressure to maintain the cued information in memory after the retro-cue (Myers et al., 2017a). To test this hypothesis, we trained models (as before, N = 30 per task variant) on variants of the original task from Experiment 1, except that now we parametrically varied the length of the post-cue delay interval, whilst holding the overall trial length constant (see **Fig 2.7A**). We then calculated the AI between the cued subspaces prior to and post retro-cue presentation and compared them between the different task variants (**Fig 2.7B,** red dots and line). We found that, in agreement with our previous results, all models maintained the cued item information in orthogonal subspaces before the retro-cue presentation, regardless of the task timings.

However, only models trained with longer post-cue delay lengths underwent the previously noted transition into a parallel subspace format prior to producing a response, whilst networks trained with shorter delays learnt to read-out cued colour information from the low-dimensional orthogonal subspaces (**Fig 2.7B,** blue dots and line). To test the statistical significance of this finding, we compared the pre- to post-cue difference in AI with a Kruskal-Wallis test, which showed a significant difference in this metric across the different post-cue delay length conditions ($H(7) = 199.7$, $p < .001$). We then followed up with custom contrasts comparing the pre- and post-cue AI at each level of the post-cue delay length factor. This revealed no significant differences with delay lengths below 3 cycles ($t(232) = 0.27$, $p = .789$; $t(232) = -1.16$, $p = .249$; $t(232) = -.61$, $p = .546$, in increasing order) and a significantly higher post-cue AI otherwise ($t(232) = 5.04$, $p < .001$; $t(232) = 14.53$, $p < .001$; $t(232) = 24.3$, $p < .001$, $t(232) = 29.52$, $p < .001$, $t(232) = 33.36$, $p < .001$, in increasing order). Note that the degrees of freedom here are computed from the ANOVA (8 conditions x 30 models = 240 datapoints, see methods).

Lastly, we examined the training time of networks (quantified as the number of training epochs completed until loss threshold was reached, NETC) from Experiment 3 and found it to differ depending on the length of the post-cue delay (one-way ANOVA with factor post-cue delay length (8 levels) with log-transformed NETC as the dependent variable, $F(7,232) = 50.74$, $p < .001$, $\eta^2 = 0.61$). A linear contrast confirmed that the training time declined alongside the post-cue delay length ($t(232) = -18.66$, $p < .001$; see also Fig. **2.7C**). These results suggest that networks using the orthogonal (short post-cue delay) solution need to complete the most training iterations to converge, as compared to the networks using parallel plane solutions (trained with long post-cue delays).

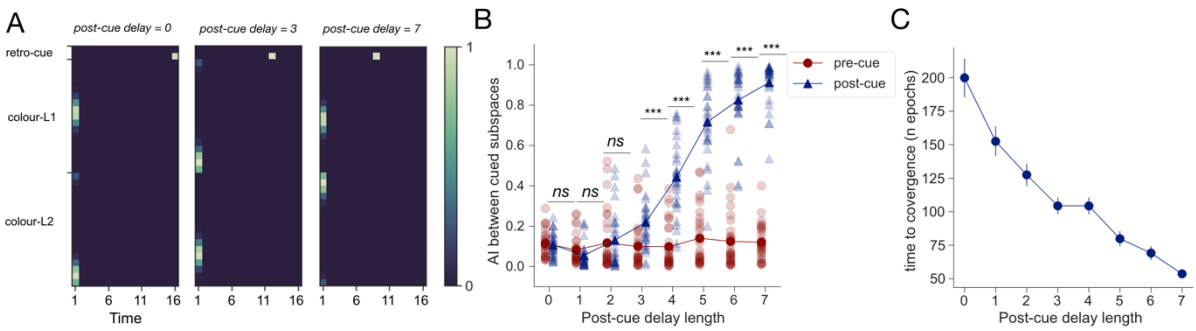


Fig 2.7. Comparison of cued item geometry across networks trained under various post-cue maintenance pressure conditions. A. Example trials for a subset of the parametric family of retro-cueing tasks used. Overall trial length was constant, and the ratio between pre- and post-cue delay lengths was varied. Conventions as in **Fig 2.1B**. **B.** Comparison of cued subspace AI during the pre- (red circles) and post-cue (blue triangles) intervals between networks trained with different post-cue delay lengths. Asterisks denote the significance levels of the contrasts described in the main text: ***: $p < .001$, **: $p < .01$, *ns*: $p >= .05$. **C.** Time to convergence (number of training epochs, y-axis) for models trained with different post-cue delay lengths (x-axis). Dots correspond to the means across models, error bars to SEM.

2.5 Discussion

In this Chapter, we describe a detailed study of the short-term memory representations that arise in RNNs during a retro-cueing task. We report that these are similar to those previously described in Panichello and Buschman (2021) for monkeys. In macaque LPFC, it was reported that prior to the presentation of a retrocue, information about individual memory items was maintained in short term memory along orthogonal axes of the neural state space. In contrast, following the retrocue, representations of the prioritised items were rotated into a common subspace. As seen across the three experiments reported in this paper, the RNN reliably captured this orthogonal-to-parallel transformation when the retrocue was followed by a memory delay with several timesteps. Building on this representational correspondence, we then characterised how and why the described geometry emerged, in

addition to revealing its mechanistic basis, thus providing testable predictions for future neurobiological studies.

At first glance, the orthogonal geometry in the pre-cue delay might seem a trivial finding, given that the information about the two stimuli is supplied to the RNN via independent (and uncorrelated) input channels. Accordingly, this geometry is present even in naïve networks (**Fig 2.4C**). However, during the pre-cue delay, over the course of training the networks first learn to arrange both mnemonic representations into parallel neural subspaces before converging back onto the orthogonal geometry. This coding format seems well suited to satisfy the computational demand of the task in the pre-cue delay – namely, the minimalization of interference between multiple memory items. This interpretation also fits well with other findings from the literature. For example, a recent sequential working memory study reported that information about individual stimuli from a sequence was also maintained in orthogonal neural subspaces in macaque LPFC (Xie et al., 2022). Similar observations have also been made in other brain regions, species and task (Flesch et al., 2021; Libby and Buschman, 2021; Nieh et al., 2021; Xie et al., 2022) as well as in neural networks (Flesch et al., 2018, 2022; Russin et al., 2022), suggesting that factorising representations into orthogonal subspaces to prevent cross-interference is a general coding principle shared by both biological and artificial networks.

Interestingly, in parallel to the monkey results from Panichello and Buschman (2021), the RNNs seem to employ two mechanisms for preventing interference between the cued and uncued items in the post-cue delay. Firstly, we found evidence for the modulation of the discriminability of the cued and uncued stimuli. In the representational geometry framework, this corresponds to the overall size of the colour rings. In agreement with

previous reports from both macaque (Panichello and Buschman, 2021) and human research (Sprague et al., 2016; Ester et al., 2018; van Loon et al., 2018; Muhle-Karbe et al., 2021), we observed an increase in colour discriminability for the cued item. Secondly, we found the angles between the cued and uncued planes to be centered around 90 degrees, suggesting that the Cued and Uncued item representations are orthogonalised. The same observation was made in macaque LPFC (Panichello and Buschman, 2021).

In the post-cue delay interval, the representations of cued memory items were arranged into parallel subspaces in both the monkey LPFC (Panichello and Buschman, 2021) and the RNN. This geometry abstracts over the now-irrelevant stimulus location and aligns the colour representations into a common template space. Consequently, the same downstream readout mechanism can be used to generate a response, regardless of the retrocue identity. Our results from Experiment 2 suggested that networks trained with variable delay intervals parallelised the Cued item subspaces to a lesser degree. However, upon further investigation, we discovered that when the networks were tested on a different test dataset (using the maximum delay interval length experienced during training), the subspace alignment index was on the same order of magnitude as reported for networks from Experiment 1. Thus, this apparent discrepancy reflects the gradual nature of the orthogonal-to-parallel transformation. Interestingly, Panichello and Buschman (2021) also report a parallel Cued neural geometry in a related task where the cue is provided prior to the stimulus presentation, at the onset of the trial. Whilst we did not model this task variant, we predict that the neural geometry developed by the RNNs under such conditions would similarly match. This prediction is based on observations of parallel low-dimensional manifolds in other tasks incorporating a response generalisation element, which have been made in both biological and artificial networks (Remington et al., 2018; Luyckx et al., 2019; Bernardi et al., 2020;

Sheahan et al., 2021). Such findings collectively imply that the parallel geometry might too be a general coding principle, its function being to facilitate the generalisation of knowledge across different behavioural contexts.

Nor does it seem like the parallel plane geometry in the post-cue delay is a trivial consequence of the way we set up the task. In Experiment 3 we show that this geometry only emerges when the cued information must be maintained in memory for several cycles prior to the response. In contrast, when there is only a short post-cue delay, the RNNs continue to maintain the cued information in the orthogonal format adopted in the pre-cue delay period. Thus, our results suggest that the parallel geometry might additionally confer a robustness to working memory maintenance, providing a normative explanation for its emergence in the context of the retro-cueing task, as well as a testable prediction for future neurobiological studies. We note that these speculations, based on our simulations, could be confirmed with exact analysis of neural network dynamics (where these become available).

One notable difference between our results and those reported in Panichello and Buschman (2021) pertains to the nature of the Cued item transformation that converts the orthogonal pre-cue geometry into parallel post-cue planes. In macaque LPFC (Panichello and Buschman, 2021), the pre- and post-cue neural subspaces were independent of one another (i.e., the post-cue subspace lied in the null space of the pre-cue subspace, and vice-versa). The authors proposed that the associated transformation might therefore have a role in output gating, whereby only the information expressed in the post-cue neural subspace is directly accessible to the downstream readout circuit (i.e., lies in its the target space). In contrast, following the retrocue input, only one of the cued memory representations was rotated in the RNNs to form the parallel plane geometry. Thus, the pre-cue and post-cue subspaces

were not independent. This difference might stem from the fact that the network output was constrained only at the last timepoint of the trial. Consequently, not transforming the other memory item would not have had a detrimental impact on network performance (such as causing it to respond prematurely).

Whilst the *Cued* and *Cued/Uncued* geometries are interpretable with respect to the task demands as outlined above, this is not the case for the *Uncued* geometry. We found the *Uncued* geometry to be variable between the different network replicants. Surprisingly however, this geometry was also highly predictive of network behaviour – in Experiment 1, it proved to be a reliable predictor of the network training latency. This relationship was negative, indicating that to reach the same level of performance, models that relied on a more parallel *Uncued* geometry needed to complete more training iterations than those that orthogonalised the uncued item subspaces. We interpret this result to be a consequence of using random network initialisation, which would predispose some networks to a more parallel *Uncued* geometry even before training commenced. If the uncued planes happened to be aligned with the readout subspace in a network, it would be more prone to responding based on the uncued item and thus need more training iterations to overcome this response bias. It is unclear whether this geometry bore any relationship to behaviour of the monkeys reported in Panichello and Buschman (2021).

The similar representational geometry with the macaque suggests that the LPFC might use a similar computational principle to solve the retro-cueing task. Thus, in later experiments we capitalised on the inherent accessibility of the RNN model and extended our analyses to two aspects not explored in Panichello and Buschman (2021) – the (i) emergence and (ii) mechanistic basis of the reported representational geometry. As artificial neural network

models are not endowed with representational schemes by hand and instead develop them through training, they offer a normative account of how representational properties can be acquired through learning in biological brains. To this end, we provide a description of the stereotypical representational geometries the RNNs develop at different stages of training. The existence of such learning transitions can be verified in vivo in future studies, to ascertain whether biological learning follows similar principles to the learning algorithms used for training RNNs.

Overall, our findings are consistent with theories proposing that retrocues benefit performance partly by triggering a representational transformation into an action-optimised format (Myers et al., 2017b). We also find evidence for the attentional strengthening hypothesis of the retro-cueing effect (Souza and Oberauer, 2016) – in Experiment 1, we observed a selective increase in colour information for the cued item in the post-cue delay, in accordance with other observations from primate recordings (Panichello and Buschman, 2021) and human neuroimaging (Sprague et al., 2016; Ester et al., 2018; Muhle-Karbe et al., 2021). On the other hand, our findings are inconsistent with accounts proposing the removal of uncued items post-cue (Souza and Oberauer, 2016) – in Experiment 1, we found that the post-cue delay was still characterised by significant information about the uncued item. The same result was reported in monkey LPFC (Panichello and Buschman, 2021). On a broader scale, this work speaks to the utility of the representational geometry approach for revealing computational mechanism in both artificial and biological networks. As seen with regards to the *Cued* geometry, characterising the geometric structure of the stimulus representations and its transformations across different stages of processing permits inferences about the nature of the computations learnt by the network. Over and above that, this framework not

only affords direct cross-modal comparisons within a context of a given paradigm but can also shed light on computational similarities across different task domains.

Chapter 3. Exploring the neural basis of retro-cueing benefit in recurrent neural networks

3.1 Abstract

When multiple items are held in short-term memory, cues that retrospectively prioritise one item over another (retro-cues) can facilitate subsequent recall. Such benefits are observed on trials where the retrocue correctly point to the item later probed for recall (*valid* retrocues), relative to performance on trials where no such cue is presented, or where it is uninformative (*neutral*) or points to the wrong item (*invalid).* The neural basis of this effect is a topic of active research. Under the attentional strengthening theory, retrocueing leads to an increase in the proportion of the limited memory resources allocated to the maintenance of the cued item in working memory, at the cost of the uncued item. An alternative account posits that retrocues additionally trigger a representational transformation that puts the cued item in an action-optimised format. In this chapter, we explored the neural basis of the retro-cueing benefit in recurrent neural networks trained on a task where the retrocues were probabilistic. Networks received two colour-location inputs, followed by a (pre-cue) memory delay, a probabilistic retrocue, a second (post-cue) delay, a deterministic probe, and a final (post-probe) delay. The task was to report the colour of the probed item. Our results suggested that the retrocueing benefit in this setting was explained by the attentional strengthening account. Accordingly, we found that the cue was followed by an increase in the representational discriminability of the cued item, relative to the uncued item. This difference was even more pronounced in the post-probe delay. Crucially, it was significantly attenuated on invalid trials, and the magnitude of this attenuation tracked the memory recall precision difference between valid and invalid trials. In contrast to our expectations, we did

not find evidence for a representational transformation after the probabilistic cue. This prediction could be tested in future neurobiological studies.

3.2 Introduction

In Chapter 2, we explored how the contents of working memory are flexibly manipulated to address changing task demands. Using recurrent neural network (RNN) simulations, we showed that models trained on a retrocueing task formed representations whose neural geometry bore a striking resemblance to the one previously observed in the lateral prefrontal cortex of non-human primates (Panichello and Buschman, 2021). Leveraging the strengths of the computational modelling approach, we then explored how and why the observed geometry might emerge, providing mechanistic and normative hypotheses that can be tested in future neurobiological studies.

In this chapter, we build upon our previous findings and directly explore how retrocues can benefit performance. Numerous studies involving human participants have demonstrated this effect, showing improved accuracy and faster reaction times on trials with retrocues that correctly point to the item later probed for recall (*valid* retrocues), relative to performance on trials where no such cue is presented, or where it is uninformative (*neutral*) or points to the wrong item (*invalid)* (Souza and Oberauer, 2016). Behavioural modelling results suggest that such valid-cue performance benefits stem from the increased probability of recalling the cued item (Murray et al., 2013; Williams et al., 2013; Souza et al., 2014, 2016; Gunseli et al., 2015; van Moorselaar et al., 2015; Wallis et al., 2015; Myers et al., 2018), with some studies also demonstrating a similar increase in recall precision (Williams et al., 2013; Gunseli et al., 2015; Makovski and Pertzov, 2015; van Moorselaar et al., 2015; Wallis et al., 2015; Myers et al., 2018).

However, despite a large body of research, the neural processes underpinning this *retrocueing effect* remain not well understood. Most proposals draw parallels with the mechanisms of top-down attention directed to external stimuli, extending the notion of biased sensory processing of the cued stimulus to mnemonic processes (Chun, 2011; Chun et al., 2011; Kiyonaga and Egner, 2013; Ma et al., 2014; Bays, 2015). Accordingly, selective attention is thought to increase the proportion of the limited memory resources allocated to the maintenance of the cued item in working memory, at the cost of the uncued items (Chun, 2011; Chun et al., 2011; Kiyonaga and Egner, 2013; Ma et al., 2014; Bays, 2015). In line with such *attentional strengthening* accounts, many neurophysiological markers of exogenous attention are also observed following the presentation of a retrocue. This includes (but is not limited to) the engagement of the same frontoparietal circuits (Griffin and Nobre, 2003b; Nobre et al., 2004; Nee and Jonides, 2009; Tamber-Rosenau et al., 2011) as well as increased activity in sensory areas responsible for processing of the cued item (Lepsien and Nobre, 2007; Higo et al., 2011; Lepsien et al., 2011; Lewis-Peacock and Postle, 2012; LaRocque et al., 2013; Kuo et al., 2014; Lewis-Peacock et al., 2015). However, a recent theoretical account theory (which we will refer to as the *transformation hypothesis*) proposes that, in addition to such attentional strengthening mechanisms, retrocues might also trigger a transformation of the cued memory item into a prospective, action-optimised format (Myers et al., 2017b). This allows for the preparation of the right behavioural policy in advance, leading to faster and more accurate responses. For example, consider the task of reporting whether the colour of the probe matches that of the cued item. A retrocue that points to the green item from the stimulus set allows the correct response to be prepared in advance, by allowing the abstract task rule (match/mismatch) to be reconfigured into something much more specific (press match if the probe is green, and mismatch otherwise). A recent EEG study provided evidence for the transformation of the cued item into a

different representational format (Muhle-Karbe et al., 2021). Firstly, using behavioural modelling, the authors showed that the decoding strength of the cued item tracked the rate of evidence accumulation, suggesting that the cued item was transformed into a decision variable. Secondly, trials where the representation of the uncued item was more similar to that of when the same item was cued were associated with increased behavioural interference. These finding are consistent with the idea that cued memory items are represented in a format that can more readily drive behavioural responses. Finally, as previously noted, the *transformation hypothesis* is also supported by our findings from Chapter 2, and the accompanying non-human data from Panichello and Buschman (2021).

In this Chapter, we trained randomly initialised RNNs with gradient descent to perform a variant of the retrocueing task where the relationship between the retrocue and probe was probabilistic. We predicted that networks trained under conditions where the retrocue was informative about the probed item would show behavioural markers similar to those previously reported in humans. This prediction was based on the fact that under such conditions, it is optimal to combine evidence from all available sources of information (O'Reilly et al., 2012). Having confirmed the existence of such valid-cue performance improvements, we investigated the mechanisms that gave rise to it in the RNN. Based on the *transformation hypothesis* outlined in (Myers et al., 2017b), we expected to see the same orthogonal-to-parallel transformation of the cued item geometry as was observed in Chapter 2. However, contrary to this prediction, we found no evidence for such reformatting after the presentation of a probabilistic retrocue. Instead, our results suggested that the principal mechanism for the retrocueing effect in this setting was through the gain modulation of the cued item representation. This finding is in agreement with the *attentional strengthening* accounts of the retrocueing effect and could be tested in future neurobiological studies.

3.3 Methods

The simulations and analyses presented in this chapter were inspired by the findings described in Chapter 2. Therefore, to avoid repetition, in cases where the methodology used in the current chapter is identical, we point the reader to the relevant sections of Chapter 2.

3.3.1 Task structure

We modified the trial structure described in the previous chapter (section 2.3.2.1) by setting the length of the pre- and post-cue delay periods to 5 cycles, as well as including a location probe (1 cycle), and a third, final memory delay of 5 cycles (post-probe delay, see **Fig. 3.1A** for two example trials). Thus, networks were presented with two colour inputs, followed by a pre-cue delay, a retro-cue input, a post-cue delay, a probe input, and a final, post-probe delay. Networks were trained to output the colour shown at the probed location and could disregard the retro-cue input. We designed 3 experimental conditions, defined by the proportion of trials on which the retro-cue and probe inputs matched: 100%, 75% and 50%. The training dataset included 512 unique trials and for each epoch, we drew a random subset on which the retro-cue and probe inputs were set to mismatch.

3.3.2 Model architecture and training

Model architecture was identical to that described in the previous chapter (section 2.3.1). The models (N = 30 per condition) were not subject to any noise, and we stopped training once the loss (summed over all the validly cued trials in the last epoch) crossed the threshold of 0.0005. A new set of models was trained for each condition (defined by the retrocue validity). To increase the statistical power of the multiple linear regression analysis, we increased the sample size to N=100 networks for each condition.

3.3.3 Hidden activity analysis

Analysis methods were identical to those described in Chapter 2. Thus, to obtain planes of best fit for a pair of conditions (such as stimulus location) and calculate the angles θ (between-plane angle) and ψ (phase alignment) between them, we followed the same steps as described in section 2.3.2.3. To calculate the alignment index (AI) between two subspaces, we applied the methodology described in section 2.3.2.4. Lastly, to calculate the colour discriminability index (CDI) for a given colour ring, we followed the same steps as outlined in section 2.3.2.6. For any analysis pertaining to the representations formed by the RNNs after the presentation of the probe input (post-probe representations), we used the hidden activity data from the last timepoint of the post-probe delay.

3.3.4 Multiple linear regression analysis

The Probed, Unprobed and Probed/Unprobed AI values for trained networks were calculated by following the same steps as outlined in Chapter 2 (section 2.3.2.8), except using the hidden activity data from the last timepoint of the post-probe delay. The values were calculated for valid trials only. The dependent variable (total number of training epochs) was box-cox transformed (separately for each validity condition) prior to being used in the multiple regression model.

3.3.5 Behavioural analysis

We sampled network choices from the last-timepoint of the trial as previously described in section 2.3.1 and then fit a 3-component mixture model to the trial-wise choice data,

separately for every model, using the Analogue Report Toolbox (https://www.paulbays.com/toolbox/) in MATLAB version 2020b (*Mathworks*). The model is fit using maximum likelihood estimation and is described by:

$$\begin{aligned} \mathrm{p}(\alpha_{reported}) = (1 - \mathrm{pU} - \mathrm{pNT})\phi_K(\alpha_{reported} - \alpha_{cued}) \\ +\mathrm{pU}\frac{1}{2\pi} + \mathrm{pNT}\phi_K(\alpha_{reported} - \alpha_{uncued}) \end{aligned} \quad (7.1)$$

where pU stands for the probability of making random guesses, pNT describes the probability of misremembering the target location and ϕ_K is a von Mises distribution function with zero mean and concentration parameter K. Probability of making the target response (pT) is the difference between 1 and the sum of pU and pNT. Thus, choices are modelled as arising from a mixture of target and non-target-guided components, as well as random guesses. The fit parameters (pT, pU, pNT and K) were compared between trials on which the retro-cue and probe inputs matched (valid) and mismatched (invalid trials), as well as across the experimental conditions (100%, 75% and 50% retro-cue validity). To this effect, for each parameter, we used a 2x3 mixed model ANOVA, with a repeated measures factor of trial type, and a between-subject factor of experimental condition. In cases where the data did not satisfy the assumptions of the mixed model ANOVA, we used a combination of non-parametric tests instead. In such cases, to test the main effect of the repeated measures factor of trial type, we concatenated the data from the two validity conditions and assessed the difference between valid and invalid trials with a Wilcoxon signed-rank test. To test the main effect of the between-subject factor of validity condition, we averaged the data between the valid and invalid trials and compared it between the two validity conditions with a Mann-Whitney U-test. Lastly, to test for an interaction between the trial type and

validity condition factors, we calculated the difference between the valid and invalid trials and compared it across the validity conditions with a Mann-Whitney U-test.

3.3.6 Software

All statistical analyses described were implemented in python using the scipy.stats package (version 1.7.1), with the exception of the ANOVA and multiple linear regression analyses, which were performed in JASP (version 0.12.2). All networks were trained using the pytorch library (version 1.10.2).

3.3.7 Statistics abbreviations

M: mean

SEM: standard error of the mean

CI: confidence interval

3.4 Results

3.4.1 Networks trained with informative retrocues show the retrocueing effect

Seeking a normative account of the retrocueing effect observed in humans, we trained RNN models on a modified version of the task from Chapter 2 (see **Fig 3.1A** for a schematic of an example trial). As before, the networks were first presented with two colour stimuli, followed by a delay, a retro-cue, and another delay. However, unlike in the experiments from Chapter 2, the cue indicated which item was trial-relevant with varying degree of validity and was followed by a deterministic probe and another memory delay. We manipulated the percentage of training trials on which the retro-cue and probe inputs matched (*retro-cue validity*). We designed 2 experimental conditions of interest, where the

retro-cue validity was set to 75 and 50%, in addition to a control condition with validity equal to 100% (this latter experiment resembles those in Chapter 2, except for differences in timing of the trial events).

When the retro-cue validity experienced during training was 75%, model performance (as measured by MSE recall) was higher for validly than invalidly cued trials (see **Fig 3.1B**). In contrast, when the information provided by the retro-cue was maximally ambiguous (validity = 50%), performance on valid and invalid trials was matched (note that in this case, even though the retro-cue signal does not correlate with the trial-relevant item, it is still possible to make a distinction between valid and invalid trials, as the identities of both the retro-cue and probe were signalled using the same pair of input nodes, valid trials refer to those on which the same node was activated twice, as shown in **Fig 3.1A**, *left*). To understand these effects, we leveraged the fact that errors on delayed-estimation tasks such as the retro-cueing task are thought to arise from a combination of imprecise reports of the target item, incorrect reports of the non-target item, and random guesses. Adopting an approach familiar from human behavioural studies, we fit a mixture model incorporating these error terms to the individual model choices (Bays et al., 2009). The model had four free parameters, corresponding to memory precision (*K*), probability of recalling the target item (*pT*), probability of recalling the non-target item (*pNT*), and probability of making random guesses (pU). We compared each fitted parameter type across conditions (validity = 75% and 50%) and trial types (*valid, invalid*) with a 2x2 mixed model ANOVA (or a non-parametric equivalent in the case of the probability parameters, which had highly unequal variances between conditions).

The memory precision parameter (K) showed a significant main effect of trial type ($F(1,58) = 144.91, p < .001$; valid = 26.9, invalid = 22.78) and a significant interaction with condition ($F(1,58) = 144.76, p < .001$), but no main effect of condition ($F(1,58) = 0.86, p = .358$). Post-hoc tests with Holm-correction revealed that recall precision was significantly higher on valid than invalid trials in the 75% validity condition ($t(29) = 17.02, p < .001$), whilst there was no difference in the 50% validity condition ($t(29) = .004, p = .997$). Furthermore, the precision parameter on invalid trials was significantly lower in the former compared to latter condition ($t(29) = -4.22, p < .001$), with an opposite trend on valid trials, which did not however reach statistical significance ($t(29) = 2.40, p = .057$).

An analogous analysis concerning the probability of recalling the target item (pT) revealed that it was significantly higher on valid than invalid trials (Wilcoxon signed-rank test, $W(59) = 1337$, p = .002; valid = .996, invalid = .994) as well as in the 75%, compared to 50% validity condition (Mann-Whitney U-test: $W(29) = 286, p = .015$; $pT = .997$ and $pT = .992$, respectively). There was also a significant interaction between the two factors (mean difference between valid and invalid trials – 75% validity: .003, 50% validity: 6.02×10^{-6}, Mann-Whitney U-test: $W(29) = 208, p < .001$). Holm-corrected post-hoc tests revealed that pT was significantly higher on valid trials in the probabilistic than in the neutral condition (Mann-Whitney U-test: W(58) = 199, p < .001), with no significant difference between the conditions on invalid trials (Mann-Whitney U-test: W(58) = 336, p = .187). In the 50% validity condition, there was no significant difference between the two trial types (Wilcoxon signed-rank test: W(29) = 224, p = .871). In contrast, in the 75% validity condition, pT was significantly higher on valid than invalid trials (Wilcoxon signed-rank test: W(29) = 430, p < .001).

Probability of random guesses (*pU*) was significantly lower on valid than invalid trials (Wilcoxon signed-rank test, *W(59)* = 571, p = .011) and in the 75%, compared to the 50% validity condition (Mann-Whitney U-test: *W*(29) = 618, *p* = .013). The interaction between the trial type and condition was also significant, with a more pronounced difference between valid and invalid trials in the 75% validity condition (Mann-Whitney U-test: *W*(29) = 637, *p* = .005; *pU* = -.002 and *pU* = -7.810 x 10^{-5} for 75% and 50% validity, respectively). Holm-corrected post-hoc tests revealed a significant between-condition difference on valid (Mann-Whitney U-test: *W*(58) = 683, *p* = .002, *pU* = 0.007 and 0.001 for 50 and 75% validity, respectively) but not invalid trials (Mann-Whitney U-test: *W*(58) = 588, *p* = .084). In the 75% validity condition, *pU* was significantly lower on valid than invalid trials (Wilcoxon signed-rank test, *W(29)* = 75, *p* = .004; *pU* = 0.001 and 0.003, respectively), whilst there was no such difference in the 50% validity condition (Wilcoxon signed-rank test, *W(29)* = 234, *p* = .984).

Lastly, the analogous analysis regarding the probability of recalling the non-target item (*pNT*) yielded a significant main effect of trial type (Wilcoxon signed-rank test, *W(59)* = 577, *p* = .013, valid = 4.68 x 10^{-4}, invalid = 8.89 x 10^{-4}) but no main effect of the validity condition (Mann-Whitney U-test: *W*(29) = 447, *p* = .971). The interaction between the two factors was significant (mean difference in *pNT* between valid and invalid trials – 75% validity: -9.14 x 10^{-4}, 50% validity: 7.0 x 10^{-5}, Mann-Whitney U-test: *W*(29) = 628, *p* = .008). Holm-corrected post-hoc tests revealed a significant between-condition difference on valid (Mann-Whitney U-test: *W*(58) = 659, *p* = .006, *pNT* = 6.65 x 10^{-4} and 0.001 for 50 and 75% validity, respectively), but not invalid trials (Mann-Whitney U-test: *W*(58) = 410, *p* = .562). In the 75% validity condition, *pNT* was significantly lower on valid than invalid trials (Wilcoxon signed-rank test, *W(29)* = 75, *p* = .002; *pNT* = 1.99 x 10^{-4} and .001, respectively),

whilst there was no such difference in the 50% validity condition (Wilcoxon signed-rank test, $W(29) = 225$, $p = .887$).

In summary, we found significant interaction effects between the validity condition and trial type for all mixture model parameters examined (shown in **Fig 3.1C**). In the 50% validity condition, there were no significant trial-type differences in the fitted parameters. In contrast, in the 75% validity condition, both the precision (K) and probability of recall (pT) of target items were significantly higher on valid than invalid trials, whilst the probability of recalling the non-target item (pNT) and making random guesses (pU) followed the opposite pattern. It is however worth noting that all the estimated probability quantities were close to ceiling (or floor, in the case of pNT and pU). These results imply that the superior performance of the neural networks on validly cued trials under probabilistic conditions can be accounted for by precision, recall, protection against intrusions of the uncued item, and nonspecific lapses.

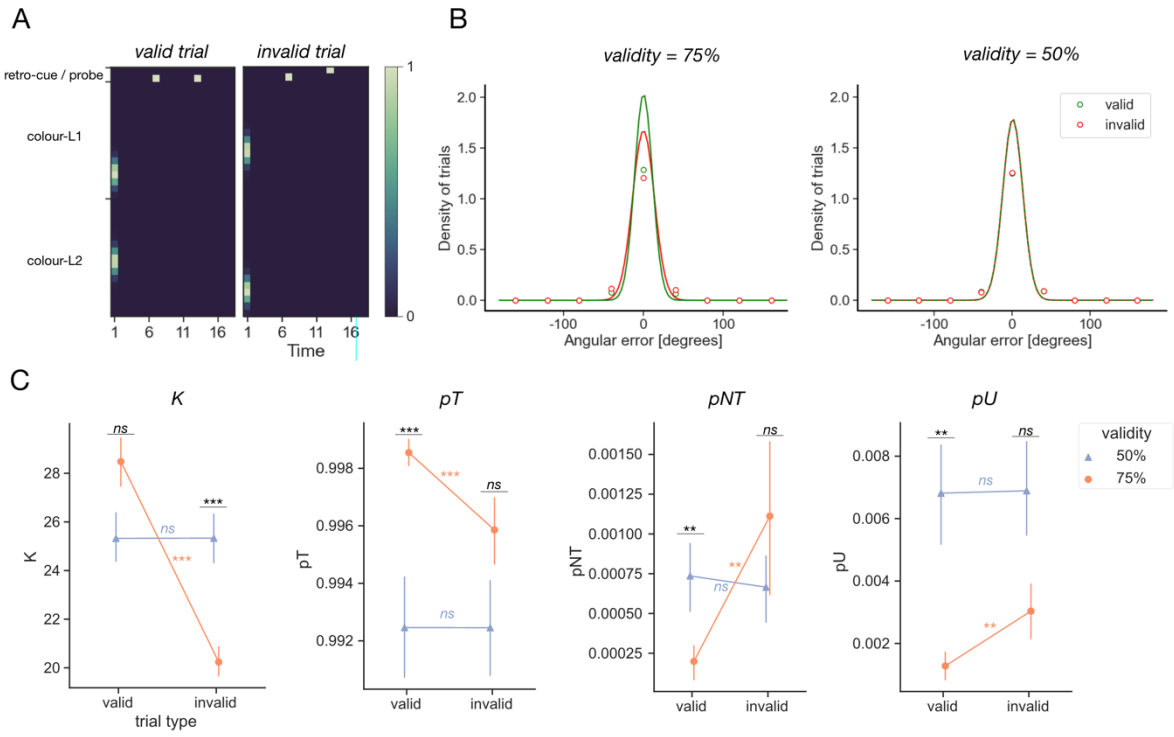


Fig 3.1. Behavioural findings from the probabilistic cue paradigm. A. Task structure. Example valid trial (where the retro-cue and probe inputs match) shown on the left, example invalid trial shown on the right. **B.** Distribution of errors made by the models after convergence (dots and spikes, *M*±*SEM*) shown with a best-fit von Mises pdf (line). Performance on trials where retro-cue matched the probe (*valid*) shown in green, mismatch (*invalid*) trials shown in red. Data plotted for models trained under 2 different cue validity conditions – 75 and 50%. Note that errors were binned into 40 degree wide bins for plotting. **C.** Comparison of the mixture model parameters (left to right panels: *K, pT, pNT* and *pU*) across cue validity conditions (50 and 75%, corresponding to lavender triangles and orange circles, respectively) and trial types (valid, invalid; shown on the x-axis). Asterisks denote the significance levels of the post-hoc tests described in the main text: ***: $p < .001$, **: $p < .01$, *ns*: $p >= .05$. Comparisons within a validity level shown in orange and lavender, comparisons across different validity levels (but for the same trial type) denoted in black.

3.4.2 Parallel Cued item geometry is only observed following deterministic retrocues

Having explored the nature of the behavioural effect observed in **Fig 3.1B**, we then compared the neural geometries formed in networks trained under different retro-cue validity conditions. We were interested in the relationship between the trial-relevant item subspaces in the different delay periods. We found that networks trained under the deterministic regime (where the retro-cue was valid on 100% of trials) kept the information in orthogonal subspaces during the pre-cue delay (v-test: $v(29) = 25.79$, $p < .001$) and underwent a configurational change into parallel planes after the retro-cue presentation (v-test: $v(29) = 25.09$, $p < .001$*)* that prevailed into the post-probe delay interval (v-test: $v(29) = 28.16$, $p < .001$; see also **Fig 3.2A**). The post-probe planes were also significantly phase-aligned (v-test: $v(29) = 29.96$, $p < .001$). In contrast, in the networks trained under non-deterministic conditions (cue validity 75% and 50%), the retro-cue did not trigger this

representational change (v-test: $v(29) = 3.45$, $p = .186$ and $v(29) = 5.34$, $p = .084$; respectively). Under these conditions, the information was transformed into the parallel format later in the trial, in the post-probe delay ($v(29) = 25.49$, $p < .001$ and $v(29) = 25.49$, $p < .001$; respectively). At that timepoint, the cued planes were also phase-aligned ($v(29) = 29.96$, $p < .001$ and $v(28) = 28.92$, $p < .001$; for cue validity 75% and 50%, respectively).

3.4.3 Neural geometries formed in the post-probe delay period are behaviourally relevant

Next, we asked whether the representational geometries developed by the trained models were behaviourally relevant. We focused on the trial period immediately preceding the response (i.e., the post-probe delay) and quantified the relationship between the *colour-L1* and *colour-L2* items when they were probed (*Probed Geometry*), unprobed (*Unprobed Geometry*), or when only one of them was probed and the other one was not (*Probed/Unprobed Geometry*). To do so, we regressed the AI metrics for the *Probed*, *Unprobed* and *Probed/Unprobed* geometries against the total number of training epochs completed by the networks, separately for each validity condition. In the 100% validity condition, the regression model explained a significant proportion of the variance in training speed across networks ($R^2 = .23$, $F(3,96) = 9.49$, $p < .001$). The AI between the *Unprobed* subspaces was positively predictive of the number of training epochs ($ß = 4.59 \times 10^{-4}$, $t(96) = 4.71$, $p < .001$). Additionally, the AI between the *Probed* subspaces bore a significant negative relationship with the training speed ($ß = -8.93 \times 10^{-4}$, $t(96) = -3.05$, $p = .003$), whilst the AI between the *Probed/Unprobed* subspaces did not contribute significantly to the model ($ß = 5.51 \times 10^{-5}$, $t(96) = 0.52$, $p = .602$). The same pattern of results was observed in the 75% ($R^2 = .42$, $F(3,96) = 23.03$, $p < .001$; AI *Unprobed*: $ß = 0.006$, $t(96) = 4.88$, $p < .001$; AI *Probed*: $ß = -0.004$, $t(96) = -3.65$, $p < .001$; AI *Probed / Unprobed*: $ß = 0.004$, $t(96) = 2.07$,

p = .041) and 50% (R^2 = .38, F(3,96) = 19.46, p < .001; AI *Unprobed*: ß = 0.046*, t(96)* = 5.52, p < .001; AI *Probed*: ß = -0.024*, t*(96*)* = -2.93, p = .004; AI *Probed / Unprobed*: ß = 7.88 x 10^{-4}*, t*(96*)* = 0.07, p = .947) validity conditions (see also **Fig 3.3B**).

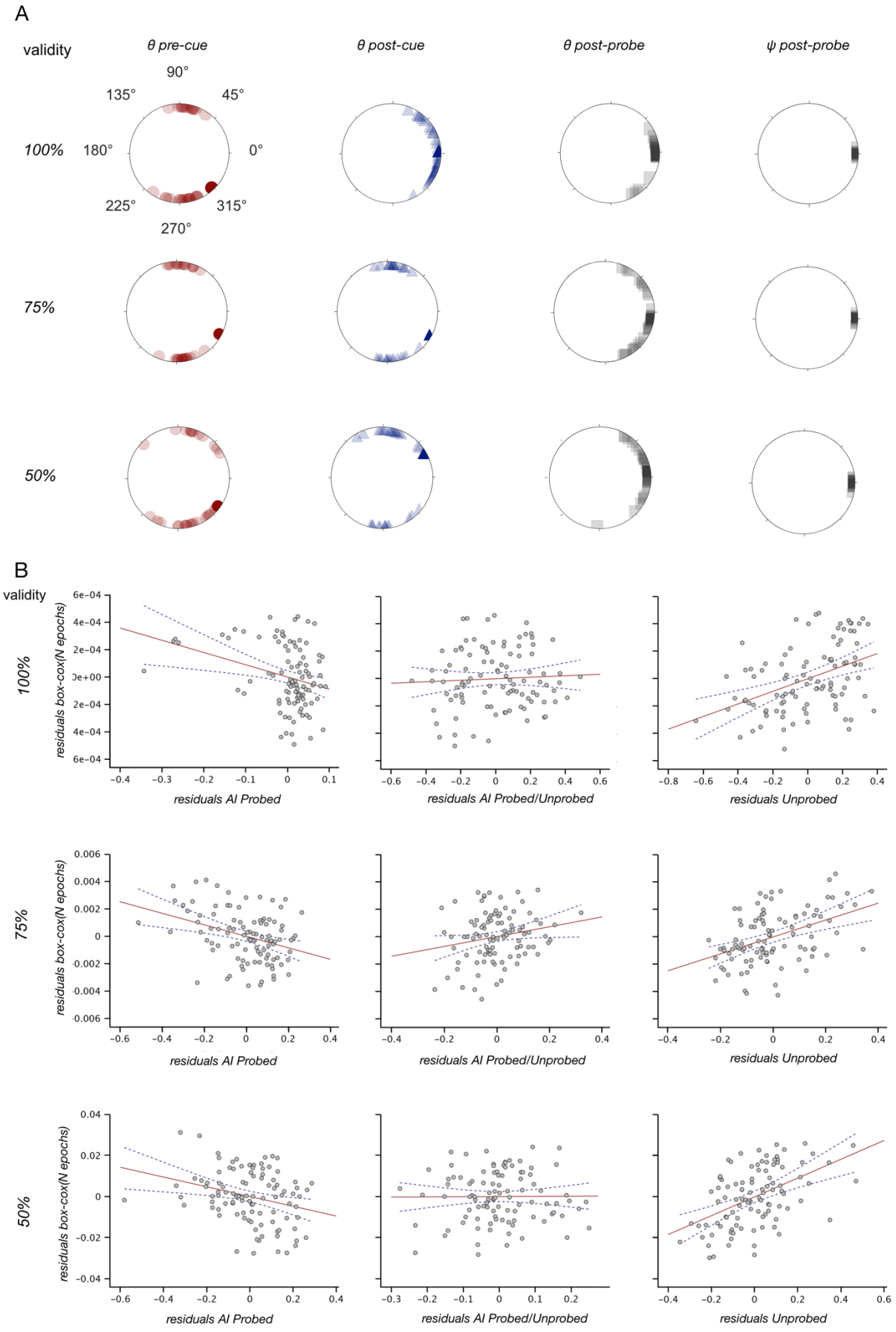

A
validity
θ pre-cue
θ post-cue
θ post-probe
ψ post-probe
90°
135°
45°
180°
0°
225°
315°
270°
100%
75%
50%
B
validity
100%
75%
50%
residuals box-cox(N epochs)
residuals AI Probed
residuals AI Probed/Unprobed
residuals Unprobed

Fig 3.3. Neural geometry findings from the probabilistic cue paradigm. A. Distribution of plane angles θ and phase-alignment angles ψ formed in the delay intervals for all trained networks. Values for the pre-cue, post-cue and post-probe delays shown as red circles, blue triangles and grey squares. Data from individual models shown in opaque, and grand averages in solid colours. Plots for models trained under trained under the deterministic (retrocue validity = 100%) and non-deterministic retrocue validity conditions (75% and 50%) ordered top to bottom rows, respectively. **B.** Results of the regression analyses predicting the number of training epochs (after a box-cox transformation) from AI *Probed*, AI *Probed/Unprobed* and AI *Unprobed* (columns, left to right*)*. Rows correspond to the three retrocue validity conditions examined (100, 75 and 50%, top to bottom). Data shown in grey dots, regression lines of best fit in red, 95% confidence interval for the slope in navy dashed lines.

3.4.4 Retrocueing effect is linked to a gain-modulation mechanism of cued item representations

Finally, we asked how networks trained under different cue validity regimes modulated the discriminability of their colour representations at different timepoints of the trial. To this end, we performed a CDI analysis analogous to the one described in Chapter 2. We calculated the CDI for the pre-cue items, as well as the cued, uncued, probed and unprobed colours (using data from the end of the appropriate delay interval), separately for valid and invalid trials. For the results shown in **Fig 3.3A**, the CDI values were (1) averaged across trials and locations to form a single pre-cue average and (2) averaged across valid and invalid trials to construct the cued and uncued averages. To assess for statistical differences between conditions, CDI data from the post-cue and post-probe timepoints was input into a repeated measures ANOVA with 3 factors: timepoint (two levels: post-cue, post-probe), item (two

levels: indicated and not indicated by the immediately preceding cue/probe input) and trial type (two levels: valid and invalid). For networks trained under 75% cue validity conditions, this analysis showed significant main effects of all factors: timepoint ($F(1,29) = 124.91$, $p < .001$, $\eta^2 = 0.25$), item ($F(1,29) = 122.51$, $p < .001$, $\eta^2 = 0.24$) and trial type ($F(1,29) = 89.62$, $p < .001$, $\eta^2 = 0.01$). The two-way interaction between factors timepoint and item was statistically significant ($F(1,29) = 132.14$, $p < .001$, $\eta^2 = 0.30$), as were the interactions between timepoint and trial type factors ($F(1,29) = 83.59$, $p < .001$, $\eta^2 = 0.01$) and item and trial type factors ($F(1,29) = 83.59$, $p < .001$, $\eta^2 = 0.01$). The three-way interaction was also statistically significant ($F(1,29) = 89.62$, $p < .001$, $\eta^2 = 0.01$). To follow up, we performed two repeated measures ANOVAs with factors item and trial type, one for each timepoint.

For the post-cue timepoint, there was a significant main effect of item ($F(1,29) = 14.94$, $p < .001$, $\eta^2 = 0.34$), with the CDI significantly higher for cued ($M = 18.6$, $SEM = 2.71$) than uncued items ($M = 5.92$, $SEM = 2.71$). There was no significant main effect of trial type ($F(1,29) = 0$, $p = 1$, $\eta^2 = 0.0$) or an interaction between item and trial type ($F(1,29) = 0$, $p = 1$, $\eta^2 = 0.0$). This is consistent with the fact that at the post-cue timepoint, the network does not yet know whether the presented retrocue is valid (i.e., matches the to-be-shown probe) or invalid. Note that the absolute lack of differences is a consequence of the network performance being fully deterministic (i.e., network activity is not subject to any hidden noise).

For the post-probe timepoint, there were significant main effects of item ($F(1,29) = 123.77$, $p < .001$, $\eta^2 = 0.77$) and trial type ($F(1,29) = 89.62$, $p < .001$, $\eta^2 = 0.02$) as well as a significant interaction between them ($F(1,29) = 83.59$, $p < .001$, $\eta^2 = 0.02$). This analysis was followed up with Bonferroni-corrected post-hoc tests. The probed colour CDI was significantly higher

on valid than invalid trials ($t(29) = 13.07$, $p < .001$), whereas the analogous difference for the unprobed colour was negative, but it did not reach statistical significance ($t(29) = -1.1$, $p \sim 1$). On valid trials, the probed item CDI was significantly higher than the unprobed item CDI *(*$t(29) = 12.74$, $p < .001$). An analogous effect was observed on invalid trials ($t(29) = 9.06$, $p < .001$).

To compare these results to the ones reported in Chapter 2, we performed 3 contrasts between the (1) pre-cue baseline (averaged across the cued and uncued items, as well as valid and invalid trials) and cued item CDI (averaged across valid and invalid trials), (2) pre-cue baseline and uncued item CDI (averaged across valid and invalid trials), and (3) cued and uncued item CDI. The cued item CDI was significantly higher than the pre-cue baseline (Wilcoxon signed rank test $W(29) = 462$, $p < .001$ one-tailed, matched rank biserial correlation = 0.99) and the uncued item CDI (Wilcoxon signed rank test $W(29) = 462$, $p < .001$ one-tailed, matched rank biserial correlation = 0.99), in agreement with the results reported in Chapter 2. A Bayesian paired-sample t-test between the uncued item CDI and the pre-cue baseline returned a BF_{10} of 29.69, indicating strong evidence for the alternative hypothesis of a difference in medians. For comparison, a two-tailed Wilcoxon signed-rank test was significant ($W(29) = 404$, $p < .001$, matched biserial rank correlation = 0.74). These results suggest that, unlike what we observed in the original paradigm from Chapter 2, the CDI of the uncued item increased relative to the pre-cue baseline.

A similar 3-way repeated measures ANOVA was performed for the CDI data from models trained under 50% cue validity conditions. This analysis yielded significant main effects of factors timepoint ($F(1,29) = 125.14$, $p < .001$, $\eta^2 = 0.31$) and item ($F(1,29) = 110.87$, $p < .001$, $\eta^2 = 0.24$), and a significant interaction between them ($F(1,29) = 110.78$, $p < .001$, η^2

= 0.24), with all other factors and interactions not reaching the threshold of statistical significance. Bonferroni-corrected post-hoc tests showed a significant difference in CDI between the probed and unprobed items ($t(29) = 14.89$, $p < .001$), but no difference between the cued and uncued items ($t(29) = .002$, $p \sim 1$). We also performed an additional contrast between the pre-cue baseline and the post-cue CDI (both averaged across items and trial types) and found no significant difference between them (paired-samples t-test $t(29) = 1,17$, $p = .251$, Cohen's $d = .214$). These results are consistent with the network not showing any sensitivity to the unreliable (fully stochastic) cue.

Lastly, we also analysed the CDI data from the control condition in which models were trained under 100% cue validity conditions. Data was entered into a 2-way repeated measures ANOVA with factors delay (3 levels: pre-cue, post-cue and post-probe) and item (two levels: indicated and not indicated by the immediately preceding cue/probe input; note this distinction is arbitrary in the pre-cue delay). The ANOVA yielded a significant main effect of both factors: timepoint ($F(2,58) = 151.0$, $p < .001$, $\eta^2 = 0.34$) and item ($F(1,29) = 149.57$, $p < .001$, $\eta^2 = 0.17$), as well as a significant interaction between them ($F(2,58) = 148.76$, $p < .001$, $\eta^2 = 0.33$). A linear contrast on the timepoint factor returned a significant positive coefficient estimate ($ß = 2298.44$, $t(58) = 15.15$, $p < .001$), indicating that the CDI increased across the three consecutive delay periods. A paired t-test contrast comparing the difference between the probed and unprobed CDI to the difference between the cued and uncued item CDI returned a significant result ($t(29) = 12.19$, $p < .001$, Cohen's $d = 2.23$), indicating that the difference between items is larger at the post-probe timepoint ($M = 6393.32$) compared to the post-cue timepoint ($M = 78.51$). To sum up, these results suggest that the discriminability of the colour items increases across the consecutive delay intervals,

and it does so more for the behaviourally relevant (i.e. cued/probed) items, compared to the behaviourally irrelevant (uncued/unprobed) ones.

The CDI results described above suggest that the networks trained under 75% cue validity regime modulate the discriminability of colour representations differentially for the cued and uncued items, and this has further consequences for the discriminability of probed and unprobed items, depending on whether the presented cue was valid or not. We were therefore interested in whether this result could offer a potential mechanistic explanation for the differences in network behaviour observed on valid and invalid trials (shown in **Fig 3.3B**). To test this hypothesis, we first calculated the probed item benefit for each trial type, defined as the difference between probed and unprobed item CDI. We then calculated the difference in this benefit between valid and invalid trials and normalised it with respect to the value on valid trials. Consequently, we correlated this metric (referred to as the normalised CDI benefit) against the difference in the corresponding mixture model parameters estimated for valid and invalid trials. Interestingly, we found that it was strongly and positively correlated with the difference in magnitude of the memory precision parameter K between valid and invalid trials (Spearman's $r(28) = .65$, $p < .001$). We also observed a slightly weaker correlation between the normalised CDI benefit and the analogous difference in probability of making random guesses parameter pU (Spearman's $r(28) = .41$, $p < .05$), and no statistically significant correlations with the other mixture model parameters (pT: Spearman's $r(28) = -.29$, $p = .121$, pNT: Spearman's $r(28) = -.15$, $p = .424$; see also **Fig 3.3B**). Thus, the valid-cue behavioural benefit observed in the RNN seems to be underpinned by a gain-modulation mechanism, whereby the increased discriminability of the cued item is followed by an even better discriminability of the probed item (relative to the unprobed item), which in turn is associated with increased recall precision.

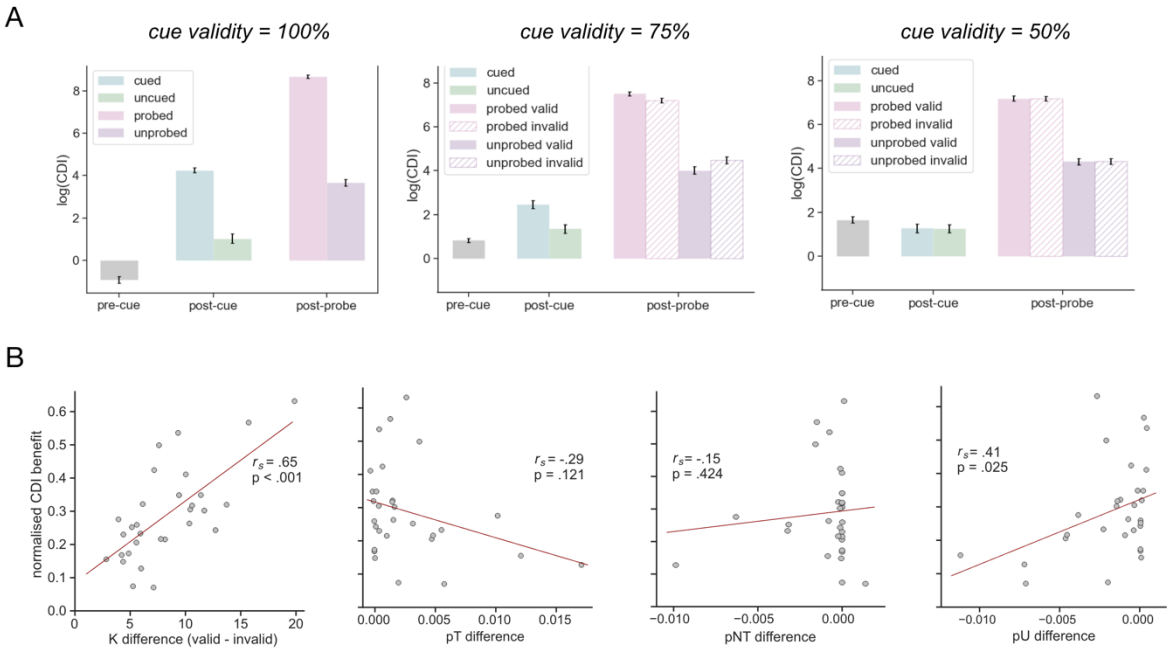


Fig 3.3 CDI findings from the probabilistic cue paradigm. A. Colour discriminability index (CDI) for the pre-cue, post-cue and post- colour planes. Pre-cue values were averaged across both locations, whilst the post-cue values for the cued and uncued subspaces were averaged across valid and invalid trials. Bars correspond to the mean across trained network, error bars denote SEM. Panels correspond to results from models trained under 100%, 75% and 50% retrocue validity conditions, from left to right. **B.** Scatterplots showing the relationship between the normalised CDI benefit (y-axis, see main text) and the difference in mixture model parameter estimates on valid and invalid trials (x-axes). Panels correspond to the memory precision parameter K, probability of choosing the target item pT, probability of choosing the non-target item pNT, and probability of making random guesses pU, from left to right. Results of a Spearman correlation between the variables shown on the x- and y-axis shown on each panel. Lines of best fit plotted in red.

3.5 Discussion

In this Chapter, we explored the mechanisms of prioritising the contents of working memory in the context of a probabilistic retrocueing task. This allowed us to measure the behavioural

consequences of retrocues and link them to their underlying mechanism. RNNs showed similar behavioural markers to those previously reported for human participants. Subsequently, we tested the predictions of existing theoretical accounts of attention to working memory and found evidence consistent with *the attentional strengthening* hypothesis of the retrocueing benefits.

RNNs trained with probabilistic retrocues captured the key behavioural effects characteristic of human participants performing similar tasks. We found that, under conditions where the cue was on average valid (75% validity paradigm), recall performance was improved on trials where the retrocue correctly pointed to the to-be-probed item and degraded otherwise, relative to the corresponding performance levels in a fully probabilistic set-up (50% validity paradigm). These effects have also been reported in human participants performing equivalent tasks (reviewed in Souza and Oberauer, 2016). Our results further showed the locus of the retrocue benefit (measured with respect to an uninformative cue baseline) to lie within an increased probability of recalling the target item as well as reduced probability of making swap or random errors. Analogous observations have been made in the human literature (Murray et al., 2013; Williams et al., 2013; Souza et al., 2014, 2016; Gunseli et al., 2015; Makovski and Pertzov, 2015; van Moorselaar et al., 2015; Wallis et al., 2015). However, we did not find a reliable increase in the target item recall precision on valid trials, which has been described in some human studies (Williams et al., 2013; Gunseli et al., 2015; Makovski and Pertzov, 2015; van Moorselaar et al., 2015; Wallis et al., 2015). This discrepancy most likely reflects the training procedures followed, which lead to matching overall performance levels on valid trials in both the 75% and 50% validity regimes.

Having confirmed that the presentation of informative retrocues indeed led to increased performance in our networks, we set out to test the predictions of the *transformation hypothesis* of the retrocueing effect. To this effect, we analysed the neural geometry of cued item representations in the post-cue delay. To our surprise, only networks trained with 100% valid retrocues were characterised by a parallel *Cued* geometry at that timepoint. In contrast, networks trained under 75 and 50% validity conditions continued to maintain the cued items in orthogonal subspaces until the post-probe delay, when the planes were rotated into a shared subspace. This finding was particularly surprising in the case of the 75% validity condition, given that the presence of the behavioural effect described above strongly suggested that the networks relied on the retrocue input to solve the task. However, upon further investigation, we discovered that this effect related to a gain-modulation mechanism, as discussed below.

We discovered differences in the amount of colour information (quantified by the CDI) between cued and uncued items for networks trained under the 100 and 75% validity conditions. More specifically, in the post-cue delay, the CDI for the uncued item was significantly lower than for the cued item. This effect carried over into the post-probe delay, where the CDI was significantly higher for probed than unprobed items, with the magnitude of this difference significantly attenuated on invalidly cued trials. Most interestingly, we observed the size of this validity effect on CDI to correlate with network performance in the 75% condition (in particular, with the precision on the output estimated from the mixture model), providing an explanation for why the network showed the retro-cueing benefit. Overall, this result is in agreement with the *attentional strengthening* theories of the retrocueing effect which assume a finite cognitive resource for WM maintenance (Chun, 2011; Chun et al., 2011; Kiyonaga and Egner, 2013; Ma et al., 2014; Bays, 2015). Under

such accounts, attentional selection of the cued memory item necessitates the withdrawal of resources from, and thus leads to a degraded memory for the uncued item.

In parallel to the analyses reported in Chapter 2, we sought to understand whether the neural geometries formed by the networks bore any relevance for their training. We focused on the geometries formed immediately prior to the response, during the post-probe delay period (namely, the *Probed, Unprobed* and *Probed/Unprobed* geometries) and found evidence for their behavioural relevance for all networks, irrespective of the retrocue validity condition. The *Unprobed* AI was a significant linear predictor of the network training latency, indicating that models that relied on a more parallel *Unprobed* geometry needed to complete more training iterations than those that orthogonalised the uncued item subspaces. Our interpretation of this surprising result follows that outlined in Chapter 2. Namely, we suspect that the random initialisation endowed some networks with a more parallel *Unprobed* geometry. If this geometry happened to be aligned with the readout subspace from the start, the networks would be more likely to respond according to the unprobed item and would therefore need more training iterations to overcome such response bias. Interestingly, we also found evidence in favour of the behavioural relevance of the *Probed* geometry. More specifically, the *Probed* AI was negatively predictive of the training latency, signifying a performance advantage in networks characterised by a more parallel geometry. Thus, a task solution that places the trial-relevant items into a common template space (and generalises over their irrelevant differences) is easier to learn.

Overall, it appears that a parallel *Cued* geometry forms part of an optimal task solution only when the retrocue is fully informative about the trial-relevant item (i.e., deterministic). Under such conditions, the network computes the decision value immediately after the

presentation of the retrocue and maintains information in this prospective format until the end of the trial. In contrast, we find that probabilistic cues do not trigger such a representational transformation into an action-optimised format in the RNN, in contrast to the theory proposed by (Myers et al., 2017b). It is currently unclear how this result bears in comparison to the activity in biological circuits. Previous human studies have yielded evidence for priority-based transformations of the cued information following the presentation of a deterministic retrocue (Yu et al., 2020; Muhle-Karbe et al., 2021). Furthermore, recordings from non-human primate PFC also provided evidence for the prospective coding of information, whereby the stimulus information is maintained in WM in a format that is more similar to the ensuing probe than to the original stimulus (Rainer et al., 1999). However, the extent to which humans process and maintain information in a prospective way has also been shown to depend on a range of factors, including but not limited to the reliability of the supplied contextual cue (reviewed in Braver et al., 2007). Future neurobiological studies should therefore directly test the neural geometry predictions from our models using an analogous paradigm in human subjects. Based on our results from Chapter 2, we believe the RNN an adequate model of information processing in macaque lateral PFC following deterministic retrocues. Thus, we considered it useful to extend it to generate experimental predictions for other conditions, such as the probabilistic paradigm.

Chapter 4. Contrasting the symbolic and connectionist accounts of rule learning

4.1 Abstract

It was previously observed that the speed of human Boolean category learning is well predicted by the extent to which the list of item-category assignments can be compressed into a simpler rule. This result was taken as evidence that human thought relies on innate logical primitives, as posited by the language of thought (LoT) account of cognition. However, connectionist models have recently enjoyed success in explaining aspects of human semantic categorisation behaviour, despite relying on an entirely separate set of assumptions. In particular, the order of semantic category acquisition in such models, from broadest to the most item-specific, and the mathematical theory behind it, bear resemblance to the observed sensitivity of human learning to Boolean category structure as described above. In this study, we thus assessed whether the latter finding was a prediction unique to the LoT account of cognition. To this end, we trained artificial neural networks on the same set of Boolean categories and assessed their behaviour (learning speed and robustness of performance to noise at test) as a function of the category logical structure. We found that, in line with our prediction, both features of network behaviour were well predicted by the 'compressibility' of a given category structure. Thus, our results challenge the previously drawn conclusion that such relationship is evidence for the symbolic character of human thought. In contrast, we show that a similar pattern of results is obtained with a connectionist system that relies on subsymbolic representations and a statistical learning mechanism.

4.2 Introduction

What is the nature of human thought? This question has been a topic of debate amongst philosophers and cognitive scientists alike for centuries, yet no consensus has been reached. With its origins tracing back to the rationalist school of mind (Rescorla, 2019), the 'language of thought' (LoT; Fodor, 1975) hypothesis posits that higher-order cognitive processes such as planning, reasoning and decision making are grounded in the use of a mental language. Under this account, information is represented in the form of symbols – that is, discrete tokens imbued with arbitrary meaning (Atkin, 2023) and thinking occurs via the manipulation of such entities according to a system of rules reminiscent of the syntax of natural language. A prominent feature of this view is its commitment to the innateness of both mental representations and processes (Fodor, 1975).

A contrasting account of cognition is offered by the connectionist framework. Its proponents view higher-order cognitive function as an emergent property of the operation of lower-level processes (McClelland et al., 2010). Instead of discrete symbols and rules, the mind is modelled as a complex web of simple undifferentiated units. Information is stored in the connection weights between the units and mental processing occurs through the gradual development of activity across the network. Importantly, neither mental representations nor processes are assumed to be innate, and instead are learnt through the gradual adjustment of connection weights as the network gains experience with its environment, emphasising the flexibility and adaptability of the connectionist system.

In the domain of category learning, rule-based accounts such as the classical theory (Bruner et al., 1956; Smith and Medin, 1981) were initially found to offer an inadequate description of human performance (Smith and Medin, 1981; Ashby and Maddox, 1998). Natural

concepts were deemed to be 'fuzzy' in nature, with category structures not easily captured by logical rules, instead exhibiting varying degrees of coherence (Rosch, 1973). Accordingly, human learning and generalisation were observed to vary across between items from the same category set, with subjects exhibiting graded judgements of category membership (e.g., reporting a sparrow to be a more typical example of a bird than a penguin) (Rosch and Mervis, 1975; Barsalou, 1985). Subsequently, rule-based theories were largely replaced by similarity-based accounts such as the prototype (Posner and Keele, 1968, 1970; Reed, 1972; Rosch, 1973, 1975; Homa et al., 1981; Smith and Medin, 1981) and exemplar (Brooks, 1978; Medin and Schaffer, 1978; Estes, 1986, 1994; Hintzman, 1986; Nosofsky, 1986; Lamberts, 2000) theories, which have since enjoyed a tremendous amount of scientific attention.

However, recently there has been a resurgence of interest in rule-based explanations of human category learning, particularly in the domain of Boolean concepts (Feldman, 2003; Goodman et al., 2008, 2011; Piantadosi et al., 2016). This can be traced back to the seminal work of Feldman (2000), who reported that the acquisition of a range of Boolean concepts was sensitive to their respective syntactic structure. More specifically, categories that could be expressed as simpler rules (e.g., based on a single stimulus dimension) were acquired much faster than those that could not be condensed in this way. Furthermore, the sensitivity of human category learning to the logical structure of categories was taken as evidence that human thinking was itself governed by something akin to the rules of logic, as articulated in the LoT hypothesis.

On the other hand, connectionist models have also enjoyed considerable success in capturing key features of human category learning and categorisation performance. In particular, deep

neural networks have recently been used to explain phenomena such as exemplar typicality effects, progressive semantic knowledge deterioration in neurological conditions and the hierarchical acquisition of semantic concepts during development (with broader categories such as animals and plants acquired before more specific ones such as birds and fish, McClelland and Rogers, 2003; Saxe et al., 2019). Notably, the latter finding is reminiscent of the results from Feldman (2000), whereby more abstract categories (e.g., those based on a single rule that discriminated between all items) were acquired faster than those based on more complex rules (based on combinations of features) or those that required memorisation. Interestingly, a variant of a connectionist network has been previously shown to exhibit a similar sensitivity of learning to problem difficulty on a small subset of conditions examined in Feldman (2000) (Kruschke, 1992), hinting at the possibility that qualitatively different learning mechanisms can nonetheless give rise to similar observations.

In this chapter, we asked whether the sensitivity of behaviour to the logical structure of Boolean categories observed in Feldman (2000) was a prediction unique to the LoT hypothesis and could only be explained by the use of explicit rule-based strategies. Based on the results of Saxe et al. (2019), we suspected that a connectionist model trained with gradient descent (a statistical learning algorithm) would be similarly sensitive to the logical structure of categories. To test this prediction, we trained simple feedforward neural networks (NNs) on the same set of learning problems as used in Feldman (2000). We reasoned that the behavioural quantity measured in the human participants (proportion of correct responses after memorising the complete set of exemplar-category assignments at once) could be operationalised as the speed of category acquisition or the susceptibility of the learnt category representations to noise. Therefore, to test our prediction that NNs would

be similarly sensitive to the logical structure of Boolean categories, we analysed their learning speed and performance robustness to noise at test as a function of problem complexity. We found both measures to be significantly predicted by the logical structure of the problem, much like was reported for humans. To put this finding into a wider perspective, we compared it against the predictions from other benchmark theories of category learning, namely the prototype and exemplar models. Lastly, we asked to what extent human Boolean category learning could be explained by predictions generated from our connectionist model and the LoT framework.

4.3 Methods

4.3.1 Task structure

We adopted the same paradigm as described for human subjects in Feldman (2000). We explored the space of Boolean category structures characterised by *D* binary features and *P* positive examples (that is, examples included in the category, as opposed to the non-category), with D ∈ {3, 4} and P ∈ {2, 3, 4}. Each D[P] combination constituted a *rule family,* and the families differed in the number of unique *rule* (or logical) *types* they comprised. For instance, consider family 3[4] (see **Table 4.1**, the remaining rule families are shown in Supplementary Information for Chapter 4). There are 6 unique ways of assigning 4 stimuli (each comprised of 3 binary features) to the same category. All possible category structures within this three-dimensional stimulus space are isomorphic to one of these 6 rule types. The rule types can be expressed as Boolean formulas that are disjunctions of the positive stimuli (with each stimulus written as a conjunction of its feature values); a description termed the disjunctive normal formula (DNF). The DNF provides the list of all stimuli included in the category; however, in many cases, there exists a much more concise

description of the rule type. Feldman (2000) derives the shortest logical formulas for all rule types he studies and introduces a measure based on their length - Boolean complexity (BC). BC is quantified as the number of variables in the shortest description of the rule type and captures its 'compressibility'.

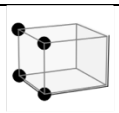
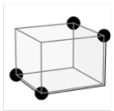
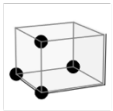
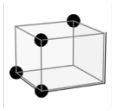
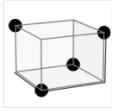
Rule type	DNF	Shortest formula	BC	Category structure 
I	$a'b'c' + a'b'c + a'bc' + a'bc$	a'	1	
II	$a'b'c' + a'b'c + abc' + abc$	$ab + a'b'$	4	
III	$a'b'c' + a'b'c + a'bc' + ab'c$	$a'(bc)' + ab'c$	6	
IV	$a'b'c' + a'b'c + a'bc' + ab'c'$	$a'(bc)' + ab'c'$	6	
V	$a'b'c' + a'b'c + a'bc' + abc$	$a'(bc)' + abc$	6	
VI	$a'b'c' + a'bc + ab'c + abc'$	$a(b'c + bc') + a'(b'c' + bc)$	10	

Table 4.1. Characterisation of rule family 3[4] adapted from Feldman (2000). Each rule type (denoted by a roman number; rows) is listed alongside its DNF, shortest Boolean formula derived by Feldman (2000), Boolean complexity (BC) and a visualisation of the positive exemplars. For increased readability, a conjunction $a \wedge b$ is written as ab, disjunction $a \vee b$ as $a + b$, and negation $\neg a$ as a'.

4.3.2 Model architecture

Each model was a three-layer, fully connected feedforward NN. The input layer consisted of *2 x D* binary units. Each stimulus dimension was thus encoded in a one-hot manner by a pair of input nodes. The hidden layer consisted of 25 units with rectified-linear (ReLU) nonlinearities, its activity given by the vector h:

$$h = ReLU(W_i u + b_h \ + \eta) \tag{1}$$

where W_i corresponds to the input-hidden weight matrix, u corresponds to the input to the network, b_h is the bias term and η is the noise term drawn from $\mathrm{N}(0, \sigma^2)$. The parameter σ was set to 0 unless otherwise stated. The hidden layer activity was readout by a linear decoder output layer consisting of 2 units (2.1) and converted into choice probabilities for the two categories with a Softmax function (2.2):

$$o \ = \ W_o h + b_o \tag{2.1}$$

$$y_i = \frac{e^{o_i}}{\sum_j^2 e^{o_j}} \tag{2.2}$$

where W_o corresponds to the hidden-output weight matrix, b_o is the bias term and y_i is the probability of making the category *i* response. We used the Xavier initialisation for all weights and biases.

4.3.3 Model training

We trained N=50 models on each rule type. Models differed in terms of their initial parameters and stimulus presentation order. Each model was reinitialised with the same weights prior to training on a new rule type and thus can be treated as a 'participant' that learns every new task without any transfer of previous knowledge. Networks were trained

in a supervised manner with stochastic gradient descent (with parameter updates performed after every trial) and cross entropy loss. Targets were the category labels. For most rule types, the two categories consisted of different numbers of exemplars, leading to an unbalanced training set. To prevent this from affecting network training, we weighted the loss between a given item-target pair by a category weight *w* equal to the ratio of the number of category stimuli and the total stimulus set size. However, we observed the same pattern of results when the category weight parameter was not used. We conducted a hyperparameter search over a range of different hidden layer sizes and learning rate values and observed no qualitative differences between conditions. The data reported in this chapter comes from networks containing 25 hidden units, with learning rate set to 0.1. In this setting, all networks were trained until the training loss value summed over the most recent epoch fell below a common threshold (set to 0.05).

4.3.4 Learning speed measure

To measure the learning speed of networks on different categorisation problems, we chose an arbitrary small loss value to use as a convergence threshold and calculated the total number of epochs it took for each network to reach it (number of epochs to convergence, NETC). The results reported here were generated with the threshold set to 0.05, but the overall pattern was not different for the different values tested.

4.3.5 Sensitivity to noise metrics

To assess how network performance was affected by the presence of noise, we evaluated the trained models (after freezing the weights and disabling gradient updates) on the training dataset under two conditions: without (*baseline*) and with white noise injected into the

hidden layer. Noise was drawn independently for each hidden unit from $\mathrm{N}(0, \sigma^2)$ (see Equation 1). We set σ to 3, which corresponded to three times the standard deviation of the hidden unit activity under baseline conditions. This noise level ensured that we could detect a difference in performance, but the networks could still perform the task with above-chance accuracy. The sensitivity to noise (NS) was then quantified as a percentage drop in test accuracy under noise (acc_{wn}), relative to baseline (acc_{bl}), conditions:

$$NS = \frac{acc_{bl} - acc_{wn}}{acc_{bl}} \cdot 100\% \quad (3)$$

4.3.6 Regression methods

To determine whether network behaviour was sensitive to problem complexity measures such as BC, we followed the same approach as reported in Feldman (2000). For each D[P] family, we averaged the dependent variable across all networks trained on a given rule type and fit a linear regression model to the resulting averages. Each regression model had a single predictor variable.

4.3.7 Prototype and exemplar models of category learning

The distance between a pair of category exemplars x_i and x_j was defined as:

$$d_{ij} = (\sum_{k=1}^{D} w_k |x_{i_k} - x_{j_k}|^r)^{1/r}$$

where w_k is the attentional weight to stimulus dimension k (set to 1/D), x_{i_k} is the stimulus value on dimension k, and r specifies the distance metric to be computed (set to 2, i.e.

Euclidean distance for both exemplar and prototype computations). The similarity of an exemplar to category A was then given by:

$$S_{iA} = \sum_{j \in A} e^{-d_{ij}}$$

Analogously, the distance and similarity between an exemplar x_i and category A prototype P_A were given by:

$$d_{iP_A} = (\sum_{k=1}^{D} w_k |x_i - P_A|^r)^{1/r}$$

$$S_{iP_A} = e^{-d_{iP_A}}$$

where P_A was defined as the average of all exemplars from category A. To compare the predictions of these formal models of category learning to BC, we calculated the ratio of the average within- and across-category similarities under each model.

4.3.8 Human participant task and data

Human data analysed in this chapter was kindly provided by Jacob Feldman and has previously been published in Feldman (2000). The dataset constituted the group means and standard errors of the mean for each rule type.

Participants were instructed that their task was to learn to categorise species of amoebas, which were characterised by different shape, number, size, and shading of their nuclei (these were the four possible stimulus dimensions). For each block (corresponding to a new rule type), they were expected to discriminate between the example and non-examples of a new species. At the beginning of the block, participants were shown a training screen listing all exemplars and their respective category membership. All examples of the new species were

listed at the top of the screen (and clearly labelled 'EXAMPLES'), whilst all non-examples were shown at the bottom of the screen (labelled 'NOT EXAMPLES'). The duration for which the training screen was shown was determined by the total number of exemplars for the rule family (set to equal 5*P* seconds), to equate the task difficulty between rule families. After viewing the training screen, the participants were shown a random sequence of all training exemplars and categorised them in turn without feedback. The number of correct responses was recorded.

Each subject learnt all rule types from a single rule family. The sample sizes were equal to 20, 22, 27, 44, 45 and 57 subjects for the 3[2], 3[3], 3[4], 4[2], 4[3] and 4[4] rule families, respectively. For each rule family, the entire set of rule types was repeated with new concepts (i.e., stimulus feature-response assignments) 8, 8, 3, 3, 2, and 1 time. This was done to equate the overall session length across different rule families, as well as maximise the number of datapoints for each participant.

4.4 Results

4.4.1 Learning speed of NNs is sensitive to Boolean complexity

In a striking correspondence to the results reported for human participants in Feldman (2000), we observed that BC was a positive linear predictor of the network learning speed (quantified as NETC) for all D[P] families examined (**Fig. 4.1E**; see also **Fig. 4.1B**). In comparison to the human data, BC explained more variance in the network learning speed for most families, with R^2 values ranging between 73.3% and 98.3% (the corresponding range for the human proportion of errors was 36.3% to 98.3%; see also **Fig. 4.1D**). The observed sensitivity of learning speed to BC was not purely driven by the difference in learning speed of linear and nonlinear problems (which often have non-convex solutions

and thus pose a greater challenge for optimisation algorithms like gradient descent), as there were systematic differences even between nonlinear problems (e.g., types V and VI in the 3[4] family, **Fig. 4.1E**). Moreover, like was observed for humans, there was further variation in the learning speed among rule types characterised by the same BC value (e.g., types III, IV and V in the 3[4] family), suggesting that perhaps, there might be other variables that affect learning speed in a systematic way in both humans and NNs.

4.4.2 Noise sensitivity in networks and Boolean complexity

Next, we asked what other features of NN behaviour might show a similar sensitivity to the complexity of problem structure. Intuitively, we would expect BC to negatively correlate with the robustness of network performance to perturbations. This intuition is based on the observation that higher-dimensional, more complex solutions to problems generalise less well (a property known as the bias-variance trade-off in the fields of statistics and machine learning; Luxburg and Schölkopf, 2011). We therefore tested how the performance of NNs after convergence was affected by the presence of random noise. We compared test accuracy of trained models under two conditions: baseline (which matched the training regime, where models were not subject to any noise) and an experimental manipulation whereby white noise was injected into the hidden layer. Our measure of choice was the difference between accuracy under the baseline and noise conditions, normalised with respect to baseline performance (a quantity we label *noise sensitivity*, or *NS*). The injection of noise degraded overall NN performance from baseline (see **Fig. 4.1C**). Importantly however, the extent of this degradation, as measured by NS, showed a positive linear association with the BC across all rule families examined. More specifically, BC explained 97, 77, 37, 83, 61, and 23% of variance in the NS for the 3[2], 3[3], 3[4], 4[2], 4[3], and 4[4] rule families, respectively

(see also **Fig. 4.1F**). Notably, the linear effect of BC on NS was somewhat weaker than the one described for network learning speed, as evidenced by lower overall R^2 values of the regression models. Note that models trained on rule families with more unequal category splits (such as the 2:6 category 1 vs 2 split in family 3[2]) were on average less sensitive to disruption by noise. This most likely reflects the higher baseline probability of being correct for such families (an agent randomly guessing the larger category 2 would be correct on 75% of trials). A similar pattern seems to be present with regards to the fits of the BC linear regression models – the R^2 values are lower for the rule families with more equal category splits. However, it is important to note that such families are also characterised by more rule types (e.g., 4, 6 and 19 types for rule families 4[2], 4[3] and 4[4], respectively).

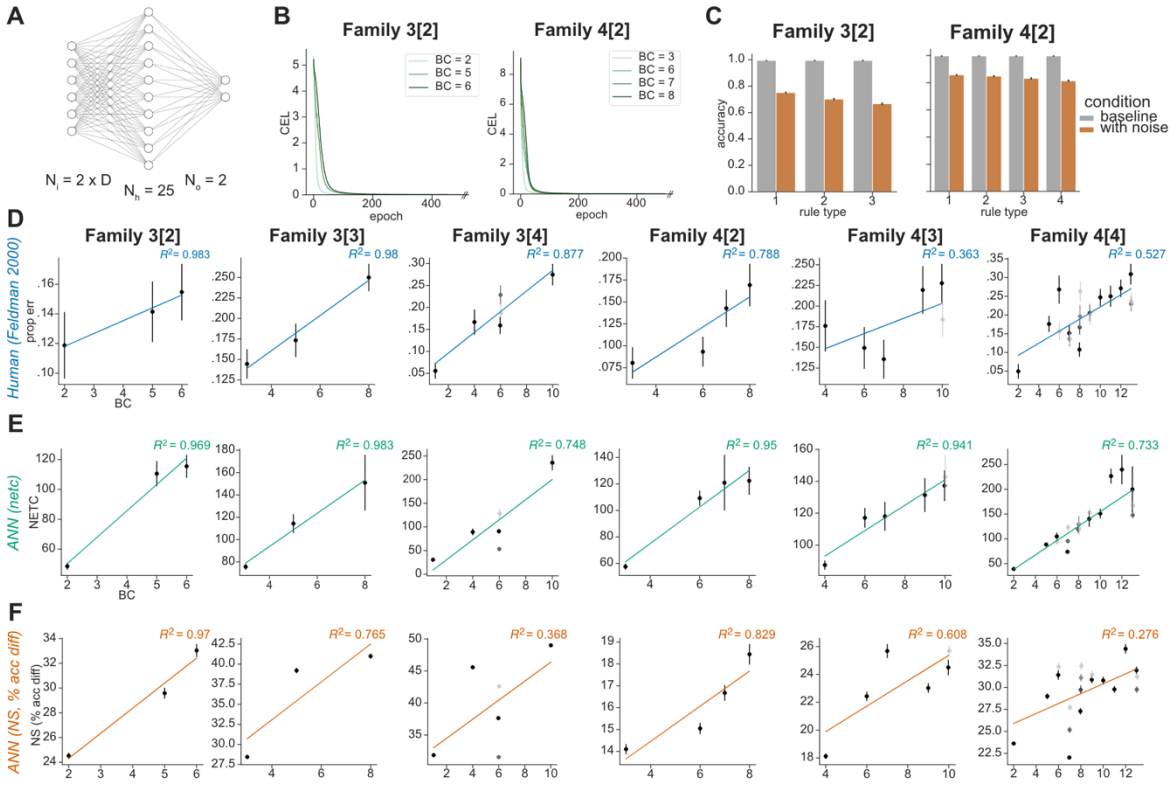


Fig 4.1. Learning speed and sensitivity to noise in networks are well predicted by Boolean complexity. A. Schematic network architecture. N_i – number of input units (twice the dimensionality of the problem D); N_h – number of hidden units (only 10 out of 25 units shown for clarity); N_o – number of output units. **B.** Training cross entropy loss (CEL) for

rule types characterised by different Boolean complexity values (different shades of green). Data shown for an example model for an example 3- and 4-dimensional rule family. Only the first 500 training epochs depicted, out of the 1000 completed by each model. **C**. Test accuracy under no noise (*baseline*, grey) and hidden noise (orange) conditions. Bars correspond to the average across all trained networks, errorbars show SEM. Data depicted for all rule types of two example D[P] families (3[2] and 4[2]). **D.** Mean +- SEM proportion of errors made for each distinct rule type by subjects from Feldman, 2000. Rule types characterised by the same Boolean complexity values plotted in shades of grey for better discriminability and easier comparison to the ANN data in E-F. Blue lines correspond to the linear regression fits predicting the mean proportion of error during the test block from the Boolean complexity. R^2 value for the regression model indicated in the top right corner. **E-F.** Analogous plots for the learning speed (quantified as number of epochs to convergence, NETC; green) and noise sensitivity (NS; orange).

As a control condition, we also assessed the performance on networks trained with hidden noise but evaluated without. In this setting, removing noise slightly improved overall test accuracy (note however that baseline test accuracy was already very close to ceiling; Fig. 2A). As before, we calculated the difference between baseline (same level of noise as experience during training) and no noise conditions, normalised by baseline performance and regressed it against BC. We found NS to be negatively predicted by BC – the improvement of performance under no noise conditions was more pronounced for problems with higher BC, even after accounting for any potential slight differences between base performance with normalisation.

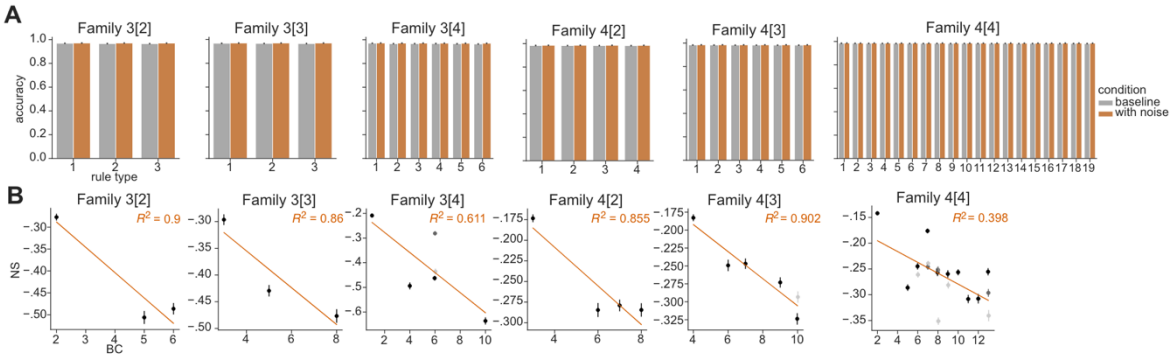


Fig 4.2. Test performance of networks trained with hidden noise (σ = 0.3). A. Comparison of the test accuracy under baseline (hidden noise with the same statistics as experienced during training; grey) and no noise (orange) conditions for every rule type of the 6 rule families examined. Bar height denotes the group average, errorbars correspond to SEM. **B.** Regression lines (orange) of best fit predicting the NS from BC of the categorisation problem for the 6 rule families examined. All conventions as in Fig. 4.1D-F.

4.4.3 Alternative models of categorisation

In the previous sections, we asked if the behaviour of a simple connectionist model would be sensitive to a putative marker of rule-based category learning mechanisms. We observed this to indeed be the case. Here we put this finding into a wider perspective by contrasting it against the predictions from other theories of category learning, which assume a distinctly different underlying mechanism. We focus on two prominent formal models: the prototype and exemplar models. Both models involve similarity computations and have been used to capture the differences in human behaviour to different members of a category (Ashby and Maddox, 2005). In the exemplar framework, categorisation decisions are based on the summed similarity of the item and all the items belonging to a given category (Brooks, 1978; Medin and Schaffer, 1978; Estes, 1986, 1994; Hintzman, 1986; Nosofsky, 1986; Lamberts,

2000). In the prototype model, the similarity of an item is instead compared to the average of all the category exemplars, i.e., the category prototype (Posner and Keele, 1968, 1970; Reed, 1972; Rosch, 1973, 1975; Homa et al., 1981; Smith and Medin, 1981). To derive a measure of the overall categorisation problem difficulty under each of these accounts (akin to the BC metric), we calculated the ratio of the average within- and across-category similarity values. Subsequently, we used these measures (i.e., the exemplar and prototype similarity ratios) as predictor variables in linear regressions predicting the NETC and NS, mirroring the BC analysis described in the previous sections.

Whilst both new predictors (the exemplar and prototype similarity ratios) explained a considerable amount of variance in learning speed (mean R^2 of 77.7 and 68.4% for the exemplar and prototype measures, respectively), the BC measure was found to best capture the network learning speed across all rule families (Fig. 3A), as evidenced by the highest R^2 values of models incorporating this predictor. However, a Kruskal-Wallis with the factor model type revealed only a trend-level significance result ($H(2)=4.53$, $p = 0.104$). The converse was true for noise sensitivity, where the prototype, exemplar and BC metrics explained on average 95.2, 94.3 and 63.6 % of variance across all rule families, respectively. An analogous Kruskal-Wallis test revealed significant between-group differences ($H(2) = 6.92$, $p = 0.031$). Bonferroni-corrected post-hoc tests showed the pairwise differences between BC and other models to be statistically significant (BC vs exemplar similarity ratio: $t(5) = -3.28$, $p = 0.015$; BC vs prototype similarity ratio: $t(5)=-3.38$, $p = 0.012$).

4.4.4 Comparison of learning speed in networks and humans

Previously, we characterised the behaviour of a connectionist model on a series of tasks previously investigated in humans. We have found that, much like was observed for human

participants, the behaviour of NNs showed features previously assumed to be diagnostic of rule-based processing. In this section, we ask a different question. Though human behaviour was found to be well predicted by Boolean complexity, there was also a degree of residual variance, particularly in the case of the categorisation problems with more features and thus a larger stimulus sets (i.e., the 4-dimensional rule families, see **Fig. 4.1D**). Thus, we asked to what extent the behaviour of human participants and NNs was similar. In other words, would the learning speed predictions derived from NNs capture human performance better than BC? To answer this question, we used the NETC as a predictor of human proportion of errors reported in Feldman (2000). We then compared the goodness of fit of this model (quantified by R^2) to the analogous model with BC as the predictor variable and found that, contrary to our prediction, NETC was a significantly worse predictor of human behaviour across the set of rule families examined (**Fig. 4.3C**; paired-sample t-test: $t(5) = -2.56$, $p = 0.975$, one-tailed).

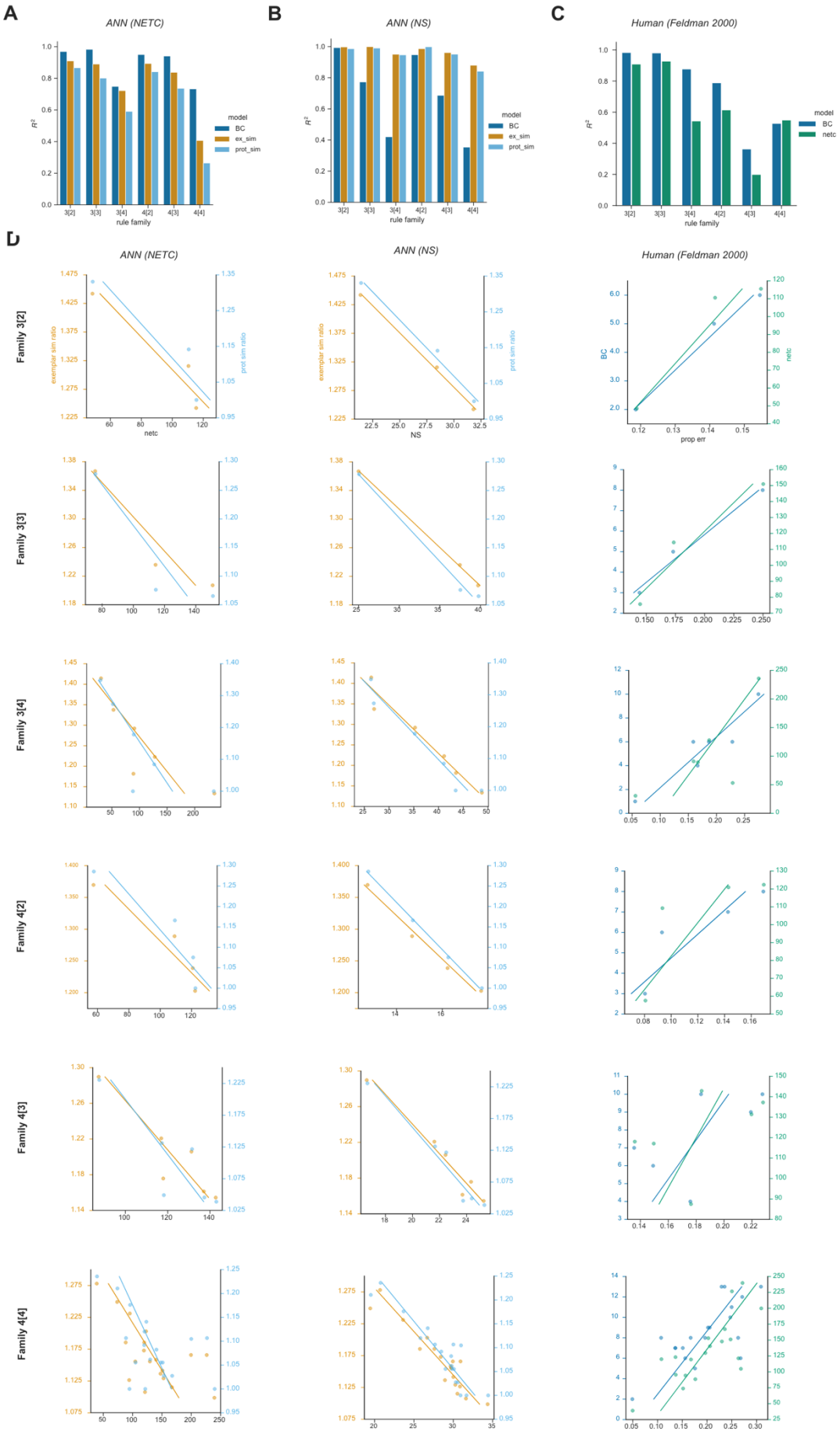

A
ANN (NETC)
B
ANN (NS)
C
Human (Feldman 2000)
R^2
rule family
3[2]
3[3]
3[4]
4[2]
4[3]
4[4]
model
BC
ex_sim
prot_sim
netc
D
Family 3[2]
Family 3[3]
Family 3[4]
Family 4[2]
Family 4[3]
Family 4[4]
exemplar sim ratio
prot sim ratio
netc
NS
prop err

Fig 4.3. Alternative accounts of category learning. A. Comparison of the R^2 values of linear regression models fit to the NETC for each rule family. The predictor variables for regression models denoted by colour: BC (navy), ex-sim (exemplar similarity ratio, light orange), prot-sim (prototype similarity ratio, baby blue). **B.** Analogous comparison of regression model fits to the NS data. **C.** Analogous comparison of linear regression fits to the human data from Feldman (2000). Models incorporating the NETC as predictor variables shown in green. **D.** Regression lines of best fit for models incorporating the exemplar similarity (light orange), prototype similarity (baby blue), BC (navy) and NETC (green) to the network (NETC, first column from the left; NS, second column) and human (third column) data. All conventions as in previous figures. Note each plot shows two independent variables plotted on separate y-axes, for compactness and ease of comparison.

4.5 Discussion

In this chapter, we reassessed some of the evidence put forward in favour of the LoT account of human Boolean category learning. We demonstrated that a simple connectionist network exhibited the same sensitivity of learning speed to Boolean complexity as was previously reported in humans, despite lacking explicit rule-based mechanisms. We then asked how well the behaviour of NNs was captured by alternative theories of human category learning. Finally, we compared the extent to which the LoT and connectionist accounts individually predicted human learning performance.

Consistent with our prediction, we observed that the learning speed in NNs was well predicted by BC across a set of 6 Boolean category families. On average, the BC linear regression model explained more variance in the learning speed of NNs than humans, which most likely reflects the absence of noise sources in the NNs (such as sensory or memory

noise). Of note, a recent study (Carcassi and Szymanik, 2022) provided corroborating evidence that the correlation between BC and learning effort in NNs can be observed for a variety of different LoTs, differing in terms of the logical operators they include. Our results further suggest that BC tends to predict the differences in the NN learning speed better than other problem difficulty metrics based on alternative theories of category learning (namely, the prototype and exemplar similarity ratio), though this difference did not reach the threshold for statistical significance. Overall, based on these findings, one might conclude that the BC metric captures a core feature of NN learning.

However, in the absence of explicit rule-based processing and hard-coded logical primitives, what mechanism might give rise to the observed association between learning speed and BC in NNs? Recent work on analytical solutions to the learning dynamics of NNs might provide an answer to this question. Saxe et al. (2019) demonstrated that the acquisition of semantic categories in NNs is driven by the singular value decomposition of the matrix capturing the category structure. The authors presented deep linear networks with item examples (such as living things: canary, dog, sunflower) and trained them to predict the item properties (e.g., 'can move, 'has feathers, 'has leaves'). Like was observed in human development studies (Keil, 1979; Mandler and McDonough, 1993), broader categories (such as animals and plants) were acquired before more specific ones (such as mammals and birds). Importantly, the mechanism behind this effect related to the correlational structure of the observed input-output correlation matrix (where inputs correspond to the items, and outputs to their properties). Networks decomposed this correlational structure into modes (or singular values) which captured the extent to which properties were shared by items from the training set. Importantly, the largest singular values (capturing the properties that discriminated

between the largest proportion of items, such as ‘has leaves’ and ‘can move’ which can be ascribed to most plants and animals, respectively) were acquired the fastest.

A similar mechanism could account for our findings, as problems with higher BC might be expected to be composed of more singular values. Given that the singular values vary in the speed at which they are acquired during learning proportionally to their order (Saxe et al., 2019), the acquisition of problems composed of more singular values would necessarily require a longer training time. Interestingly, a different line of research suggests that in the case of Boolean categories, NNs might also exhibit a structural bias towards solutions with low complexity (Mingard et al., 2020). The authors provide a theoretical and empirical demonstration that simple NNs initialised from symmetric weight distributions (such as a normal or uniform distribution) possess an a priori simplicity bias, whereby untrained models are biased towards solutions with lower entropy (preferring solutions in which the relative assignment of inputs to categories is more uneven). Future work could characterise to what extent the learning speed effect we report for the 6 Boolean category families examined might be a consequence of the principles governing network learning dynamics versus a simplicity bias at initialisation.

With respect to human performance reported by Feldman (2000), our results showed that the learning speed predictions based on the NN training results did not capture more variance in the human behaviour than BC. There are several explanations for this result. Firstly, human Boolean category learning might not rely on similar learning algorithms and representations as those used by NNs. However, an alternative explanation would pertain to the lack of granularity in the human data provided by Feldman (2000). Given the small number of datapoints and the variance in the mean estimates for each rule type, the

comparison between BC and NETC conducted at the group-level might have lacked sensitivity. Our intention was to address this point by collecting a new human dataset. Unfortunately, this did not prove possible within the timeframe of the DPhil degree.

In summary, our work suggests that the Boolean complexity is not a quantity best suited for discriminating between the LoT and connectionist accounts of Boolean category learning. In support of this, we demonstrate that the sensitivity of performance to BC need not imply the use of innate, symbolic representations and rule-based learning strategies. Instead, it can also be underpinned by distributed, graded representations that emerge through the application of statistical learning algorithms. As corroborated through other recent work (Carcassi and Szymanik, 2022), this result is not specific to the model architecture and logical primitives assumed in the LoT used in our experiment. The extent to which human categorisation behaviour is better captured by a connectionist or symbolic account remains an open question for future research. Interestingly, connectionist models have recently enjoyed successes in other domains traditionally thought to require symbolic computations, such as planning and decision making, language, reasoning, and mathematics (Brown et al., 2020; Dasgupta et al., 2022; Lample and Charton, 2019; Schrittwieser et al., 2020; Santoro et al., 2017). Such findings raise interesting questions about the degree to which the processes governing human intelligent behaviour might be innate versus learnt, as well as dependent on symbolic versus subsymbolic representations, continuing the long-standing debate that motivated our research.

Chapter 5. Unsupervised learning and generalisation of abstract task rules

5.1 Abstract

Humans can regulate their behaviour in a context-dependent way, which is thought to rely on the use of abstract rules defining the appropriate stimulus response contingencies. Such rules have been shown to be actively maintained in working memory. Interestingly, recent perspectives on working memory function suggest that its role might extend beyond simple maintenance and include context-dependent computations, thus implementing rule-based processing in the brain. In contrast to rule implementation, much less is known about how and when abstract rules are learnt directly from experience. Recent studies suggest that humans might have a natural tendency to search for abstract rules during learning, even when the learning problem does not require such a strategy. In this Chapter, I describe behavioural experiments that tested whether humans naturally build abstract rules when learning a simple instrumental task, and whether such tendencies can be further modulated by engaging context-dependent processing in working memory. Subjects learnt a set of four stimulus-response mappings that were described by an exclusive-or (XOR) structure and had the opportunity to transfer it to a new set of stimuli. Each stimulus consisted of two features, and we manipulated how they were presented to the participants in a between-subject design. Thus, stimulus features were presented jointly, in temporal succession, or separated by a working memory delay, for the subjects in the control group (CG), temporal segmentation (TS) and working memory (WM) groups, respectively. Our results showed that subjects in the control group exhibited RT markers of rule-dependent processing, suggesting that they had naturally decomposed the XOR problem into a two-step decision process. Furthermore, the extent to which individual subjects showed this tendency was

positively correlated with their learning speed improvement on the second learning problem. However, we found no evidence that our experimental WM manipulation encouraged the use of hierarchical strategies above the natural tendency to do so. Surprisingly, the magnitude of the RT markers of rule-dependent processing did not correlate with the transfer benefit on the second XOR problem in the WM or TS group. In agreement with previous studies, our results suggest that humans might have a natural tendency to learn abstract rules from experience and transfer them across contexts. However, the involvement of context-dependent processing in working memory remains an open question.

5.2 Introduction

The natural environment is complex and dynamic, and actions that lead to positive outcomes in one setting might not be appropriate under different circumstances. To thrive, intelligent agents must therefore be able to not only flexibly adjust their behaviour depending on the current context, but also learn such behavioural policies directly from experience.

For example, imagine you are busy at work and your mobile phone screen flashes with an incoming call notification. If the caller is your favourite (albeit *very chatty*) friend, you should wait until the working day is over to catch up. However, if the person trying to get in touch is your boss, rejecting the call might carry severe consequences. The reverse might be true when you are relaxing at home on the weekend – you might decide that the social interaction with your friend is exactly what you need, whilst discussing work matters can wait until Monday. To learn when it is appropriate to answer the call from your friend or boss, one can treat each new situation (i.e., the conjunction of the physical location and caller identity) as a unique case. However, an alternative strategy is to recognise that your location sets the behavioural *context* for which behaviours are appropriate and which are not, and

thus carries a special status. In other words, your workplace can be thought of as an abstract, *higher-order* rule which determines the set of actions one should take in response to specific stimuli (i.e., lower-order stimulus-response mappings). Whilst such a *hierarchical strategy* can seem overly complicated at first, it might carry a key behavioural advantage. More specifically, building and conditioning a whole behavioural repertoire on the higher-order concept of a 'workplace' can permit the future reuse of such knowledge should the circumstances change. For example, should we choose to start a new job, we only need to link our new office to the abstract 'workplace' rule to be able to generalise our knowledge to this new setting, thus obviating the need to learn everything 'from scratch'.

The mechanisms underlying such flexible, context-dependent behaviour have been extensively studied in both psychology and neuroscience under the moniker of cognitive control. One approach has been to use the task-cueing paradigm, whereby at the outset of the experiment, participants receive instructions about the nature of the rules governing the stimulus-response contingencies (e.g., high/low and even/odd number judgements (Sudevan and Taylor, 1987; Meiran, 1996)). Then, participants are further cued about the currently relevant rule at the beginning of each trial (e.g., via a visual cue), prior to seeing the target stimulus (such as a digit) and making their response (e.g., right key press if the digit is even, or left when it is low). Multiple studies using this paradigm have demonstrated that performance on trials characterised by a rule switch (from even/odd to high/low, or vice versa) is slower (and sometimes less accurate), relative to trials on which it stays the same (Monsell, 2003; Kiesel et al., 2010; Vandierendonck et al., 2010). This effect has been termed the 'switch cost' and is thought to stem from a combination of the reconfiguration of a new task-set (that is, the configuration of perceptual, cognitive, and motor processes as well as mental representations that enables the person to act according to the new task rule)

and interference from the old task-set (Kiesel et al., 2010; Vandierendonck et al., 2010). Neurally, task-sets have been linked to the prefrontal cortex in both monkeys (Sakai, 2008) and humans (Richter and Yeung, 2014). Importantly, they have also been shown to be independent of the perceptual features that accompany them (such as the specific cue instructing them), and thus constitute abstract rule representations (Woolgar et al., 2011).

In contrast to their implementation, relatively little is known about how or when such abstract rules are learnt directly from experience. Previous work suggests that when subjects are required to learn a large set of stimulus-response mappings which can be condensed down to a smaller set of abstract rules, participants are able to discover and use the second strategy to speed up learning (Badre et al., 2010). Furthermore, the search for and use of such rules is accompanied by sustained prefrontal activation (Badre et al., 2010; Crescentini et al., 2011). These findings suggest that when learning many simple stimulus-response mappings, participants might preferentially rely on abstract rules to reduce cognitive load.

It is also possible that participants would show a similar propensity to search for abstract rules when the task structure does not suggest their existence, and there is no immediate advantage to doing so (e.g., the working memory load is low). Interestingly, Collins and Frank (2013) found this to be the case, as evidenced by the behavioural markers of task-switching (including reaction time switch costs). The authors theorised that such predisposition to build abstract rules might be an adaptive mechanism which permits the future generalisation of knowledge to novel but related contexts. Accordingly, in their study, the strength of this tendency correlated with the degree of performance improvement whilst learning a new problem that shared the same structure. Further support for such innate predisposition to search for higher-order strategies comes from developmental studies,

which show that such tendencies can already be observed in infants and might involve prefrontal cortical activity (Werchan et al., 2015, 2016). Other studies directly contrasted performance between subjects instructed to use abstract task rules with those asked to learn a small set of stimulus-response mappings from trial-wise feedback. Importantly, these experiments did not find evidence for spontaneous rule induction in the latter group, even when the learning problem could be simplified by adopting such a strategy (Dreisbach et al., 2007; Dreisbach and Haider, 2008; Forrest et al., 2014). Overall, these results cast some doubt on the generality of findings (and conclusions) from Collins and Frank (2013), inviting further investigation into the precise conditions under which the uninstructed reliance on hierarchically-structured, rule-based strategies is observed.

In this Chapter, we were interested in investigating the conditions under which humans create and use abstract rules when learning a small set of stimulus-response mappings through trial and error. Firstly, we were interested in the extent to which such tendencies are innate, as proposed by Collins and Frank, (2013). Secondly, we were also interested in the potential role of working memory engagement in abstract rule formation, the motivation for which we review below.

WM involvement has been demonstrated even in simple stimulus-response learning tasks where it is thought to augment learning (Collins and Frank, 2012, 2018; Collins et al., 2017). Furthermore, prefrontal cortex activity and WM function have been previously linked to more prospective forms of learning that go beyond relying on simple stimulus-response associations (model-based learning, Lee and Seo, 2007; Smittenaar et al., 2013; Doll et al., 2016). Convergent neuroimaging evidence shows prefrontal activation during the search for and use of abstract task rules (Badre et al., 2010). Accordingly, WM functionality has long

been considered a key part of cognitive control, with a role in the active maintenance of abstract rules (Miller and Cohen, 2001; Wallis et al., 2001). However, more recent perspectives on WM argue that its function extends beyond simple maintenance and to include active manipulation of information to support goal-directed behaviour. One example is the 'activity-silent' framework (Stokes, 2015) that highlights the role of WM in state-dependent processing, which closely resembles the notion of hierarchical processing outlined in the first paragraph. That is, the rule maintained in a short-term change to synaptic plasticity actively biases the processing of incoming sensory information, enabling context-dependent switches between different behavioural policies (Buonomano and Maass, 2009). Based on these considerations, we hypothesised that separating the contextual cue from the subsequently presented target with a memory delay, thus forcing the subjects to encode it into WM, might prime the fronto-parietal network for context-dependent processing and thus promote the acquisition of hierarchical representations.

However, to be able to disentangle this effect from the possible influences of the segmentation of task events in time (which might cue subjects to the existence of a rule hierarchy) and trial length (and thus longer processing time), we included another experimental manipulation. More specifically, we reasoned that merely presenting the context and target cues in temporal succession (rather than jointly, as was done by Collins and Frank, 2013) might affect subjects' propensity to rely on abstract rules. This could be because events in real-world environments unfold over time, and thus temporal cues might provide important information about the nature of the underlying environmental dependencies. Thus, we expected that separating the stimulus features in time might promote hierarchical learning and/or constrain the nature of the hierarchy in the case of multiple candidate structures.

In this Chapter, I describe experiments that tested (1) whether human subjects have an innate predisposition to build abstract rules when trained on simple stimulus-response mappings and (2) the extent to which it can be promoted by engagement of context-dependent processing in WM. Subjects were assigned to one of three experimental groups: control (joint stimulus feature presentation, CG), temporal segmentation (TS) or WM and trained on a simple instrumental learning task. Following the acquisition of four stimulus-response mappings, participants completed an evaluation block in which they continued performing the task without feedback. This allowed us to assess the RT signatures of hierarchical processing independently of learning. Subjects were subsequently tested on a new learning problem with the opportunity to generalise previously acquired abstract rule knowledge. We found that subjects in the CG naturally exhibited RT markers of hierarchical processing, and their extent correlated with the magnitude of performance increase on a second related task. This suggests that subjects naturally induced abstract rules during the first problem and transferred them to the second task. In contrast to our predictions, we found that engaging context-dependent processing in WM only constrained but did not promote the formation of hierarchical rule sets, relative to both the CG and TS conditions. Furthermore, the use of such hierarchical processing in the TS and WM groups was not positively correlated with the transfer benefits when subjects were tested on a new problem with the same underlying structure.

5.3 Materials and Methods

5.3.1 Subjects

The experiments were approved by the University of Oxford Central University Research Ethics Committee (approval no. R50750/RE003). Subjects were adults recruited via the online platform Prolific. We set an acceptance criterion of 70% accuracy during the last 30% of trials of Baseline learning phase of the experiment (see section 5.3.2 Experiment Protocol). For Study 1, we continued to recruit subjects until we reached an N = 70 in each group. We collected data from 274 subjects (112 female/162 male) in total, and included 210 subjects in the analysis, majority of which were between 21 and 30 years old (for more details on the final inclusion numbers and other statistics see Figure S1 and Table S1 in Supplementary Information for Chapter 5). Further exclusions were made for some analyses and are specified in the relevant sections. Study 2 was conducted for validation purposes. We continued to recruit subjects until we reached an N = 100 in each group (we included only the TS and WM conditions). We collected data from 269 subject in total (109 female/155 male). For more details, refer to Figure S2 and Table S2 in Supplementary Information for Chapter 5.

5.3.2 Stimuli

Each stimulus consisted of a top and bottom object. Objects were black icons drawn from 2 categories: fruit and vegetables. Objects from one category were used during the Baseline learning phase, and from the other one – during the Transfer phase (see section 5.3.3 - Experimental protocol below). Category order and top/bottom assignments were randomised between participants. The whole object set used in the experiment is shown in **Fig. 5.1A**.

5.3.3 Experiment protocol

The experiments were written in JavaScript with the Raphaël Vector Graphics Library and run in forced-fullscreen mode. The participants were instructed that their objective was to learn the correct response mapping for each visual stimulus. They were offered a monetary performance-based bonus to encourage task compliance and engagement. In Study 2, subjects were also offered a 'speedy performance bonus' contingent on reaching the accuracy criterion and dependent on how they ranked within their cohort. The experiment from Study 1 consisted of 3 distinct phases: Baseline learning, Evaluation, and Transfer. During Baseline learning, subjects used trial wise feedback to learn the response mappings for a set of 4 visual stimuli. During the Evaluation phase, they continued performing the same task, but without receiving trial-to-trial feedback. Finally, in the Transfer phase of the experiment, subjects saw 4 new visual stimuli constructed from the previously unseen object category and were asked to learn the appropriate stimulus-response contingencies for the new stimulus set.

At the beginning of the experiment, participants were instructed that there were 2 possible response rules ('Left' and 'Right') and asked to learn which one was correct for each top-bottom object combination. Importantly, every top object and every bottom object were part of both response rules. For example, the 'banana' top object from the example learning set in **Figure 5.1C** was part of the 'Left' rule when combined with the 'pear' bottom object. However, it was part of the 'Right' rule when combined with the 'apple' bottom object. Therefore, the learning problem had an exclusive-or (XOR) structure, but this was never explicitly revealed to the participants. Importantly, this structure is inherently hierarchical, in the sense that one object can be viewed as a higher-order context cue, and the other one

as a lower- order target, constraining higher- and lower-order action policies, respectively. However, representing the learning problem in this way is not essential for successful task acquisition – subjects can also treat each top-bottom object combination as a separate entity. This corresponds to a 'flat' strategy, or 4 unique 'AND' logic gates. Importantly, despite being associated with different learning exemplars, both Baseline and Transfer learning problems shared the underlying XOR rule structure. Therefore, if the participants discovered it during Baseline learning, they could leverage this knowledge to bootstrap their transfer performance. Trial sequences for each phase of the experiment were pseudo-randomised between participants. Specifically, we generated 2 unique pseudo-random trial sequences. Each subject saw one of the sequences during Baseline learning, and the other one during Transfer. The Evaluation phase sequence was drawn randomly from a set of 2. The numbers of subjects for each possible sequence order and combination were matched. During a given experiment phase, each stimulus was presented 20 or 21 times (giving a total of 81 trials). All trial transitions were equiprobable.

Subjects were randomly assigned to one of 3 experimental groups: control (CG), temporal segmentation (TS) and working memory (WM). The basic structure of the experiment described above was shared across all conditions, which only differed in terms of the trial structure (shown in **Fig. 5.1B**). In the CG condition, at the start of the trial subjects were presented with both stimulus objects and allowed to make their response immediately. In contrast, in the TS and WM groups, participants were first presented with the top stimulus object (from now on referred to as the *context cue*), followed by a filled (object stayed on the screen) or unfilled (blank screen) delay, respectively. After the delay, the bottom object (referred to as the *target cue*) was shown on the screen and subjects were free to give their response. Trial timings were matched in the TS and WM conditions and were as follows:

cue presentation - 500ms, delay - 750ms, feedback - 1000ms, ITI - 1000ms. In all conditions, the experiment was self- paced, but subjects were encouraged to only take breaks in-between phases. In the CG condition, the cue presentation length was therefore entirely dependent on the timing of the subject's response. The ITI and feedback timings were the same as in the other conditions.

The experiment from Study 2 was in most respects identical to that from Study 1, with the only differences being a longer evaluation phase (97 trials), unique pseudo-randomised trial sequences for each subject, and only 2 experimental groups: TS and WM.

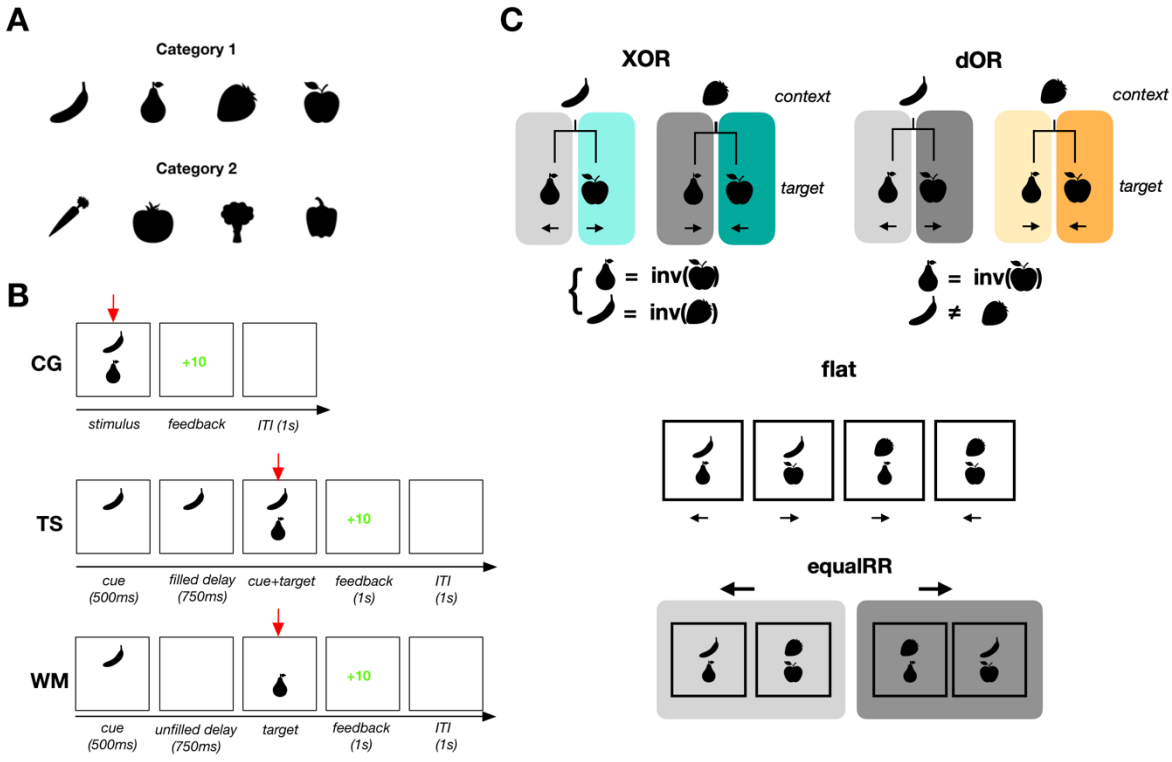


Figure 5.1. Task design. A. Object sets used in the experiments. **B**. Trial structure for each of the experimental groups. Red arrow corresponds to the trial stage at which subjects were allowed to make a response. **C.** Possible behavioural strategies used to solve the Baseline learning problem. Shaded boxes indicate abstract state groupings. **XOR**: Participants treat

one of the objects as the higher-order context cue, and the other one as the target cue. Targets within a given context are characterised by opposite response mappings (shaded boxes with the same opacity), and so are the corresponding targets in different contexts (shaded boxes with the same hue). **Double-or (dOR)**: Participants group together stimuli that share a context object, but do not capitalise on the inverse relationship between the two contexts. **Flat:** each stimulus is treated as a separate entity (i.e., a conjunction of the top and bottom object) and forms a state. **Equal response ratio (equalRR):** participants bind together stimuli that share a response mapping creating two abstract states.

5.3.4 Analysis

Data was analysed in MATLAB (R2018b, Mathworks), with the exception for the ANOVA, logistic regression and Bayesian statistics analyses, which were conducted in JASP (Version 0.11.1). We excluded all trials with reaction times (RTs) longer than 10 seconds and trials where the RT was more than 3 standard deviations above the log-mean for a given subject. For the RT analyses reported in sections 5.4.1 and 5.5.1., we excluded all error trials and trials following an incorrect response. Furthermore, we also excluded subjects missing more than 5 trials for either of the trial categories (i.e., cue-switch and target-switch trials, see **Fig. 5.2A**).

5.4 Results of Study 1

Subjects were trained on two consecutive learning problems (which we refer to as the Baseline and Transfer problems, respectively), which involved independent stimulus sets but shared an underlying abstract XOR task structure. Subjects were randomly assigned into one of three conditions, which differed in terms of the trial structure and corresponded to the control group (CG), task segmentation (TS) group, and working memory (WM) group

(see **Fig. 5.1B**). Our hypothesis stated that presenting the two objects that constituted a task stimulus in temporal succession rather than jointly (as was experienced by subjects in the CG) would cue the subjects to the existence of a rule hierarchy and/or shape its nature. Furthermore, we expected that introducing a WM delay in-between the two task events would further encourage hierarchical processing by priming the frontoparietal network for context-dependent processing. This would have further consequences on the learning of the second (Transfer) problem.

Firstly, we note the possible behavioural strategies subjects could have used to solve the Baseline learning problem. As noted above, both Baseline and Transfer problems had an exclusive-or (XOR) structure, but this knowledge was never revealed to the participants, and neither was it essential for the successful acquisition of either task (given the small set sizes). Furthermore, learning about each stimulus-response mapping separately from the others (i.e., in a **'flat'** way, **Fig. 5.1C**) was in principle just as effective during Baseline learning. However, abstracting the problem structure during Baseline learning and applying it during the Transfer phase would provide performance benefits over the flat strategy at that stage of the experiment. More specifically, our task was designed to encourage the subjects to structure the stimulus/state space in a hierarchical fashion, whereby one of the objects is treated as a higher-order context cue, and the other one as a lower-order target cue. Having applied this two-step process should allow the participant to realise that: 1) within a given context, the two target cues are associated with different responses and 2) the two contexts are inversely related, i.e., the response mappings for the targets under context 1 are reversed in context 2 (we refer to this strategy as the **XOR** structure, see **Fig. 5.1C**). Reapplying this abstract problem structure during the Transfer phase should in turn permit rapid learning by the ways of inference - equipped with the knowledge about how the 4 stimulus types (context

1-target 1, context 1-target 2, etc.) relate to one another, the participants should be able to infer all the new stimulus-response mappings on the basis of a single trial outcome. However, it is also possible to assign the different stimulus objects to different levels of hierarchy but fail to capitalize on the inverse relationship between contexts. We will refer to this somewhat myopic version of the XOR strategy as 'double-OR' (**dOR, Fig. 5.1C**). Under this model, the participant treats the different contexts as separate entities rather than inverses, which limits the extent of inference that can be made on the basis of a single trial and thus slows down transfer learning. Finally, the XOR and dOR strategies both structure the stimulus (state) space hierarchically with regards to the stimulus objects. However, it is also possible to perform state abstraction based on the frequency of the observed response mappings whilst treating both stimulus objects as equivalent (**equalRR, Fig. 5.1C**). A subject using this strategy would discover that, within a learning phase, two out of the 4 stimuli are paired with the same key response (e.g. 'left'), and the other two - with the opposite response. This would again allow the inference of correct response mappings for the unseen stimuli during transfer, albeit requiring more observations than the XOR strategy.

We begin by analysing subjects' reaction times during the Evaluation phase of the experiment, to determine whether there is evidence that subjects adopt a hierarchical learning strategy to solve the Baseline learning problem. Next, we focus on the learning of the Transfer problem, by comparing its acquisition speed to the Baseline problem (*transfer benefit*), as well as analysing the response patterns subjects make on the second trial. Lastly, we test whether there is an association between the RT signature of hierarchical learning (RT switch cost) and the extent of the transfer benefit. Across all the analyses, we expect to see a graded pattern of results, whereby the markers of hierarchical strategy use and

knowledge transfer will be more pronounced in the TS group than in the CG group, and even more so pronounced in the WM group.

5.4.1 RT signatures of hierarchical learning

To quantify and compare the RT signatures of hierarchical processing between conditions, we first categorised each trial based on how the stimulus objects shown compared to those of the trial immediately preceding it (see Fig. 5.2A). This gave rise to 4 trial categories: 'no switch' (NS), 'target switch' (TS), 'context switch' (CS) and 'double switch' (DS). For each subject, we then calculated the median RT within each trial category. These measures were then used to construct the *RT switch cost* according to the following formula:

$$RT_{cost} = \frac{median_{CS} - median_{TS}}{median_{NS}} \cdot 100\,\%$$

Given previous findings from task switching literature (namely, the RT cost on rule-switch trials, Monsell, 2003; Kiesel et al., 2010; Vandierendonck et al., 2010), we predicted the following: if a participant used a hierarchical strategy (XOR or dOR), they should exhibit higher RTs on CS trials than on TS trials, leading to a positive RT cost. In contrast, a participant employing a 'flat' strategy (i.e., treating both stimulus objects as equal) should not exhibit such difference. Therefore, if subjects used the hierarchical strategy, we would expect their respective group mean RT cost to be significantly larger than zero. In line with our predictions, mean RT cost was not significantly different from zero in the CG (Wilcoxon signed-rank test: $W(51) = 0.39$, $p = .349$, one-tailed), unlike the TS and WM groups (TS: one-sample t-test: $t(58) = 4.34$, $p < .001$; WM: Wilcoxon signed-rank test: $W(58) = 4.01$, $p < .001$; see also **Fig. 5.2B**).

The RT cost analysis described above had two caveats owing to the differences in the trial structure between the CG and other conditions. Firstly, because the CG subjects could make their response immediately after the trial start, unlike their TS and WM counterparts, the median RTs were significantly different between the groups (one-way ANOVA on log RTs: $F(2) = 16.88$, $p < .001$). More specifically, the median RTs in the CG (1.23s, N = 52) were significantly longer than in either the TS or WM groups (TS = 0.93s, N = 59, WM = 0.85s, N = 59; Tukey-corrected post-hoc tests: CG vs TS $t(109) = 4.49$, $p < .001$; CG vs WM $t(109) = 5.51$, $p < .001$). This difference meant we could not directly compare the RT cost values between the temporal delay conditions and the CG. However, the comparison between the TS and WM groups was not subject to similar difficulties, as their respective trial structures were more closely matched. Accordingly, we observed no significant difference in the log RT magnitudes between those groups (TS vs WM post-hoc t-test: $t(116) = 1.06$, $p = .154$). We therefore tested the second part of our hypothesis, which stated that the WM manipulation should promote hierarchy use over and above the effects of temporal segmentation alone. However, there was no such effect (median TS RT cost = 14.26, median WM = 9.43, Mann-Whitney U test: $U = 1713$, $p = .453$, $N_1 = N_2 = 59$). To be able to disambiguate between a true null finding and a false negative, we also performed a Bayesian Mann-Whitney test, which revealed moderate evidence for the null hypothesis of no difference between groups ($W = 1718$, $BF_{10} = 0.216$).

The second caveat of the RT cost analysis was as follows. We expected the sequential stimulus object presentation in the TS and WM groups to align which stimulus object was treated as the context (top object) and which as the target cue (bottom) across subjects. However, participants from the CG could, in principle, use the opposite object-level assignment (which would be consistent with the findings of Collins and Frank, 2013). Such

subjects would still show a non-zero RT cost effect but in the opposite direction to that assumed in the analysis, which could lead to a false negative result of the positive effect contrast. Therefore, to account for the effects of this factor on our ability to compare the RT signatures of hierarchical processing between groups, we constructed a metric which was agnostic to the direction of the RT cost effect: RT cost consistency. This analysis only assumed that subjects relying on a hierarchical strategy should show a consistent direction of the RT cost effect. For each subject, we performed 10 independent data splits. We calculated the RT cost values for each data half and compared their signs, coding matches as '1' and mismatches as '0'. We then averaged across the 10 data splits, giving us a final RT cost consistency score for each participant. A subject showing a perfectly consistent RT cost effect would therefore be assigned a score of 1, and a subject showing a random RT cost effect - a score of 0.5. To test whether subjects in each group showed a consistent RT cost effect, we tested if the medians of the group-wise consistency scores were significantly larger than the hypothesised value of 0.5. Subjects in the CG showed a significant positive effect (Wilcoxon signed-rank test: $W(51) = 767.5$, $p = .03$, one-tailed), as did subjects in the TS (Wilcoxon signed-rank test: $W(58) = 1048$, $p = .01$, one-tailed), whereas subjects in the WM condition showed a trend-level effect (Wilcoxon signed-rank test: $W(58) = 795.5$, $p = .063$, one-tailed). However, we did not observe a significant difference between groups (Kruskal-Wallis test: $H(2) = 0.20$, $p = .907$; see also **Fig. 5C**). Thus, whilst we found that subjects in all groups showed a consistent direction of RT switch cost effects, we did not

observe the strength of such tendency to differ between the experimental conditions.

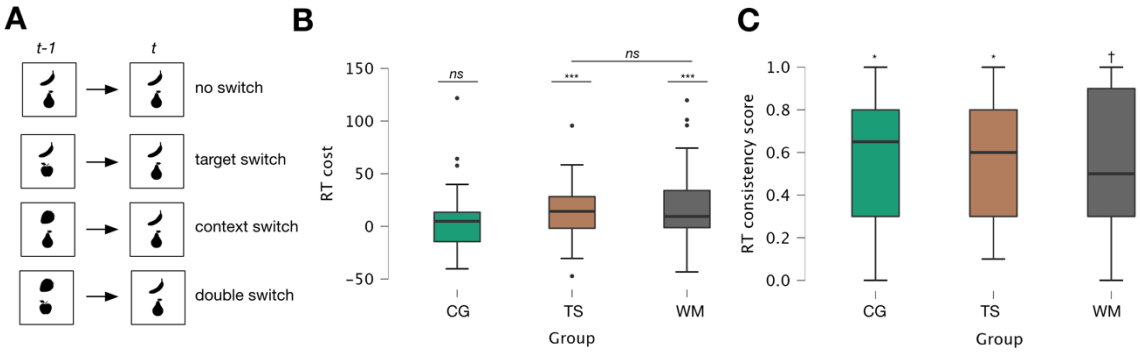


Figure 5.2 RT signatures of hierarchical strategy use. A. Trial type categorisation under the stimulus object-based hierarchy where the top object signifies the context cue. If neither stimulus object changes from trial t-1 to trial t, the trial is categorised as 'no switch'. If only the target cue (bottom object) changes, the trial is categorised as a 'target switch' trial. If only the context cue (top object) changes, the trial belongs to the 'context switch' category. If both objects change, the trial is labelled a 'double switch' trial. **B.** Boxplots of the RT cost scores for each group. Filled circles are outlier data points. **C.** Boxplots of the switch consistency scores for each experimental group. Significance values correspond to the contrasts described in the main text (ns: non-significant, ***: $p < .001$, *: $p < .05$, †: $p < .1$).

5.4.2 Performance benefits during transfer

We now shift focus to the Transfer learning performance. We start by comparing accuracy between the Baseline and Transfer problems, expecting the latter to be acquired much more quickly (we will refer to this as the *transfer benefit*). However, it is important to note that a relative increase in the speed of acquisition of the Transfer problem might not necessarily be a consequence of transferring the abstract XOR problem structure across tasks, as per our hypothesis. Instead, it might reflect non-specific effects such as an improvement in low-level motor skills. However, there is no a priori reason to expect the extent of such non-

specific effects to differ between the experimental groups. Therefore, we focus on the second part of our hypothesis which states that the magnitude of the transfer benefit should be highest for the WM group, followed by the TS and CG, reflecting the decreasing use and transfer of abstract problem structure.

To test this prediction, we binned the trial-wise accuracies from each learning phase into bins of 10 trials and entered the data into a Mixed-Model ANOVA with 3 factors: (1) a repeated measures (RM) factor Trial Bin (8 levels), (2) a RM factor Phase (2 levels: Baseline and Transfer learning), and (3) a between-subject factor Experimental group (3 levels: CG, TS, WM). This analysis revealed a main effect of Phase ($F(1) = 138.01$, $p < .001$), Trial Bin ($F(5.97) = 146.64$, $p < .001$, Greenhouse-Geisser corrected) and a Phase x Trial Bin interaction ($F(5.62) = 24.19$, $p < .001$, Greenhouse-Geisser corrected), but no main effect of Group ($F(2) = 0.43$, $p = .653$). The other two-way interaction terms (Phase x Group, Trial Bin x Group) were not statistically significant. The Phase x Trial Bin x Group interaction term was near the threshold for statistical significance ($F(11.24) = 1.71$, $p = .064$, Greenhouse-Geisser corrected) and corresponded to our experimental prediction of a difference in the extent of transfer benefits between the experimental groups. To investigate these results further, we used a combination of simple contrasts and simple main effects analyses. Firstly, a simple contrast revealed a significantly higher average accuracy during Transfer than Baseline phase ($t(207) = 11.75$, $p < .001$). Accuracy also increased as a function of time on task (linear contrast: $t(1449) = 31.04$, $p < .001$, quadratic contrast: $t(1449) = -7.27$, $p < .001$). To gain some insight into the trend-level three-way interaction, we looked at the main effect of Phase within each Trial Bin, separately for each group. This revealed a significant difference between Transfer and Baseline performance for every trial bin in the CG (all p-values < .05). However, this difference was more short-lived in the TS

and WM groups, where the transfer performance benefit seemed to be reliable mostly in trial bins 1-5 (TS: all p-values < .01; WM: all p-values < .001).

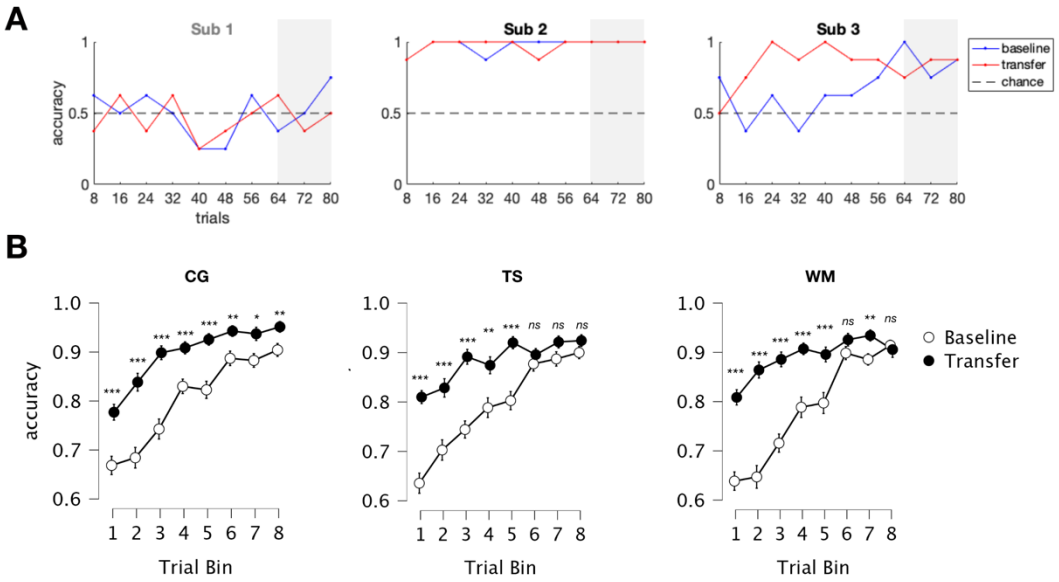


Figure 5.3 Comparison of the Baseline and Transfer learning performance. A. Data from 3 example subjects from the CG, each datapoint is an average of 8 trials. Accuracy data from the last 24 trials (30%) of baseline learning was used for calculating the inclusion criterion and corresponds to the grey shaded box. Subject 1 did not reach the criterion and was excluded from further analyses. **B.** Group-average learning curves for Baseline (white) and Transfer (black) learning. Average accuracy ± SEM plotted against trial bin number. Significance values correspond to the results of the simple main effects analysis described in the main text.

5.4.3 Transfer choice patterns

In the previous section, we compared learning performance between the Transfer and Baseline phases. We were interested in contrasting the magnitude of the transfer benefit across the three experimental conditions, expecting it to increase from CG to TS and WM groups. However, as noted before, this analysis did not directly address the hypothesised

underlying cause of the transfer benefit, namely - hierarchical processing. In this section, we analyse subjects' response patterns on the second trial of the Transfer phase, in an effort to disambiguate about the behavioural strategies subjects used to solve the Transfer problem.

The previously introduced behavioural strategies (flat, equalRR, dOR and XOR) differ in terms of the amount of information shared between different trial types, which in turn restricts the number of new stimulus-response mappings that can be inferred on the basis of a single trial outcome. Adopting the same trial classification as described for the RT_{cost} analysis (shown in **Fig. 5.2A**) and applying it to the second-seen unique trial of the Transfer phase, we can compare the predictions for response probabilities under each strategy. Focusing on the *target-switch* and *context-switch* conditions, we see that only the XOR model predicts significantly higher-than-chance accuracy on both of these trial types (**Fig. 5.4A**). Therefore, if we observe this pattern in the group-level data, we can conclude that a proportion of subjects within the group is using the XOR strategy for transfer. Importantly, subjects adopting a simple win-stay / lose-switch heuristics would be expected to make the opposite responses to those predicted by the use of the XOR strategy (i.e., the accuracy on those trials should be significantly below chance).

Thus, we grouped subjects according to the classification of the second-seen unique Transfer phase trial they experienced and compared their aggregate choice patterns against chance level (50%), separately for each group. We first focused on the *context-switch* trials, which are uniquely diagnostic of the XOR strategy use. We found significantly above-chance accuracy in all experimental groups (binomial test results: **CG** group: proportion = 0.68, N = 44, p = .011; **TS** group: proportion = 0.68, N = 37, p = .023; **WM** group: proportion = 0.65, N = 37, p = 0.049; see also **Fig. 5.4B**, dark grey bars). An analogous analysis

concerning the *target-switch* trials returned the same pattern of results (**CG** group: proportion = 0.73, N = 26, p = .015; **TS** group: proportion = 0.73, N = 33, p = .007; **WM** group: proportion = 0.75, N = 33, p = .004; see also **Fig. 5.4B**, light grey bars). Next, to compare these proportions between groups, we fed the data into a logistic regression with accuracy as the dependent variable, and group and trial type as predictors. Overall, the model did not explain a significant proportion of variance in the data (McFadden $R^2 < .001$, $p \sim 1$). We did not observe a significant effect of the trial type (odds ratio = 1.0, 95%CI = [0.31, 3.25]). The analysis also did not reveal any significant between-group differences (TS: odds ratio = 0.93, 95%CI = [0.30, 2.87];WM: odds ratio = 1.05, 95% CI = [0.35, 3.13]) or group x trial type interactions (TS x Trial type: odds ratio = 1.19, 95%CI = [0.24, 6.02]; WM x Trial type: odds ratio = 0.91, 95%CI = [0.18, 4.68]). Collectively, these results suggest the existence of subjects using the XOR strategy in all experimental groups, with no difference in their relative proportions between the CG, TS, and WM conditions.

5.4.4 Association between the hierarchical strategy use during Baseline problem and magnitude of transfer benefit

In our last line of analysis, we sought to answer the question whether the extent of hierarchical strategy use during the Baseline problem (as quantified by the RT cost metric) was positively associated with the magnitude of the transfer benefit shown by individual subjects. To this end, we correlated the absolute RT cost with the transfer benefit (defined as the difference in average accuracy between Transfer and Baseline learning phases). We found a significant positive correlation between the two measures in the CG (Spearman r = .37, p < .003, one-tailed, N = 52), but no such association in the TS (Spearman r = .11, p = .200, one-tailed, N = 59) and WM (Spearman r = -.24, p = .964, one-tailed, N = 59) groups (see also Fig. **5.4C**). Excluding outlier datapoints did not change this pattern of results.

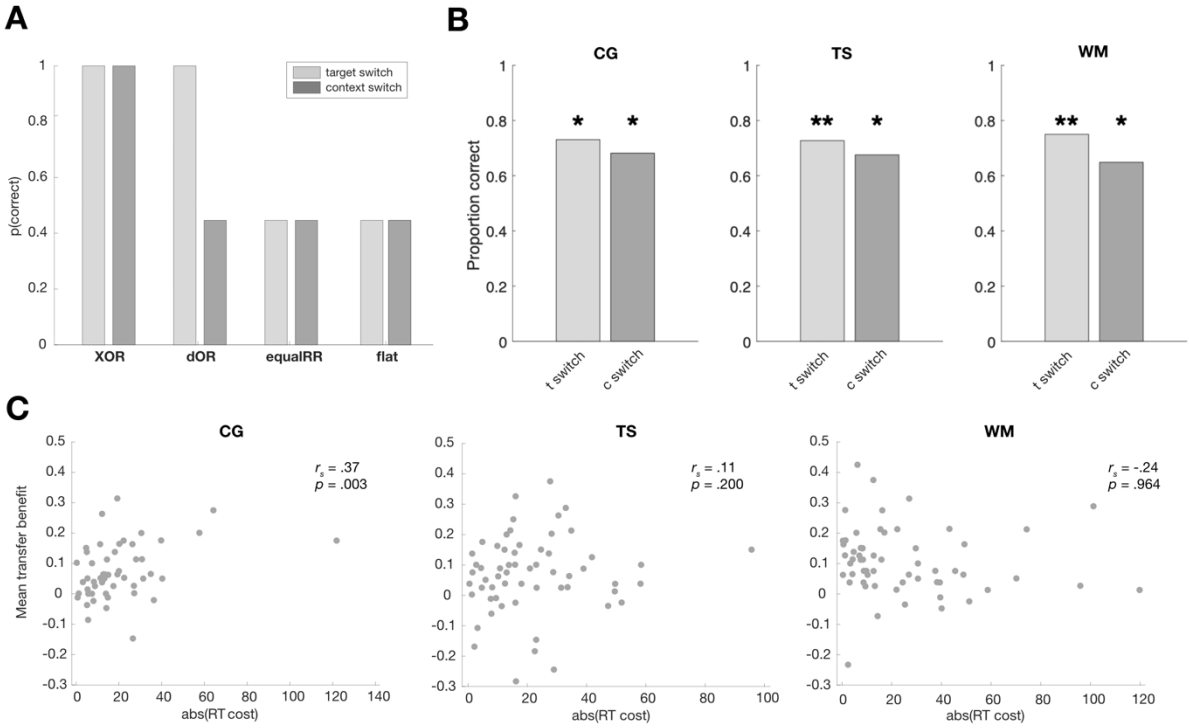


Figure 5.4. Reuse of the XOR problem structure to bootstrap Transfer learning.

A. Predicted probability of being correct under each behavioural strategy (XOR, dOR, equalRR and flat) on target-switch (light grey) and context-switch (dark grey) trials. Predictions assume no noise. Group-averaged accuracy values on second-unique transfer trial should approach the predictions proportionally to the number of subjects using a particular strategy within the group**. B**. Corresponding observed group-average proportion correct values on target-switch (t switch, light grey) and context-switch (c switch, dark grey) trials. Significance values correspond to a binomial test testing if the observed proportion is significantly greater than chance (0.5). Sample sizes were (from left to right): 26, 44, 33, 37, 33, 37**. C.** Scatterplots showing the mean transfer benefit (y-axis) against the absolute RT cost (x-axis) for each experimental group. Results of Spearman correlations between the two measures indicated on each panel.

5.5 Results of the replication study (Study 2)

Considering the small differences in transfer benefit effects between groups which did not reach the threshold for statistical significance, we decided to conduct a validation study to see whether they would replicate. Furthermore, given the caveats of comparing the temporal delay conditions to the CG (resultant from the different trial structures), we decided to increase our statistical power and avoid those confounds by focusing on the TS-WM comparison only.

5.5.1 Signatures of hierarchical learning - RT cost analysis

We replicated the results of the RT cost analysis from Study 1.1. Median RT cost values were significantly larger than zero for both subject groups (Wilcoxon signed-rank test; TS: $W(79) = 2630$, $p < .001$, one-tailed; WM: $W(79)$ signed rank = 2582, $p < .001$, one-tailed). However, we failed to see the predicted group difference (Mann-Whitney U-test WM > TS: U = 3155, $p = 0.494$, see also **Fig. 5.5A**). As in Study 1.1, we also performed the Bayesian Mann-Whitney U-test, which showed moderate evidence for the null hypothesis of no difference between groups (U = 3155, $BF_{01} = 5.53$).

5.5.2 Performance benefits during transfer

As was the case for the previous subject cohort, both TS and WM subjects showed a performance increase with time spent on task, as well as better performance during Transfer over Baseline learning (**see Fig. 5.5B**). These effects were statistically significant (Mixed model ANOVA with RM factors Phase (Baseline, Transfer), Trial Bin (8 levels) and between-subjects factor Group; Phase*:* $F(1) = 142.69$, $p < .001$; Trial Bin: $F(5.89) = 163.14$, $p < .001$, Greenhouse-Geisser corrected, Phase x Trial bin: $F(5.65) = 21.62$, $p < .001$,

Greenhouse-Geisser corrected). However, in contrast to the previous findings, we also found a significant group difference ($F(1) = 7.00$, $p = .009$), as well as significant Phase x Group and Phase x Trial Bin x Group interactions ($F(1) = 5.50$, $p = .02$ and $F(5.65) = 2.99$, $p = .008$, Greenhouse-Geisser corrected, respectively). A simple contrast testing the difference between groups showed that WM subjects were on average worse than their TS counterparts ($t(198) = -2.65$, $p = .009$). Inspection of the binned accuracy data suggested that this difference was particularly pronounced in the Baseline phase (see **Fig. 5.5B**, third panel from the left), with no pronounced differences in the Transfer performance (**Fig. 5.5B**, last panel). As a consequence, the WM subjects showed a larger overall transfer benefit which was spread across all experimental trials. In contrast, the transfer benefit was more short-lived in the TS group. This observation was supported by the results of the simple main effects analysis, looking at the effects of Phase within each Trial Bin, separately for each group (TS: significant Transfer-Learning difference only for trial bins 1-5 with all *F*-values > 8.8 and *p*-values <. 004; WM: significant difference for all trial bins with all F-values > 6.82 and p-values <= .01; see also **Fig. 5.5B**).

5.5.3 Transfer choice probability

As in Study 1, we analysed the group-wise accuracy on the second-seen unique trial as previously described and found them to be significantly above chance for both target- and context-switch trials in the TS group (binomial test; *target-switch* trials: proportion = 0.9, N = 40, $p < .001$, *cue-switch* trials: proportion = 0.66, N = 38, $p = .037$). However, in contrast to the findings from Study 1.1, we found the probability of making the correct choice to be significantly higher than chance only for *target-switch* but not the *context-switch* trials in the WM group (binomial test; *target-switch* trials: proportion = 0.87, N = 31, $p < .001$, *cue-switch* trials: proportion = 0.59, N = 27, $p = .221$). Furthermore, in both groups, subjects

who experienced a switch in the target object first were on average more accurate that those who experienced a context-switch (see **Fig. 5.5C**). To test if this effect was statistically significant, we performed a logistic regression, analogous to the one used in Study 1.1, predicting the subjects' accuracy from the Trial Type (*target*- and *context-switch*) and Group (TS and WM). Overall, the model captured a small but significant proportion of variance in the data (McFadden R^2 = .089, p = .004). Furthermore, the observed Trial Type effect was statistically significant (odds ratio = 0.21, 95% CI = [0.06, 0.73]). We did not observe a significant difference between groups (odds ratio = 0.75, 95%CI = [0.17, 3.27]) or a Group x Trial Type interaction (odds ratio = 1.01, 95%CI = [0.17, 6.05]). Collectiely, these results imply equal use of the dOR strategy in both TS and WM groups.

5.5.4 Association between the hierarchical strategy use during Baseline problem and magnitude of transfer benefit

Finally, we wanted to directly test for a positive association between the extent of hierarchical strategy use during the Baseline problem and the increased speed of learning of the Transfer problem. As before, we correlated the absolute RT cost magnitude with the transfer benefit. However, the results did not show a positive correlation between the two variables in either group (Spearman correlation; TS: r = .15, p = .187, N = 82; WM: r = -.03, p = .893, N = 82; see also **Fig. 5.5D**)

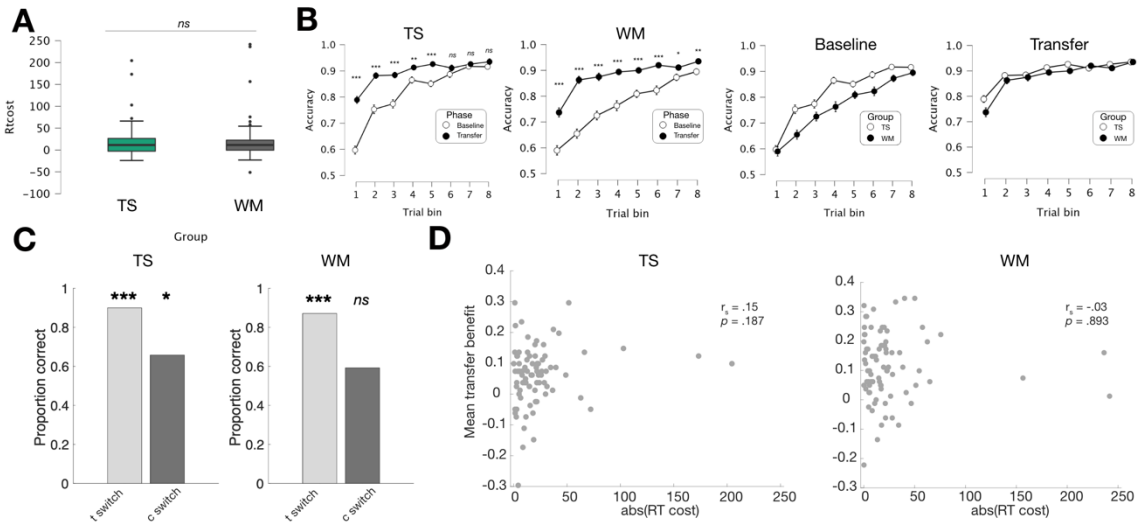


Figure 5.5. Summary of the results of validation Study 2. A. Boxplots showing the distribution of RT cost values in both groups. Significance value corresponds to the results of the one-tailed Mann-Whitney U-test described in the main text. **B.** Aggregate accuracy learning curves from Study 2. First two panels from the left show a comparison between Baseline (white circles) and Transfer (black circles) learning for the TS and WM groups, respectively. The other two panels show a comparison between the TS (white circles) and WM (black circles) groups for each learning phase, respectively. Accuracy data plotted each trial bin (constituting an average of 10 consecutive trials). Error bars correspond to SEM. Significance values correspond to the results of the simple main effects analysis described in the main text. **C.** Observed group average accuracy values for the target-switch (light grey bars) and context-switch (dark grey bars) trials. The sample sizes were (from left to right): 40, 38, 31, and 27. Significance values correspond to a binomial test for an effect significantly greater than chance (50%). **D.** Scatterplots showing the relationship between absolute magnitude of RT cost (x-axis) and the mean transfer benefit (y-axis) for subjects in each group (left to right: TS and WM). Results of a Spearman correlation between the two variables indicated on each panel.

5.6 Discussion

We conducted two online experiments to investigate whether humans naturally extract an abstract problem structure when learning a small set of simple stimulus-response mappings and transfer it to a new learning problem. We further asked whether such tendencies could be strengthened by promoting context-dependent processing in WM, over and above the potential effects of segmenting the stimulus feature presentation in time. Subjects were trained on a simple XOR rule with one set of stimuli and probed on their ability to transfer it to a completely new learning set. We employed a between-subject design, comparing the TS and WM manipulations to joint stimulus feature presentation (CG).

To assess whether subjects induced abstract rule hierarchies when learning the Baseline problem, we searched for the known RT signatures of rule-based processing whilst participants performed the task without feedback. Subjects in all groups showed a consistent asymmetry in the RTs between trials where only the context cue changed, relative to trials where only the target cue changed. Importantly, this asymmetry cannot be explained in terms of simple perceptual or response change effects (as they were matched between the two conditions), leading us to conclude that the subjects had indeed learnt to solve the Baseline problem in a hierarchical fashion. Similar findings have been previously reported by Collins and Frank (2013). In their study, subjects were also trained on four stimulus-response mappings from feedback, with the stimulus features presented concurrently. The authors found that participants spontaneously treated one stimulus feature as the higher-order contextual cue, as evidenced in part by the existence of RT costs. Interestingly, other studies have not always found the same rules (Dreisbach et al., 2006, 2007; Dreisbach and Haider,

2008). For example, in a series of experiments, Dreisbach and colleagues demonstrated that subjects instructed to learn in a flat' way (i.e., utilise a memory-based strategy) did not exhibit RT costs, in contrast to those informed about the existence of abstract task rules (Dreisbach et al., 2006, 2007; Dreisbach and Haider, 2008). Importantly, there are a number of differences between the paradigm used in these studies and the ones explored here and in Collins and Frank (2013). Firstly, the two task rules used in the studies by Dreisbach and colleagues were associated with non-overlapping stimulus sets. Furthermore, these task rules were categorisation rules pertaining to the semantic properties of the task stimuli (e.g., judging whether the written word represented an animal or not). In contrast, in our study, all task stimuli were drawn from the same semantic category. Lastly, and most importantly, new stimulus-response mappings were introduced sequentially in a block-wise manner, to purposefully prevent participants on picking up on the task rules.

With regards to our experimental manipulations, our results suggested that the nature of the constructed hierarchy (i.e., the assignment of stimulus features to hierarchy levels) was aligned with the order of feature presentation in the TS and WM groups. However, contrary to our prediction, we found no evidence for an increased use of the hierarchical strategies in those groups relative to the CG. More specifically, the consistency of the switch cost direction was not significantly different between all groups. Neither did we observe any significant increase in the RT switch cost magnitude between the TS and WM conditions. One possible explanation for this lack of differences is that the engagement of state-dependent processing does not promote the use and transfer of hierarchical, rule-based strategies. However, it is possible that this type of processing was engaged in both TS and WM subject groups. More specifically, whilst controlling for both simple segmentation of task events in time and trial length (and thus processing time), the TS manipulation did not

prevent subjects from engaging their working memory during the filled delay. More specifically, participants in the TS group could have, in principle, encoded the first stimulus object into their working memory, and used it for context-dependent processing of the second object. Therefore, possible future designs could employ an 'iconic' delay control condition instead, whereby the subjects would experience a very short (100 ms) blank delay in-between stimulus object presentations, precluding the engagement of working memory. The lack of observed differences relative to CG might have a similar explanation. As our study used two discrete stimulus objects arranged on top of one another and the task was self-paced, it is possible that the subjects in the CG processed and encoded the two objects sequentially. Such strategy could result in context-dependent processing at the neural level. This interpretation is consistent with our finding of consistent RT switch costs in this group. Thus, future designs could use bivalent stimuli whose constituent features are less easily separable (e.g., coloured shapes), in combination with a response time pressure, to provide maximal contrast with the WM manipulation.

Next, we analysed subjects' performance on the second learning problem. We found that all subjects acquired the second problem much faster than the first one, as evidenced by significantly higher average accuracy and faster time to reach asymptotic performance for the second learning set. Furthermore, there was a subtle but not statistically significant interaction between the experimental phase (baseline or transfer), time on task and group. Further investigation revealed that the positive transfer effect was mostly confined to the first half of the learning trials in the TS and WM conditions, whereas the same effect was more long-lasting in the control condition. Results of the validation study confirmed the existence of transfer benefits in both TS and WM groups, but also highlighted an undesirable difference between the two conditions, namely an increased overall task difficulty in the

case of the latter. This resulted in lower baseline performance in the WM group, but a comparable level of transfer performance. Therefore, from the reduced baseline, there was a relatively greater improvement during transfer. However, the locus of this difference is difficult to disentangle. A favourable interpretation would posit that, in line with our predictions, the WM manipulation encouraged rule abstraction and generalisation over and above the effects of TS. However, the reduced baseline performance also permits another explanation (more consistent with the lack of differences in RT switch costs between the two groups): WM subjects might have engaged a meta-cognitive mechanism to adapt to the increased task demands during baseline learning. This could for example be learning to actively pay more attention to the first stimulus object before it disappeared. Consequentially, the initial performance disadvantage would no longer be present during the transfer phase, leading to matched performance between groups. Unfortunately, dissociating between these two possibilities was not possible in our experimental design.

It is important to note that the observed transfer benefit could in principle reflect non-specific transfer effects (such as a relative improvement in motor skills), rather than the reuse of the abstract XOR problem structure. However, we found that the overall magnitude of performance improvement between the two problems was positively correlated with the magnitude of the RT switch cost in the CG. This suggests that at least some subjects in the CG group transferred the XOR strategy across tasks. Surprisingly, we found no such association in the TS and WM groups. Whilst it is possible that the transfer benefit in those groups corresponded to non-specific effects only, the results of the transfer choice patterns suggested that subjects did indeed rely on the XOR strategy. Another possibility is that the the control group could have been characterised by more meaningful subject-wise variance (whereby a lower proportion of subjects relied on a hierarchical strategy). However, the lack

of significant differences in RT cost consistency between groups argues against this possibility. Alternatively, the CG might have been more sensitive to switch costs, given that the context and target cues were presented jointly. Accordingly, switch costs are known to decrease in magnitude depending on the length of the cue-target interval (Monsell, 2003; Kiesel et al., 2010; Vandierendonck et al., 2010).

Further evidence for the transfer of the XOR rule across tasks comes from the analysis of subjects' choice patterns on the second unique trial of the transfer phase. We generated predictions for the probability of being correct on that trial contingent on both the trial type and the behavioural strategy (or abstract rule structure) the subjects had acquired during initial learning. The results of this analysis suggest that all three experimental groups included subjects using the XOR strategy to transfer between tasks, but their proportions did not significantly differ between groups. Interestingly, the same analysis in the replication study revealed a potential use of a related dOR strategy in the WM group. This was based on the findings of chance-level performance on context-switch trials in those subjects. However, this result could also simply reflect the increased difficulty of making this type of inference compared to the simpler target-switch condition. In line with this interpretation, a similar reduction in average accuracy for this trial type was also observed in the TS group. Accordingly, the results of the logistic regression did not reveal a significant interaction between the group and trial type.

Overall, our results suggest that human subjects might have a natural tendency to express the XOR problem structure in hierarchical terms. Thus, our results can be seen as a replication of the findings of Collins and Frank (2013) in a task with a different associative structure (XOR). Furthermore, we observed that, in the control group, the extent of

hierarchical processing during the first learning problem (as indexed by the RT switch cost) was positively associated with the improvement in learning of a second task, characterised by a separate stimulus set but the same underlying structure. Thus, it appears that, subjects using a hierarchical strategy are not only capable of rule generalisation based on the repeating stimulus-response mappings (as explored by Collins and Frank, (2013), but might also transfer the abstract relational (XOR) structure across tasks with no shared perceptual features. However, in order to be able to draw any stronger conclusions about people's natural tendency to abstract and generalise, the experiments should be replicated whilst addressing the methodological shortcomings outlined above.

Chapter 6. General Discussion

In this thesis, I examined different aspects of flexible, rule-based behaviour using a combination of human behavioural experiments and neural network modelling. In Chapters 2 and 3, I studied how information in short-term memory is manipulated to match changing task demands. In Chapter 4 and 5, I investigated abstract rule learning from an algorithmic and mechanistic perspective, respectively. In the following sections, I first summarise the results and contributions of each chapter, before discussing the limitations of the work described in this thesis and possible future research directions.

5.1 Chapter summaries and contributions

In **Chapter 2**, I trained recurrent neural networks (RNNs) on a retrocueing task which required manipulation of information maintained in short-term memory. I showed that networks trained to minimise memory recall error with gradient descent develop a geometry of internal representations that is a striking match to that previously observed in macaque lateral prefrontal cortex (LPFC). Because this geometry was developed independently by the RNNs (through training and only for a task that was a direct analogue of that performed by the animals) and by the brain (through evolution and/or learning), the observed convergence implies that it is necessary for performing the retrocueing task and addressing the behavioural demands it models. I then built on this result by characterising how and why the described geometry emerged in the RNNs, in addition to revealing its mechanistic basis, thus providing testable predictions for future neurobiological studies.

The results of this chapter contribute to our understanding of how neural circuits such as the LPFC support flexible, context-dependent behaviour in several ways. Firstly, I analysed

aspects of the RNN model that are not readily accessible to biological researchers, such as network connectivity. Therefore, I extended the results from Panichello and Buschman (2021) by providing a mechanistic basis for the neural representations and computations the authors described in the macaque LPFC. This prediction could be tested in biological brains when the appropriate technology to measure neural connectivity is developed.

The described results also add to our general understanding of the representational formats used by information processing systems (including both biological and artificial networks) to achieve specific computational goals. Orthogonal task representations are seen across a range of conditions that share the requirement for information from different sources to be kept separate (Flesch et al., 2018, 2021, 2022; Libby and Buschman, 2021; Nieh et al., 2021; Russin et al., 2022; Xie et al., 2022). Similarly, parallel low-dimensional manifolds have been observed in a variety of scenarios that require the agent to generalise their response across different contexts (Remington et al., 2018; Luyckx et al., 2019; Bernardi et al., 2020; Sheahan et al., 2021). Our results also suggest that in the context of working memory tasks, a parallel geometry might additionally confer a robustness to maintenance. Interestingly, as posited by recent normative theories of cognitive control (Musslick and Cohen, 2021), we also observed that it is accompanied by a learning speed advantage, whereby the models that relied on a parallel post-cue geometry required the least training iterations, compared to models that used separate, task-specific representations. From this perspective, the findings from Chapter 2 extend our understanding of the general coding principles used by information processing systems, as well as the normative reasons behind them.

In **Chapter 3**, I described the results of a simulation study that extended the retrocueing paradigm to a case where the cue was probabilistic. In this setting, the cued item information

in working memory was prioritised, but could nonetheless turn out to be irrelevant for guiding behaviour at a later probe stage (i.e., on trials where the cue was invalid). Our RNN model captured the behavioural retrocueing benefit characteristic of human participants performing this task - namely, the significant improvement in memory performance on validly cued trials, relative to invalidly cued trials (Souza and Oberauer, 2016). Subsequently, we searched for the neural basis of this effect in the RNN. Based on previous theoretical considerations (Myers et al., 2017b), as well as our results from Chapter 2 and the accompanying non-human primate data from Panichello and Buschman (2021), we expected to see a similar orthogonal-to-parallel transformation of the cued memory item following the probabilistic cue. However, we found this not to be the case. Instead, cued item information was maintained in an orthogonal format, and was only transformed into the parallel geometry at the end of the trial, after the deterministic probe. Our results further suggested that the retrocueing benefit was underlined by a gain modulation mechanism. More specifically, we observed the representational discriminability to be significantly higher for the cued item compared to the uncued item, a difference that was even more pronounced in the post-probe delay. Crucially, this difference was significantly attenuated on invalidly cued trials, and the magnitude of such attenuation correlated with the difference in memory precision between the two trial types. This result fits well with the attentional strengthening accounts of the retrocueing effect. Such accounts argue that selective attention increases the proportion of the limited working memory resources allocated to the maintenance of the cued item, at the cost of the uncued items (Chun, 2011; Chun et al., 2011; Kiyonaga and Egner, 2013; Ma et al., 2014; Bays, 2015). The lack of the orthogonal-to-parallel transformation after a probabilistic cue suggests that this task solution is only optimal for fully deterministic conditions. This prediction could be tested experimentally in future neurobiological studies.

In **Chapter 4**, I examined how logical rules can be learnt from an algorithmic perspective. It was previously observed that human performance on Boolean categorisation problems is well predicted by the extent to which a category can be reduced to a simpler, abstract rule (as measured by its Boolean complexity). Our results challenged the previously drawn conclusion that such relationship constituted strong evidence for the symbolic character of human thought, as posited by the language of thought (LoT) accounts of cognition (Feldman, 2000). In contrast, we demonstrated that a similar behavioural sensitivity to the Boolean complexity could be observed in artificial neural networks (ANNs) trained on the same set of categorisation problems. Importantly, ANN models rely on a markedly different set of assumptions (the representational format and learning algorithm) than that inherent to the LoT account. This suggests that the sensitivity of categorisation performance to Boolean complexity does not index rule-based processes involving symbolic representations in humans and should not be used for such purpose. Interestingly, in the domain of semantic category learning, learning dynamics of gradient descent have previously been shown to follow the correlational structure of the learning problem, whereby the features that are shared by the largest amount of learning exemplars are learnt the fastest (Saxe et al., 2014). Our results suggest that a similar mechanism might also underlie Boolean category learning, inviting future investigations into this topic.

Finally, in **Chapter 5**, I once again tackled the problem of abstract rule learning, this time from a mechanistic perspective. I described behavioural experiments that aimed to answer two experimental questions. Firstly, we wished to assess the extent to which human subjects showed a natural predisposition to learn abstract rules from experience and transfer them to a new context. Secondly, we wished to test whether such tendencies could be further promoted through the engagement of state-dependent processing in working memory. With regards to the first question, our results showed that subjects in the control group exhibited

reaction time (RT) markers of rule-dependent processing, suggesting that they had decomposed the XOR problem into a two-step decision process. Furthermore, the extent to which individual subjects showed this tendency was positively correlated with their learning speed improvement on the second learning problem (which also had an XOR structure). With regards to our second prediction, we found no evidence that our experimental manipulation encouraged the use of hierarchical strategies above the natural tendency to do so. Surprisingly, the magnitude of the RT markers of rule-dependent processing did not correlate with the transfer benefit on the second XOR problem in the WM or TS group. The results of this chapter make a small contribution towards our understanding of abstract rule learning by showing that human subjects can learn the abstract XOR structure through trial and error. The RT switch cost analysis further suggests that subjects naturally treat the XOR problem in a hierarchical way, and this tendency correlates with their ability to transfer it to a new learning problem. Similar observations have been previously made in other settings (Collins and Frank, 2013). The limitations of the study and the strength of evidence in support of this conclusion are discussed in the next section.

5.2 Limitations, outstanding questions and future directions

As argued above, the results from **Chapter 2** make several contributions to our understanding of the neural mechanisms of flexible, rule-based behaviour. However, as is the case for any study, the described experiments also have their limitations.

First of such limitations pertains to the finding that only models trained with a longer post-cue delay interval develop a parallel *Cued* item geometry. In contrast, models not required to maintain the cued item information in memory for long use an orthogonal plane solution. One interpretation of this finding is that a parallel cue solution might render the memories

more robust for maintenance. However, it is also possible that there are temporal constraints on the representational transitions from orthogonal to parallel formats in the RNN. This possibility could be explored in future theoretical studies. It would also be interesting to see if the other task geometries (*Uncued*, *Cued/Uncued*) are affected by the length of the delay period.

Another outstanding question is whether the extent of the attentional strengthening of the cued item (reflected in its higher colour discriminability, relative to the uncued item) shows any relationship to network behaviour. In Chapter 3, we saw that it was positively correlated with the memory recall precision for a related, probabilistic task. Thus, it would be interesting to investigate whether a similar effect was present in the paradigm explored in Chapter 2.

In **Chapter 3**, we were interested in the neural mechanisms of prioritisation of items in working memory. Our results suggested that the retrocueing benefit observed on validly cued trials related to a gain modulation mechanism. However, this conclusion was based on the results of a correlational analysis. Therefore, to draw stronger conclusions about the causal role of this mechanism, we could extend our findings by using a variant of the recurrent neural network where the gain of subpopulation of neurons is controlled by separate hyperparameters. After training such networks on the task, we could then causally manipulate the gain modulation of the cued and uncued items prior to the response (for example, by swapping the parameters between the two subpopulations of neurons encoding the items) and test whether this affects the differences in behaviour on valid and invalid trials.

One limitation of the results described in **Chapter 4** pertains to the format of the behavioural dataset supplied. We wished to assess how well the variability in human categorisation

behaviour could be accounted for by the predictions grounded in the connectionist framework of cognition. To do so, we regressed the learning speed predictions generated from our neural networks against the behavioural data and compared the proportion of variance explained by this model to the regression model including the Boolean complexity predictor. We did not find any evidence that human behaviour was better captured by the network-generated learning speed predictions. However, it is possible that this analysis lacked the required statistical power, as we only had access to the group mean and standard error values for the experimental conditions. This point could be addressed by collecting a new human dataset on the same task and performing the same analysis at the individual subject level.

The results of **Chapter 5** are subject to a number of caveats. We observed a consistent direction of RT switch costs in the control group. The RT cost magnitude was also significantly correlated with the magnitude of performance improvement on the second learning problem (i.e., the transfer benefit). Collectively, these findings suggest that participants might have extracted abstract problem structure during the first learning task and transferred it to the second problem. However, to strengthen the evidence in support of this conclusion, we could have performed a similar correlational analysis between the switch consistency metric and transfer benefit. In terms of improvements to the experimental design, we could have included another evaluation phase, following the second learning problem. The existence of similar RT switch costs in this phase of the experiment would corroborate our interpretation that the transfer benefits were underlined by the use of a hierarchical strategy. Similarly, including a negative transfer condition would have helped us further assess the specificity of the transfer benefit.

With regards to the innateness of the hierarchical strategy use, it is also possible that it might have reflected the nature of the stimuli used in our experiment. More specifically, it is possible that using two discrete stimulus objects presented at the top and bottom might have encouraged people to treat them as different levels of a hierarchy. Thus, to strengthen the evidence in support of our conclusion that subjects naturally build rule hierarchies, we should strive to replicate this finding in a separate study, using less separable stimulus dimensions (such as coloured shapes).

Lastly, regarding the involvement of context-dependent processing in rule abstraction, it is possible that it could account for the behaviour of subjects in the control group. Because our task was self-paced, subjects were free to process and encode the stimulus features sequentially, and thus could have relied on context-dependent processing in working memory to solve the task. To investigate this possibility (as well as increase the contrast between the control group and the working memory manipulation), we could use a timed task. Introducing an RT pressure could prevent subjects from pursuing such a strategy, encouraging them to instead encode the conjunctions of stimulus features.

Most generally, one limitation of the results described in the chapters of this thesis pertains to the use of artificial neural networks. Such models are composed of simplified units that do not capture all the features of biological neurons and rely on learning rules that are thought to be biologically implausible (such as backpropagation). However, several recent perspectives argue that neural networks might operate at an appropriate level of abstraction to model information processing in biological brains (Richards et al., 2019; Doerig et al., 2022). Additionally, biological detail can be incorporated as needed, in a hypothesis-driven way. Lastly, a large body of literature shows that the backpropagation algorithm can be approximated in a biologically plausible way, for example by incorporating simple

assumptions about cellular structures, short-term synaptic plasticity, or long-range connections into the networks, to name a few (O'Reilly, 1996; Körding and König, 2001; Roelfsema and van Ooyen, 2005; Lillicrap et al., 2016; Bengio et al., 2017; Guerguiev et al., 2017; Scellier and Bengio, 2017; Whittington and Bogacz, 2017, 2019; Sacramento et al., 2018; Pozzi et al., 2019).

5.4 Conclusion

In summary, in this thesis, I investigated different aspects of flexible, intelligent behaviour. I demonstrated that neural networks capture features of neural activity and behaviour of biological agents. I then leveraged the strengths of the modelling approach to investigate features not easily measured in biology, and generated predictions for future studies. Taken together, it is hoped that the work presented in this thesis furthers our understanding of the principles behind flexible, intelligent behaviour and sheds new light on their mechanistic implementation in neural circuits.

References

Aron AR, Monsell S, Sahakian BJ, Robbins TW (2004) A componential analysis of task-switching deficits associated with lesions of left and right frontal cortex. Brain 127:1561–1573.

Asaad WF, Rainer G, Miller EK (1998) Neural activity in the primate prefrontal cortex during associative learning. Neuron 21:1399–1407.

Asaad WF, Rainer G, Miller EK (2000) Task-specific neural activity in the primate prefrontal cortex. J Neurophysiol 84:451–459.

Ashby FG, Maddox WT (1998) Stimulus categorization. In: Measurement, judgment, and decision making, pp 251–301 Handbook of perception and cognition (2nd ed.). San Diego, CA, US: Academic Press.

Ashby GF, Maddox TW (2005) Human category learning. Annual review of psychology 56 Available at: https://pubmed.ncbi.nlm.nih.gov/15709932/ [Accessed December 18, 2022].

Atkin A (2023) Peirce's Theory of Signs. In: The Stanford Encyclopedia of Philosophy, Spring 2023. (Zalta EN, Nodelman U, eds). Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/archives/spr2023/entries/peirce-semiotics/ [Accessed April 19, 2023].

Baddeley AD, Hitch G (1974) Working Memory. In: Psychology of Learning and Motivation (Bower GH, ed), pp 47–89. Academic Press. Available at: https://www.sciencedirect.com/science/article/pii/S0079742108604521 [Accessed April 13, 2023].

Badre D, Kayser AS, D'Esposito M (2010) Frontal cortex and the discovery of abstract action rules. Neuron 66:315–326.

Baeg EH, Kim YB, Huh K, Mook-Jung I, Kim HT, Jung MW (2003) Dynamics of population code for working memory in the prefrontal cortex. Neuron 40:177–188.

Banino A et al. (2018) Vector-based navigation using grid-like representations in artificial agents. Nature 557:429–433.

Barak O, Tsodyks M (2014) Working models of working memory. Curr Opin Neurobiol 25:20–24.

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

Barsalou LW (1985) Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. Journal of Experimental Psychology: Learning, Memory, and Cognition 11:629–654.

Baumeister RF, Bratslavsky E, Muraven M, Tice DM (1998) Ego depletion: Is the active self a limited resource? Journal of Personality and Social Psychology 74:1252–1265.

Bays PM (2015) Spikes not slots: noise in neural populations limits working memory. Trends in cognitive sciences 19 Available at: https://pubmed.ncbi.nlm.nih.gov/26160026/ [Accessed March 25, 2023].

Bays PM, Catalao RFG, Husain M (2009) The precision of visual working memory is set by allocation of a shared resource. J Vis 9:7.1-11.

Bechtel W, Abrahamsen A (1991) Connectionism and the mind: An introduction to parallel processing in networks. Cambridge, MA, US: Basil Blackwell.

Bengio Y, Mesnard T, Fischer A, Zhang S, Wu Y (2017) STDP-Compatible Approximation of Backpropagation in an Energy-Based Model. Neural Comput 29:555–577.

Bernardi S, Benna MK, Rigotti M, Munuera J, Fusi S, Salzman CD (2020) The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. Cell:S0092867420312289.

Blum RA (1952) Effects of subtotal lesions of frontal granular cortex on delayed reaction in monkeys. AMA Archives of Neurology & Psychiatry 67:375–386.

Botvinick M, Braver T (2015) Motivation and Cognitive Control: From Behavior to Neural Mechanism. Annu Rev Psychol 66:83–113.

Braver TS (2012) The variable nature of cognitive control: a dual mechanisms framework. Trends Cogn Sci 16:106–113.

Braver TS, Gray JR, Burgess GC (2007) Explaining the many varieties of working memory variation: Dual mechanisms of cognitive control. In: Variation in working memory, pp 76–106. New York, NY, US: Oxford University Press.

Broadbent DE (1957) A mechanical model for human attention and immediate memory. Psychol Rev 64:205–215.

Brody CD, Hernández A, Zainos A, Romo R (2003) Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. Cereb Cortex 13:1196–1207.

Brooks LR (1978) Nonanalytic Concept Formation and Memory for Instances. In: Cognition and Categorization (Rosch E, Lloyd B, eds), pp 3–170. Lawrence Elbaum Associates.

Brown T et al. (2020) Language Models are Few-Shot Learners. In: Advances in Neural Information Processing Systems, pp 1877–1901. Curran Associates, Inc. Available at: https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html [Accessed December 19, 2022].

Bruner, Jerome S., Goodnow, Jacqueline J., Austin, George A. (1956) A Study of Thinking. Wiley & Sons, Incorporated, John. Available at: https://www.taylorfrancis.com/books/mono/10.4324/9781315083223/study-

thinking-jerome-bruner-jacqueline-goodnow-george-austin [Accessed December 18, 2022].

Buckley MJ, Mansouri FA, Hoda H, Mahboubi M, Browning PGF, Kwok SC, Phillips A, Tanaka K (2009) Dissociable components of rule-guided behavior depend on distinct medial and prefrontal regions. Science 325:52–58.

Bunge SA (2004) How we use rules to select actions: a review of evidence from cognitive neuroscience. Cogn Affect Behav Neurosci 4:564–579.

Bunge SA, Kahn I, Wallis JD, Miller EK, Wagner AD (2003) Neural circuits subserving the retrieval and maintenance of abstract rules. J Neurophysiol 90:3419–3428.

Buonomano DV, Maass W (2009) State-dependent computations: spatiotemporal processing in cortical networks. Nat Rev Neurosci 10:113–125.

Burgess GC, Braver TS (2010) Neural Mechanisms of Interference Control in Working Memory: Effects of Interference Expectancy and Fluid Intelligence. PLoS One 5:e12861.

Buschman TJ, Siegel M, Roy JE, Miller EK (2011) Neural substrates of cognitive capacity limitations. Proceedings of the National Academy of Sciences 108:11252–11255.

Carcassi F, Szymanik J (2022) Neural Networks Track the Logical Complexity of Boolean Concepts. Open Mind (Camb) 6:132–146.

Carter CS, Mintun M, Cohen JD (1995) Interference and Facilitation Effects during Selective Attention: An H215O PET Study of Stroop Task Performance. NeuroImage 2:264–272.

Carter EC, Kofler LM, Forster DE, McCullough ME (2015) A series of meta-analytic tests of the depletion effect: Self-control does not seem to rely on a limited resource. Journal of Experimental Psychology: General 144:796–815.

Cavanagh SE, Towers JP, Wallis JD, Hunt LT, Kennerley SW (2018) Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex. Nat Commun 9:3498.

Chaisangmongkon W, Swaminathan SK, Freedman DJ, Wang X-J (2017) Computing by Robust Transience: How the Fronto-Parietal Network Performs Sequential, Category-Based Decisions. Neuron 93:1504-1517.e4.

Christophel TB, Klink PC, Spitzer B, Roelfsema PR, Haynes J-D (2017) The Distributed Nature of Working Memory. Trends Cogn Sci 21:111–124.

Chun MM (2011) Visual working memory as visual attention sustained internally over time. Neuropsychologia 49 Available at: https://pubmed.ncbi.nlm.nih.gov/21295047/ [Accessed March 25, 2023].

Chun MM, Golomb JD, Turk-Browne NB (2011) A Taxonomy of External and Internal Attention. Annual Review of Psychology 62:73–101.

Cohen JD (2017) Cognitive Control: Core Constructs and Current Considerations. In: The Wiley Handbook of Cognitive Control, pp 1–28. John Wiley & Sons, Ltd. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118920497.ch1 [Accessed April 14, 2023].

Cohen JD, Dunbar K, McClelland JL (1990) On the control of automatic processes: a parallel distributed processing account of the Stroop effect. Psychol Rev 97:332–361.

Cohen JD, Servan-Schreiber D (1992) Context, cortex, and dopamine: A connectionist approach to behavior and biology in schizophrenia. Psychological Review 99:45–77.

Collins AGE, Albrecht MA, Waltz JA, Gold JM, Frank MJ (2017) Interactions Among Working Memory, Reinforcement Learning, and Effort in Value-Based Choice: A New Paradigm and Selective Deficits in Schizophrenia. Biol Psychiatry 82:431–439.

Collins AGE, Frank MJ (2012) How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. Eur J Neurosci 35:1024–1035.

Collins AGE, Frank MJ (2013) Cognitive control over learning: Creating, clustering, and generalizing task-set structure. Psychological Review 120:190–229.

Collins AGE, Frank MJ (2018) Within- and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. Proc Natl Acad Sci USA 115:2502–2507.

Constantinidis C, Franowicz MN, Goldman-Rakic PS (2001) The sensory nature of mnemonic representation in the primate prefrontal cortex. Nat Neurosci 4:311–316.

Constantinidis C, Funahashi S, Lee D, Murray JD, Qi X-L, Wang M, Arnsten AFT (2018) Persistent Spiking Activity Underlies Working Memory. J Neurosci 38:7020–7028.

Courtney SM, Petit L, Maisog JMa, Ungerleider LG, Haxby JV (1998) An Area Specialized for Spatial Working Memory in Human Frontal Cortex. Science 279:1347–1351.

Crescentini C, Seyed-Allaei S, De Pisapia N, Jovicich J, Amati D, Shallice T (2011) Mechanisms of Rule Acquisition and Rule Following in Inductive Reasoning. J Neurosci 31:7763–7774.

Crick F (1989) The recent excitement about neural networks. Nature 337:129–132.

Cromer JA, Roy JE, Miller EK (2010) Representation of Multiple, Independent Categories in the Primate Prefrontal Cortex. Neuron 66:796–807.

Dasgupta I, Lampinen AK, Chan SCY, Creswell A, Kumaran D, McClelland JL, Hill F (2022) Language models show human-like content effects on reasoning. Available at: http://arxiv.org/abs/2207.07051 [Accessed December 7, 2022].

D'Esposito M, Postle BR (1999) The dependence of span and delayed-response performance on prefrontal cortex. Neuropsychologia 37:1303–1315.

di Pellegrino G, Wise SP (1993) Visuospatial versus visuomotor activity in the premotor and prefrontal cortex of a primate. J Neurosci 13:1227–1243.

Dias R, Robbins TW, Roberts AC (1996) Primate analogue of the Wisconsin card sorting test: Effects of excitotoxic lesions of the prefrontal cortex in the marmoset. Behavioral Neuroscience 110:872–886.

Doerig A, Sommers R, Seeliger K, Richards B, Ismael J, Lindsay G, Kording K, Konkle T, Van Gerven MAJ, Kriegeskorte N, Kietzmann TC (2022) The neuroconnectionist research programme. Available at: http://arxiv.org/abs/2209.03718 [Accessed April 21, 2023].

Doll BB, Bath KG, Daw ND, Frank MJ (2016) Variability in Dopamine Genes Dissociates Model-Based and Model-Free Reinforcement Learning. J Neurosci 36:1211–1222.

Dreisbach G, Goschke T, Haider H (2006) Implicit task sets in task switching? Journal of Experimental Psychology: Learning, Memory, and Cognition 32:1221–1233.

Dreisbach G, Goschke T, Haider H (2007) The role of task rules and stimulus–response mappings in the task switching paradigm. Psychological Research 71:383–392.

Dreisbach G, Haider H (2008) That's what task sets are for: shielding against irrelevant information. Psychol Res 72:355–361.

Dunbar K, Sussman D (1995) Toward a cognitive account of frontal lobe function: simulating frontal lobe deficits in normal subjects. Ann N Y Acad Sci 769:289–304.

Duncan J (1986) Disorganisation of behaviour after frontal lobe damage. Cognitive Neuropsychology 3:271–290.

Duncan J, Johnson R, Swales M, Freer C (1997) Frontal lobe deficits after head injury: Unity and diversity of function. Cognitive Neuropsychology 14:713–741.

Duncan J, Miller EK (2002) Cognitive focus through adaptive neural coding in the primate prefrontal cortex. In: Principles of frontal lobe function, pp 278–291. New York, NY, US: Oxford University Press.

Durstewitz D, Seamans JK, Sejnowski TJ (2000) Neurocomputational models of working memory. Nat Neurosci 3:1184–1191.

Elsayed GF, Lara AH, Kaufman MT, Churchland MM, Cunningham JP (2016) Reorganization between preparatory and movement population responses in motor cortex. Nature Communications 7:13239.

Ester EF, Nouri A, Rodriguez L (2018) Retrospective Cues Mitigate Information Loss in Human Cortex during Working Memory Storage. J Neurosci 38:8538–8548.

Ester EF, Sprague TC, Serences JT (2015) Parietal and Frontal Cortex Encode Stimulus-Specific Mnemonic Representations during Visual Working Memory. Neuron 87:893–905.

Estes WK (1986) Array models for category learning. Cognitive Psychology 18:500–549.

Estes WK (1994) Classification and cognition. New York, NY, US: Oxford University Press.

Feldman J (2000) Minimization of Boolean complexity in human concept learning. Nature 407:630–633.

Feldman J (2003) The Simplicity Principle in Human Concept Learning. Current Directions in Psychological Science 12:227–232.

Flesch T, Balaguer J, Dekker R, Nili H, Summerfield C (2018) Comparing continual task learning in minds and machines. Proceedings of the National Academy of Sciences 115:E10313–E10322.

Flesch T, Juechems K, Dumbalska T, Saxe A, Summerfield C (2021) Rich and lazy learning of task representations in brains and neural networks. bioRxiv.

Flesch T, Juechems K, Dumbalska T, Saxe A, Summerfield C (2022) Orthogonal representations for robust context-dependent task performance in brains and neural networks. Neuron 110:1258-1270.e11.

Fodor JA (1975) The Language of Thought. Harvard University Press.

Fodor JA, Pylyshyn ZW (1988) Connectionism and cognitive architecture: A critical analysis. Cognition 28:3–71.

Forrest CLD, Monsell S, McLaren IPL (2014) Is performance in task-cuing experiments mediated by task set selection or associative compound retrieval? J Exp Psychol Learn Mem Cogn 40:1002–1024.

Funahashi S, Bruce CJ, Goldman-Rakic PS (1989) Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. J Neurophysiol 61:331–349.

Fuster JM (2001) The Prefrontal Cortex—An Update. Neuron 30:319–333.

Fuster JM, Alexander GE (1971) Neuron activity related to short-term memory. Science 173:652–654.

Fuster JM, Bodner M, Kroger JK (2000) Cross-modal and cross-temporal association in neurons of frontal cortex. Nature 405:347–351.

Goldman-Rakic PS (1995a) Cellular basis of working memory. Neuron 14:477–485.

Goldman-Rakic PS (1995b) Cellular basis of working memory. Neuron 14:477–485.

Goldman-Rakic PS, Bates JF, Chafee MV (1992) The prefrontal cortex and internally generated motor acts. Current Opinion in Neurobiology 2:830–835.

Goodman ND, Tenenbaum JB, Feldman J, Griffiths TL (2008) A rational analysis of rule-based concept learning. Cogn Sci 32:108–154.

Goodman ND, Ullman TD, Tenenbaum JB (2011) Learning a theory of causality. Psychological Review 118:110–119.

Göthe K, Oberauer K, Kliegl R (2016) Eliminating dual-task costs by minimizing crosstalk between tasks: The role of modality and feature pairings. Cognition 150:92–108.

Grafman J (1989) Plans, Actions, and Mental Sets: Managerial Knowledge Units in the Frontal Lobes. In: Integrating Theory and Practice in Clinical Neuropsychology, 1st ed.

Grafman J (2002) The structured event complex and the human prefrontal cortex. In: Principles of frontal lobe function, pp 292–310. New York, NY, US: Oxford University Press.

Griffin IC, Nobre AC (2003a) Orienting Attention to Locations in Internal Representations. Journal of Cognitive Neuroscience 15:1176–1194.

Griffin IC, Nobre AC (2003b) Orienting attention to locations in internal representations. J Cogn Neurosci 15:1176–1194.

Guerguiev J, Lillicrap TP, Richards BA (2017) Towards deep learning with segregated dendrites. Elife 6:e22901.

Gunseli E, van Moorselaar D, Meeter M, Olivers CNL (2015) The reliability of retro-cues determines the fate of noncued visual working memory representations. Psychon Bull Rev 22:1334–1341.

Hajonides JE, Nobre AC, van Ede F, Stokes MG (2021) Decoding visual colour from scalp electroencephalography measurements. NeuroImage 237:118030.

Haller M, Case J, Crone NE, Chang EF, King-Stephens D, Laxer KD, Weber PB, Parvizi J, Knight RT, Shestyuk AY (2018) Persistent neuronal activity in human prefrontal cortex links perception and action. Nat Hum Behav 2:80–91.

Hebb DO (1949) The organization of behavior; a neuropsychological theory. Oxford, England: Wiley.

Herculano-Houzel S (2012) The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. Proc Natl Acad Sci U S A 109 Suppl 1:10661–10668.

Higo T, Mars RB, Boorman ED, Buch ER, Rushworth MFS (2011) Distributed and causal influence of frontal operculum in task control. Proceedings of the National Academy of Sciences 108:4230–4235.

Hintzman DL (1986) “Schema abstraction” in a multiple-trace memory model. Psychological Review 93:411–428.

Homa D, Sterling S, Trepel L (1981) Limitations of exemplar-based generalization and the abstraction of categorical information. Journal of Experimental Psychology: Human Learning and Memory 7:418–439.

Inagaki HK, Fontolan L, Romani S, Svoboda K (2019) Discrete attractor dynamics underlies persistent activity in the frontal cortex. Nature 566:212–217.

Jacobsen CF (1935) Functions of frontal association area in primates. Archives of Neurology & Psychiatry 33:558–569.

Jimura K, Locke HS, Braver TS (2010) Prefrontal cortex mediation of cognitive enhancement in rewarding motivational contexts. Proc Natl Acad Sci U S A 107:8871–8876.

Kahneman D (1973) Attention and effort. Englewood Cliffs: Prentice-Hall.

Keil FC (1979) Semantic and Conceptual Development: An Ontological Perspective. Cambridge, MA: Harvard University Press.

Kerns JG, Cohen JD, MacDonald AW, Cho RY, Stenger VA, Carter CS (2004) Anterior Cingulate Conflict Monitoring and Adjustments in Control. Science 303:1023–1026.

Kiesel A, Steinhauser M, Wendt M, Falkenstein M, Jost K, Philipp AM, Koch I (2010) Control and interference in task switching--a review. Psychol Bull 136:849–874.

Kiyonaga A, Egner T (2013) Working memory as internal attention: toward an integrative account of internal and external selection processes. Psychonomic bulletin & review 20 Available at: https://pubmed.ncbi.nlm.nih.gov/23233157/ [Accessed March 25, 2023].

Koch I (2009) The role of crosstalk in dual-task performance: evidence from manipulating response-code overlap. Psychological research 73 Available at: https://pubmed.ncbi.nlm.nih.gov/18443819/ [Accessed April 14, 2023].

Kool W, Shenhav A, Botvinick MM (2017) Cognitive Control as Cost-Benefit Decision Making. In: The Wiley Handbook of Cognitive Control, pp 167–189. John Wiley & Sons, Ltd. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118920497.ch10 [Accessed April 16, 2023].

Körding KP, König P (2001) Supervised and unsupervised learning with two sites of synaptic integration. J Comput Neurosci 11:207–215.

Kruschke JK (1992) ALCOVE: An exemplar-based connectionist model of category learning. Psychological Review 99:22–44.

Kubota K, Niki H (1971) Prefrontal cortical unit activity and delayed alternation performance in monkeys. J Neurophysiol 34:337–347.

Kumar S, Joseph S, Gander PE, Barascud N, Halpern AR, Griffiths TD (2016) A Brain System for Auditory Working Memory. J Neurosci 36:4492–4505.

Kuo B-C, Stokes MG, Murray AM, Nobre AC (2014) Attention biases visual activity in visual short-term memory. J Cogn Neurosci 26:1377–1389.

Kurzban R (2010) Does the Brain Consume Additional Glucose during Self-Control Tasks? Evolutionary Psychology 8:244–259.

Lamberts K (2000) Information-accumulation theory of speeded categorization. Psychological review 107 Available at: https://pubmed.ncbi.nlm.nih.gov/10789196/ [Accessed December 18, 2022].

Lample G, Charton F (2019) Deep Learning for Symbolic Mathematics. Available at: http://arxiv.org/abs/1912.01412 [Accessed December 19, 2022].

LaRocque JJ, Lewis-Peacock JA, Drysdale AT, Oberauer K, Postle BR (2013) Decoding attended information in short-term memory: an EEG study. J Cogn Neurosci 25:127–142.

Lee D, Seo H (2007) Mechanisms of reinforcement learning and decision making in the primate dorsolateral prefrontal cortex. Ann N Y Acad Sci 1104:108–122.

Lepsien J, Nobre AC (2007) Attentional modulation of object representations in working memory. Cereb Cortex 17:2072–2083.

Lepsien J, Thornton I, Nobre AC (2011) Modulation of working-memory maintenance by directed attention. Neuropsychologia 49:1569–1577.

Leshno M, Lin VYa, Pinkus A, Schocken S (1993) Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural Networks 6:861–867.

Levy J, Pashler H (2001) Is dual-task slowing instruction dependent? J Exp Psychol Hum Percept Perform 27:862–869.

Lewis-Peacock JA, Drysdale AT, Postle BR (2015) Neural Evidence for the Flexible Control of Mental Representations. Cereb Cortex 25:3303–3313.

Lewis-Peacock JA, Postle BR (2012) Decoding the internal focus of attention. Neuropsychologia 50:470–478.

Libby A, Buschman TJ (2021) Rotational dynamics reduce interference between sensory and memory representations. Nat Neurosci 24:715–726.

Lillicrap TP, Cownden D, Tweed DB, Akerman CJ (2016) Random synaptic feedback weights support error backpropagation for deep learning. Nat Commun 7:13276.

Luria AR (1980) Higher Cortical Functions in Man. Boston, MA: Springer US. Available at: http://link.springer.com/10.1007/978-1-4615-8579-4 [Accessed April 17, 2023].

Luxburg U von, Schölkopf B (2011) Statistical Learning Theory: Models, Concepts, and Results. In: Handbook of the History of Logic (Gabbay DM, Hartmann S, Woods J, eds), pp 651–706 Inductive Logic. North-Holland. Available at: https://www.sciencedirect.com/science/article/pii/B9780444529367500161 [Accessed April 19, 2023].

Luyckx F, Nili H, Spitzer B, Summerfield C (2019) Neural structure mapping in human probabilistic reward learning. Elife 8 Available at: https://www.ncbi.nlm.nih.gov/pubmed/30843789.

Ma WJ, Husain M, Bays PM (2014) Changing concepts of working memory. Nature neuroscience 17:347.

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

MacLeod CM, Dunbar K (1988) Training and Stroop-like interference: Evidence for a continuum of automaticity. Journal of Experimental Psychology: Learning, Memory, and Cognition 14:126–135.

Makovski T, Pertzov Y (2015) Attention and memory protection: Interactions between retrospective attention cueing and interference. The Quarterly Journal of Experimental Psychology 68:1735–1743.

Mandler JM, McDonough L (1993) Concept formation in infancy. Cognitive Development 8:291–318.

Mansouri FA, Freedman DJ, Buckley MJ (2020) Emergence of abstract rules in the primate brain. Nat Rev Neurosci 21:595–610.

Mante V, Sussillo D, Shenoy KV, Newsome WT (2013) Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature 503:78.

Masse NY, Yang GR, Song HF, Wang XJ, Freedman DJ (2019) Circuit mechanisms for the maintenance and manipulation of information in working memory. Nat Neurosci 22:1159–1167.

McClelland JL (2010) Emergence in Cognitive Science. Topics in Cognitive Science 2:751–770.

McClelland JL, Botvinick MM, Noelle DC, Plaut DC, Rogers TT, Seidenberg MS, Smith LB (2010) Letting Structure Emerge: Connectionist and Dynamical Systems Approaches to Cognition. Trends Cogn Sci 14:348–356.

McClelland JL, Rogers TT (2003) The parallel distributed processing approach to semantic cognition. Nature Reviews Neuroscience 4:310.

McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics 5:115–133.

Medin DL, Schaffer MM (1978) Context theory of classification learning. Psychological Review 85:207–238.

Meiran N (1996) Reconfiguration of processing mode prior to task performance. Journal of Experimental Psychology: Learning, Memory, and Cognition 22:1423–1442.

Meyer DE, Kieras DE (1997a) A computational theory of executive cognitive processes and multiple-task performance: Part 2. Accounts of psychological refractory-period phenomena. Psychological Review 104:749–791.

Meyer DE, Kieras DE (1997b) A computational theory of executive cognitive processes and multiple-task performance: Part 1. Basic mechanisms. Psychol Rev 104:3–65.

Meyer T, Qi X-L, Constantinidis C (2007) Persistent discharges in the prefrontal cortex of monkeys naive to working memory tasks. Cereb Cortex 17 Suppl 1:i70-76.

Meyers EM, Freedman DJ, Kreiman G, Miller EK, Poggio T (2008) Dynamic Population Coding of Category Information in Inferior Temporal and Prefrontal Cortex. Journal of Neurophysiology 100:1407–1419.

Miller EK, Cohen JD (2001) An Integrative Theory of Prefrontal Cortex Function. Annu Rev Neurosci 24:167–202.

Milner B (1963) Effects of Different Brain Lesions on Card Sorting: The Role of the Frontal Lobes. Archives of Neurology 9:90–100.

Mingard C, Skalse J, Valle-Pérez G, Martínez-Rubio D, Mikulik V, Louis AA (2020) Neural networks are a priori biased towards Boolean functions with low entropy. Available at: http://arxiv.org/abs/1909.11522 [Accessed December 18, 2022].

Minsky M, Papert S (1969) Perceptrons. Oxford, England: M.I.T. Press.

Mishkin M (1957) Effects of small frontal lesions on delayed alternation in monkeys. Journal of Neurophysiology 20:615–622.

Monchi O, Petrides M, Petre V, Worsley K, Dagher A (2001) Wisconsin Card Sorting revisited: distinct neural circuits participating in different stages of the task identified by event-related functional magnetic resonance imaging. J Neurosci 21:7733–7741.

Mongillo G, Barak O, Tsodyks M (2008) Synaptic theory of working memory. Science 319:1543–1546.

Monsell S (2003) Task switching. Trends Cogn Sci (Regul Ed) 7:134–140.

Moore TL, Schettler SP, Killiany RJ, Rosene DL, Moss MB (2009) Effects on executive function following damage to the prefrontal cortex in the rhesus monkey (Macaca mulatta). Behav Neurosci 123:231–241.

Muhle-Karbe PS, Myers NE, Stokes MG (2021) A Hierarchy of Functional States in Working Memory. J Neurosci 41:4461–4475.

Murray AM, Nobre AC, Clark IA, Cravo AM, Stokes MG (2013) Attention restores discrete items to visual short-term memory. Psychol Sci 24:550–556.

Murray JD, Bernacchia A, Roy NA, Constantinidis C, Romo R, Wang X-J (2017) Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. PNAS 114:394–399.

Musslick S, Cohen JD (2021) Rationalizing constraints on the capacity for cognitive control. Trends in Cognitive Sciences 25:757–775.

Musslick S, Saxe A, Hoskin AN, Reichman D, Cohen JD (2020) On the Rational Boundedness of Cognitive Control: Shared Versus Separated Representations. Available at: https://psyarxiv.com/jkhdf/ [Accessed April 14, 2023].

Musslick S, Saxe A, Özcimder K, Dey B, Henselman G, Cohen JD (2017) Multitasking capability versus learning efficiency in neural network architectures. In. Cognitive Science Society. Available at: https://ora.ox.ac.uk/objects/uuid:5d234221-e892-4765-8741-775ae0b74043 [Accessed April 14, 2023].

Myers NE, Chekroud SR, Stokes MG, Nobre AC (2018) Benefits of flexible prioritization in working memory can arise without costs. J Exp Psychol Hum Percept Perform 44:398–411.

Myers NE, Stokes MG, Nobre AC (2017a) Prioritizing Information during Working Memory: Beyond Sustained Internal Attention. Trends in Cognitive Sciences 21:449–461.

Myers NE, Stokes MG, Nobre AC (2017b) Prioritizing Information during Working Memory: Beyond Sustained Internal Attention. Trends Cogn Sci (Regul Ed) 21:449–461.

Nagahama Y, Okada T, Katsumi Y, Hayashi T, Yamauchi H, Oyanagi C, Konishi J, Fukuyama H, Shibasaki H (2001) Dissociable mechanisms of attentional control within the human prefrontal cortex. Cereb Cortex 11:85–92.

Nakahara K, Hayashi T, Konishi S, Miyashita Y (2002) Functional MRI of macaque monkeys performing a cognitive set-shifting task. Science 295:1532–1536.

Nee DE, Jonides J (2009) Common and Distinct Neural Correlates of Perceptual and Memorial Selection. Neuroimage 45:963–975.

Nieder A (2005) Counting on neurons: the neurobiology of numerical competence. Nat Rev Neurosci 6:177–190.

Nieder A, Dehaene S (2009) Representation of number in the brain. Annu Rev Neurosci 32:185–208.

Nieh EH, Schottdorf M, Freeman NW, Low RJ, Lewallen S, Koay SA, Pinto L, Gauthier JL, Brody CD, Tank DW (2021) Geometry of abstract learned knowledge in the hippocampus. Nature:1–5.

Nobre AC, Coull JT, Maquet P, Frith CD, Vandenberghe R, Mesulam MM (2004) Orienting attention to locations in perceptual versus mental representations. J Cogn Neurosci 16:363–373.

Norman DA, Bobrow DG (1975) On data-limited and resource-limited processes. Cognitive Psychology 7:44–64.

Nosofsky RM (1986) Attention, similarity, and the identification–categorization relationship. Journal of Experimental Psychology: General 115:39–57.

O'Reilly JX, Jbabdi S, Behrens TEJ (2012) How can a Bayesian approach inform neuroscience? Eur J Neurosci 35:1169–1179.

O'Reilly RC (1996) Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. Neural Computation 8:895–938.

Orhan AE, Ma WJ (2017) Efficient probabilistic inference in generic neural networks trained with non-probabilistic feedback. Nat Commun 8:138.

Orhan AE, Ma WJ (2019) A diverse range of factors affect the nature of neural representations underlying short-term memory. Nat Neurosci 22:275–283.

Panichello MF, Buschman TJ (2021) Shared mechanisms underlie the control of working memory and attention. Nature 592:601–605.

Panzeri S, Macke JH, Gross J, Kayser C (2015) Neural population coding: combining insights from microscopic and mass signals. Trends Cogn Sci 19:162–172.

Pashler H (1994) Dual-task interference in simple tasks: data and theory. Psychol Bull 116:220–244.

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

Perret E (1974) The left frontal lobe of man and the suppression of habitual responses in verbal categorical behaviour. Neuropsychologia 12:323–330.

Pessoa L, Gutierrez E, Bandettini PA, Ungerleider LG (2002) Neural Correlates of Visual Working Memory: fMRI Amplitude Predicts Task Performance. Neuron 35:975–987.

Piantadosi ST, Tenenbaum JB, Goodman ND (2016) The logical primitives of thought: Empirical foundations for compositional cognitive models. Psychological Review 123:392–424.

Pinker S, Prince A (1988) On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. Cognition 28:73–193.

Posner MI, Keele SW (1968) On the genesis of abstract ideas. Journal of Experimental Psychology 77:353–363.

Posner MI, Keele SW (1970) Retention of abstract ideas. Journal of Experimental Psychology 83:304–308.

Posner MI, Snyder CRR (1975) Attention and Cognitive Control. In: Information Processing and Cognition: The Loyola Symposium (Solso RL, ed). Lawrence Erlbaum.

Pozzi I, Bohté S, Roelfsema P (2019) A Biologically Plausible Learning Rule for Deep Learning in the Brain. Available at: http://arxiv.org/abs/1811.01768 [Accessed April 21, 2023].

Quian Quiroga R, Panzeri S (2009) Extracting information from neuronal populations: information theory and decoding approaches. Nat Rev Neurosci 10:173–185.

Rainer G, Rao SC, Miller EK (1999) Prospective Coding for Objects in Primate Prefrontal Cortex. J Neurosci 19:5493–5505.

Rajalingham R, Issa EB, Bashivan P, Kar K, Schmidt K, DiCarlo JJ (2018) Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. J Neurosci 38:7255–7269.

Rao SC, Rainer G, Miller EK (1997) Integration of What and Where in the Primate Prefrontal Cortex. Science 276:821–824.

Ravi S, Musslick S, Hamin M, Willke TL, Cohen JD (2021) Navigating the Trade-Off between Multi-Task Learning and Learning to Multitask in Deep Neural Networks. Available at: http://arxiv.org/abs/2007.10527 [Accessed April 14, 2023].

Reed SK (1972) Pattern recognition and categorization. Cognitive Psychology 3:382–407.

Remington ED, Narain D, Hosseini EA, Jazayeri M (2018) Flexible Sensorimotor Computations through Rapid Reconfiguration of Cortical Dynamics. Neuron 98:1005–1019.

Rescorla M (2019) The Language of Thought Hypothesis. In: The Stanford Encyclopedia of Philosophy, Summer 2019. (Zalta EN, ed). Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/archives/sum2019/entries/language-thought/ [Accessed April 19, 2023].

Richards BA et al. (2019) A deep learning framework for neuroscience. Nat Neurosci 22:1761–1770.

Richter FR, Yeung N (2014) Neuroimaging Studies of Task Switching. In: Task Switching and Cognitive Control (Grange J, Houghton G, eds), pp 0. Oxford University Press. Available at: https://doi.org/10.1093/acprof:osobl/9780199921959.003.0010 [Accessed April 20, 2023].

Rigotti M, Barak O, Warden MR, Wang X-J, Daw ND, Miller EK, Fusi S (2013) The importance of mixed selectivity in complex cognitive tasks. Nature 497:585–590.

Roelfsema PR, van Ooyen A (2005) Attention-gated reinforcement learning of internal representations for classification. Neural Comput 17:2176–2214.

Romo R, Brody CD, Hernández A, Lemus L (1999) Neuronal correlates of parametric working memory in the prefrontal cortex. Nature 399:470–473.

Rosch E (1975) Cognitive reference points. Cognitive Psychology 7:532–547.

Rosch E, Mervis CB (1975) Family resemblances: Studies in the internal structure of categories. Cognitive Psychology 7:573–605.

Rosch EH (1973) Natural categories. Cognitive Psychology 4:328–350.

Rose NS, LaRocque JJ, Riggall AC, Gosseries O, Starrett MJ, Meyering EE, Postle BR (2016) Reactivation of latent working memories with transcranial magnetic stimulation. Science 354:1136–1139.

Rosenblatt F (1958) The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review 65:386–408.

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323:533–536.

Rumelhart DE, McClelland JL (1987) Learning the past tenses of English verbs: Implicit rules or parallel distributed processing? In: Mechanisms of language aquisition, pp 195–248. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Rumelhart DE, McClelland JL, PDP Research Group (1987) Parallel Distributed Processing. MIT Press Available at: https://mitpress.mit.edu/9780262680530/parallel-distributed-processing/ [Accessed January 24, 2023].

Russin J, Zolfaghar M, Park SA, Boorman E, O'Reilly RC (2022) A Neural Network Model of Continual Learning with Cognitive Control. arXiv:220204773 [cs, q-bio] Available at: http://arxiv.org/abs/2202.04773 [Accessed March 7, 2022].

Sacramento J, Costa RP, Bengio Y, Senn W (2018) Dendritic cortical microcircuits approximate the backpropagation algorithm. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp 8735–8746 NIPS'18. Red Hook, NY, USA: Curran Associates Inc.

Sagiv Y, Musslick S, Niv Y, Cohen JD (2020) Efficiency of learning vs. processing: Towards a normative theory of multitasking. Available at: http://arxiv.org/abs/2007.03124 [Accessed April 14, 2023].

Sakai K (2008) Task Set and Prefrontal Cortex. Annu Rev Neurosci 31:219–245.

Sakai K, Passingham RE (2003) Prefrontal interactions reflect future task operations. Nature neuroscience 6 Available at: https://pubmed.ncbi.nlm.nih.gov/12469132/?dopt=Abstract [Accessed April 21, 2023].

Sakai K, Rowe JB, Passingham RE (2002) Active maintenance in prefrontal area 46 creates distractor-resistant memory. Nat Neurosci 5:479–484.

Salvucci DD, Taatgen NA (2008) Threaded cognition: An integrated theory of concurrent multitasking. Psychological Review 115:101–130.

Santoro A, Raposo D, Barrett DG, Malinowski M, Pascanu R, Battaglia P, Lillicrap T (2017) A simple neural network module for relational reasoning. In: Advances in Neural Information Processing Systems. Curran Associates, Inc. Available at: https://proceedings.neurips.cc/paper/2017/hash/e6acf4b0f69f6f6e60e9a815938aa1ff-Abstract.html [Accessed April 21, 2023].

Saxe AM, McClelland JL, Ganguli S (2014) Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv. Available at: http://arxiv.org/abs/1312.6120 [Accessed May 19, 2022].

Saxe AM, McClelland JL, Ganguli S (2019) A mathematical theory of semantic development in deep neural networks. Proceedings of the National Academy of Sciences 116:11537–11546.

Scellier B, Bengio Y (2017) Equilibrium Propagation: Bridging the Gap between Energy-Based Models and Backpropagation. Front Comput Neurosci 11:24.

Schmidt BK, Vogel EK, Woodman GF, Luck SJ (2002) Voluntary and automatic attentional control of visual working memory. Perception & Psychophysics 64:754–763.

Schneider W, Shiffrin RM (1977) Controlled and automatic human information processing: I. Detection, search, and attention. Psychological Review 84:1–66.

Schrittwieser J, Antonoglou I, Hubert T, Simonyan K, Sifre L, Schmitt S, Guez A, Lockhart E, Hassabis D, Graepel T, Lillicrap T, Silver D (2020) Mastering Atari, Go, chess and shogi by planning with a learned model. Nature 588:604–609.

Shaffer LH (1975) Multiple attention in continuous verbal tasks. In: Attention and performance V, pp 157–167 Available at: https://scholar.google.com/scholar_lookup?title=Multiple%20attention%20in%20continuous%20verbal%20tasks&author=L.%20Shaffer&publication_year=1975&pages=157-167 [Accessed April 14, 2023].

Shallice T, Burgess PW (1991) Deficits in strategy application following frontal lobe damage in man. Brain 114 (Pt 2):727–741.

Sheahan H, Luyckx F, Nelli S, Teupe C, Summerfield C (2021) Neural state space alignment for magnitude generalization in humans and recurrent networks. Neuron 109:1214-1226.e8.

Shiffrin RM, Schneider W (1977) Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. Psychological Review 84:127–190.

Smith EE, Medin DL (1981) Categories and Concepts: Cambridge, MA: Harvard University Press.

Smittenaar P, FitzGerald THB, Romei V, Wright ND, Dolan RJ (2013) Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. Neuron 80:914–919.

Souza AS, Oberauer K (2016) In search of the focus of attention in working memory: 13 years of the retro-cue effect. Atten Percept Psychophys 78:1839–1860.

Souza AS, Rerko L, Lin H-Y, Oberauer K (2014) Focused attention improves working memory: implications for flexible-resource and discrete-capacity models. Atten Percept Psychophys 76:2080–2102.

Souza AS, Rerko L, Oberauer K (2016) Getting more from visual working memory: Retro-cues enhance retrieval and protect from visual interference. Journal of Experimental Psychology: Human Perception and Performance 42:890–910.

Spaak E, Watanabe K, Funahashi S, Stokes MG (2017) Stable and Dynamic Coding for Working Memory in Primate Prefrontal Cortex. J Neurosci 37:6503–6516.

Speer NK, Jacoby LL, Braver TS (2003) Strategy-dependent changes in memory: effects on behavior and brain activity. Cogn Affect Behav Neurosci 3:155–167.

Sprague TC, Ester EF, Serences JT (2016) Restoring Latent Visual Working Memory Representations in Human Cortex. Neuron 91:694–707.

Stokes MG (2015) "Activity-silent" working memory in prefrontal cortex: a dynamic coding framework. Trends Cogn Sci (Regul Ed) 19:394–405.

Stokes MG, Kusunoki M, Sigala N, Nili H, Gaffan D, Duncan J (2013) Dynamic Coding for Cognitive Control in Prefrontal Cortex. Neuron 78:364–375.

Stroop JR (1935) Studies of interference in serial verbal reactions. Journal of Experimental Psychology 18:643–662.

Sudevan P, Taylor DA (1987) The cuing and priming of cognitive operations. Journal of Experimental Psychology: Human Perception and Performance 13:89–103.

Tamber-Rosenau BJ, Esterman M, Chiu Y-C, Yantis S (2011) Cortical Mechanisms of Cognitive Control for Shifting Attention in Vision and Working Memory. J Cogn Neurosci 23:2905–2919.

Telford CW (1931) The Refractory Phase of Voluntary and Associative Responses. Journal of Experimental Psychology 14:1.

Tombu M, Jolicoeur P (2003) A central capacity sharing model of dual-task performance. J Exp Psychol Hum Percept Perform 29:3–18.

van Loon AM, Olmos-Solis K, Fahrenfort JJ, Olivers CN (2018) Current and future goals are represented in opposite patterns in object-selective cortex de Lange F, Frank MJ, eds. eLife 7:e38677.

van Moorselaar D, Gunseli E, Theeuwes J, Olivers CNL (2015) The time course of protecting a visual memory representation from perceptual interference. Frontiers in Human Neuroscience 8.

Vandierendonck A, Liefooghe B, Verbruggen F (2010) Task switching: interplay of reconfiguration and interference control. Psychol Bull 136:601–626.

Vendrell P, Junqué C, Pujol J, Jurado MA, Molet J, Grafman J (1995) The role of prefrontal regions in the Stroop task. Neuropsychologia 33:341–352.

Wallis G, Stokes M, Cousijn H, Woolrich M, Nobre AC (2015) Frontoparietal and Cingulo-opercular Networks Play Dissociable Roles in Control of Working Memory. J Cogn Neurosci 27:2019–2034.

Wallis JD, Anderson KC, Miller EK (2001) Single neurons in prefrontal cortex encode abstract rules. Nature 411:953–956.

Wang X-J (2021a) 50 years of mnemonic persistent activity: quo vadis? Trends Neurosci 44:888–902.

Wang X-J (2021b) 50 years of mnemonic persistent activity: quo vadis? Trends in Neurosciences 44:888–902.

Watanabe K, Funahashi S (2014) Neural mechanisms of dual-task interference and cognitive capacity limitation in the prefrontal cortex. Nat Neurosci 17:601–611.

Watanabe M (1996) Reward expectancy in primate prefrontal neurons. Nature 382:629–632.

Welford AT (1967) Single-channel operation in the brain. Acta Psychologica 27:5–22.

Werchan DM, Collins AGE, Frank MJ, Amso D (2015) 8-month-old infants spontaneously learn and generalize hierarchical rules. Psychol Sci 26:805–815.

Werchan DM, Collins AGE, Frank MJ, Amso D (2016) Role of Prefrontal Cortex in Learning and Generalizing Hierarchical Rules in 8-Month-Old Infants. J Neurosci 36:10314–10322.

Whittington JCR, Bogacz R (2017) An Approximation of the Error Backpropagation Algorithm in a Predictive Coding Network with Local Hebbian Synaptic Plasticity. Neural Comput 29:1229–1262.

Whittington JCR, Bogacz R (2019) Theories of error back-propagation in the brain. Trends in Cognitive Sciences 23:235–250.

Wickens C, Derrick W, Berringer D, Micalizzi J (1980) The Structure of Processing Resources: Implications for Task Configuration and Workload. Proceedings of the Human Factors Society Annual Meeting 24:253–256.

Williams M, Hong SW, Kang M-S, Carlisle NB, Woodman GF (2013) The benefit of forgetting. Psychonomic Bulletin & Review 20:348–355.

Wilson FA, Scalaidhe SP, Goldman-Rakic PS (1993) Dissociation of object and spatial processing domains in primate prefrontal cortex. Science 260:1955–1958.

Wolff MJ, Jochim J, Akyürek EG, Stokes MG (2017) Dynamic hidden states underlying working-memory-guided behavior. Nat Neurosci 20:864–871.

Woolgar A, Thompson R, Bor D, Duncan J (2011) Multi-voxel coding of stimuli, rules, and responses in human frontoparietal cortex. Neuroimage 56:744–752.

Xie Y, Hu P, Li J, Chen J, Song W, Wang X-J, Yang T, Dehaene S, Tang S, Min B, Wang L (2022) Geometry of sequence working memory in macaque prefrontal cortex. Science 375:632–639.

Yamins DLK, DiCarlo JJ (2016) Using goal-driven deep learning models to understand sensory cortex. Nat Neurosci 19:356–365.

Yu Q, Teng C, Postle BR (2020) Different states of priority recruit different neural representations in visual working memory. PLOS Biology 18:e3000769.

Zaksas D, Pasternak T (2006) Directional signals in the prefrontal cortex and in area MT during a working memory for visual motion task. J Neurosci 26:11726–11742.

Zarahn E, Aguirre GK, D'Esposito M (1999) Temporal isolation of the neural correlates of spatial mnemonic processing with fMRI. Cognitive Brain Research 7:255–268.

Zhou X, Zhu D, Qi X-L, Lees CJ, Bennett AJ, Salinas E, Stanford TR, Constantinidis C (2013) Working memory performance and neural activity in prefrontal cortex of peripubertal monkeys. J Neurophysiol 110:2648–2660.

Zipser D, Andersen RA (1988) A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. Nature 331:679–684.

Supplementary Information for Chapter 2

This appendix contains additional results for networks from Experiment 2. Networks were evaluated on a test dataset with the delay interval length set to the maximum one used during training (i.e., 9 cycles), to see if the Cued Geometry varied from the one described for networks from Experiment 1 . Two measures of the geometry are reported - plane angles θ (**Fig. S1A**) and subspace alignment index (AI, **Fig. S1B**).

The results were highly similar to those reported for networks from Experiment 1. That is, across all trained models, the angles θ between the two cued subspaces were significantly clustered around the 90° mark in the pre-cue delay (v-test results: $v(29) = 22.46$, $p < .001$) and 0° in the post-cue delay (circular mean = 7.88°; $v(29) = 24.79$, $p < .001$). The mean absolute difference between pre- and post-cue angles was also significantly different from 0° (one-sample test for mean angle, $M = 103.34°$, 95% $CI = 75.39°, 131.28°$).

Similarly, the mean pre-cue AI (AI = 0.19 and 0.22 for 2 and 3 dimensions, respectively) was significantly lower than the mean post-cue AI (AI = 0.88 and 0.91, respectively) in both target dimensionalities (one-tailed one-sample t-test: $t(29) = 25.33$, $p < .001$, Cohen's d = 4.62; Wilcoxon signed-rank test $W(29) = 465$, $p < .001$).

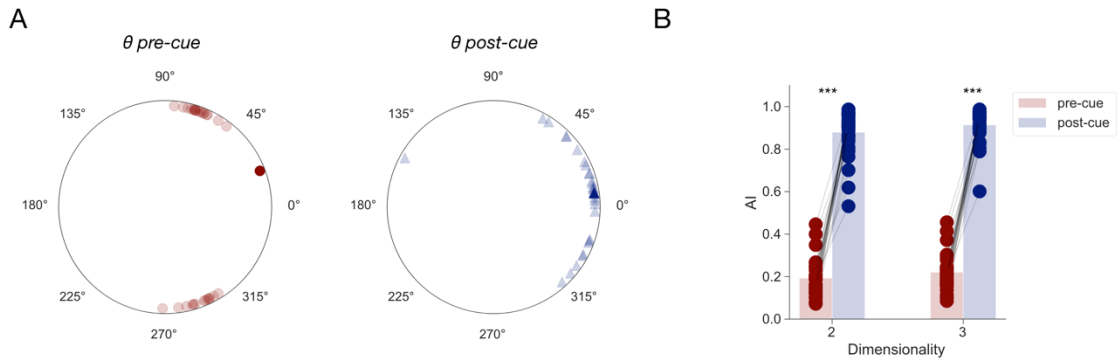


Fig. S1 Cued Geometry in networks from Experiment 2 tested on trials with the longest trained delay interval length (9 cycles), A. Plane angles θ at the end of the pre-cue (left, red markers) and post-cue delay interval (right, navy markers). Values for individual models and grand means plotted as transparent and opaque markers, respectively. **B.** Subspace AI values for the pre-cue (red) and post-cue (navy) subspaces for 2- and 3-dimensionallsubspaces. Values for individual models plotted as dots, grand means as bars. Significance values refer to the pre-cue – post-cue contrast reported in the section above. ***: $p < .001$

Supplementary Information for Chapter 4

This appendix contains information about the Boolean category structures for rule families 3[2], 3[3], 4[2], 4[3], and 4[4].

Rule type	DNF	Shortest formula	BC	Category structure 
I	$a'b'c' + a'b'c$	$a'b'$	2	
II	$a'b'c' + a'bc$	$a'(b'c' + bc)$	5	
III	$a'b'c' + abc$	$a'b'c' + abc$	6	

Table 1. Characterisation of rule family 3[2] adapted from Feldman (2000). Each rule type (denoted by a roman number; rows) is listed alongside its DNF, shortest Boolean formula derived by Feldman (2000), Boolean complexity (BC) and a visualisation of the positive exemplars. For increased readability, a conjunction $a \wedge b$ is written as ab, disjunction $a \vee b$ as $a + b$, and negation $\neg a$ as a'.

Rule type	DNF	Shortest formula	BC	Category structure 
I	$a'b'c' + a'b'c + a'bc'$	$a'(bc)'$	3	
II	$a'b'c' + a'b'c + abc'$	$a'b' + abc'$	5	

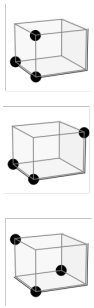
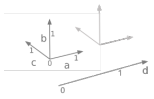
III	$a'b'c' + a'bc + ab'c$	$a'(b'c' + bc) + ab'c$	8	

Table 2. Characterisation of rule family 3[3] adapted from Feldman (2000).



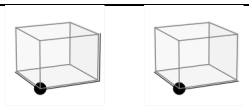
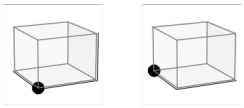
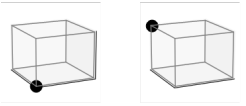
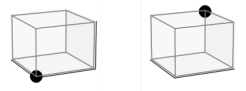
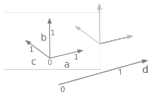
Rule type	DNF	Shortest formula	BC	Category structure
I	$a'b'c'd' + a'b'c'd$	$a'b'c'$	3	
II	$a'b'c'd' + a'b'cd$	$a'b'(c'd' + cd)$	6	
III	$a'b'c'\mathrm{d}' + a'bcd$	$a'(b'c'd' + bcd)$	7	
IV	$a'b'c'd' + abcd$	$a'b'c'd' + abcd$	8	

Table 3. Characterisation of rule family 4[2] adapted from Feldman (2000).



Rule type	DNF	Shortest formula	BC	Category structure
I	$a'b'c'd' + a'b'c'd + a'b'cd'$	$a'b'(cd)'$	4	

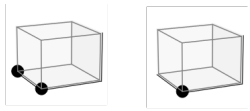
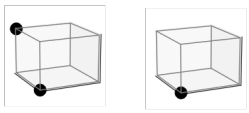
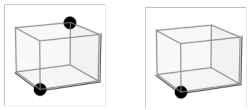
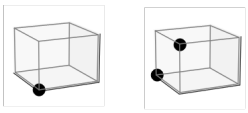
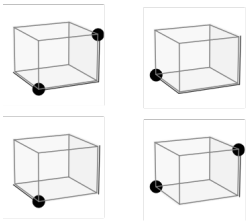
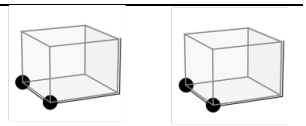
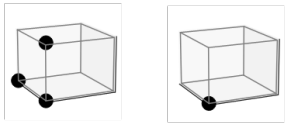
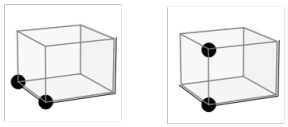
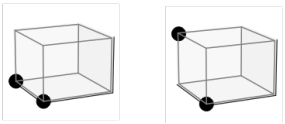
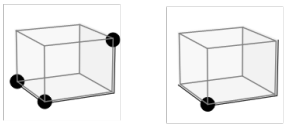
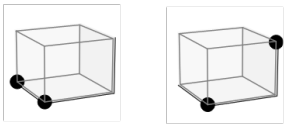
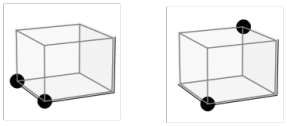
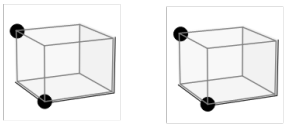
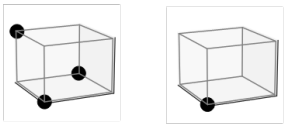
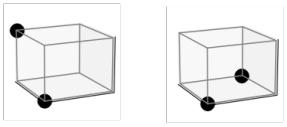
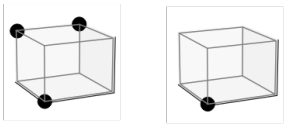
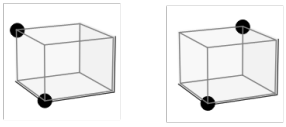
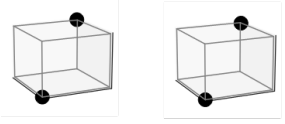
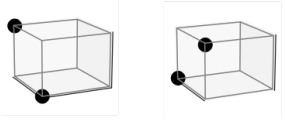
II	$a'b'c'd' + a'b'c'd + a'bcd'$	$a'(b'c' + bcd')$	6	
III	$a'b'c'd' + a'b'c'd + abcd'$	$a'b'c' + abcd'$	7	
IV	$a'b'c'd' + a'b'cd + a'bc'd$	$a'(b'(c'd' + cd) + bc'd)$	9	
V	$a'b'c'd' + a'b'cd + abc'd'$	$a'b'(c'd' + cd) + abc'd'$	10	
VI	$a'b'c'd' + a'b'cd + abc'd$	$a'b'(c'd' + cd) + abc'd$	10	

Table 4. Characterisation of rule family 4[3] adapted from Feldman (2000).

Rule type	DNF	Shortest formula	BC	Category structure
I	$a'b'c'd' + a'b'c'd + a'b'cd' + a'b'cd$	$a'b'$	2	
II	$a'b'c'd' + a'b'c'd + a'b'cd' + a'bc'd'$	$a'(b'(cd)' + bc'd')$	7	
III	$a'b'c'd' + a'b'c'd + a'b'cd' + a'bc'd$	$a'(b'(cd)' + bdc')$	7	
IV	$a'b'c'd' + a'b'c'd + a'b'cd' + a'bcd$	$a'(b'(cd)' + bcd)$	7	
V	$a'b'c'd' + a'b'c'd + a'b'cd' + abc'd'$	$a'b'(cd)' + cd) + abc'd'$	8	
VI	$a'b'c'd' + a'b'c'd + a'b'cd' + abc'd$	$a'b'(cd)' + abc'd$	8	
VII	$a'b'c'd' + a'b'c'd + a'b'cd' + abcd$	$a'b'(cd)' + abcd$	8	
VIII	$a'b'c'd' + a'b'c'd + a'bcd' + a'bcd$	$a'(bc + b'c')$	5	
IX	$a'b'c'd' + a'b'c'd + a'bcd' + ab'cd'$	$cd'(a'b + ab') + a'b'c'$	9	
X	$a'b'c'd' + a'b'c'd + a'bcd' + ab'cd$	$a'(b'c' + bcd') + ab'cd$	10	
XI	$a'b'c'd' + a'b'c'd + a'bcd' + abcd'$	$bcd' + a'b'c'$	6	
XII	$a'b'c'd' + a'b'c'd + a'bcd' + abcd$	$bc(a'd' + ad) + a'b'c'$	9	
XIII	$a'b'c'd' + a'b'c'd + abcd' + abcd$	$abc + a'b'c'$	6	
XIV	$a'b'c'd' + a'b'cd + a'bc'd + a'bcd'$	$a'(b(c'd + cd') + b'(c'd' + cd))$	11	

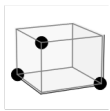
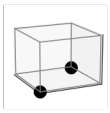
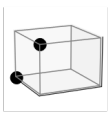
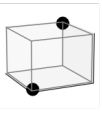
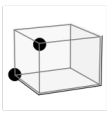
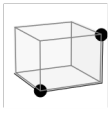
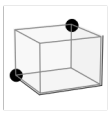
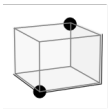
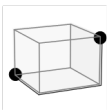
XV	$a'b'c'd' + a'b'cd + a'bc'd + ab'c'd$	$a'(b'(c'd' + cd) + bc'd) + ab'c'd$	13		
XVI	$a'b'c'd' + a'b'cd + a'bc'd + ab'cd'$	$a'(b'(c'd' + cd) + bc'd) + ab'cd'$	13		
XVII	$a'b'c'd' + a'b'cd + a'bc'd + abcd'$	$a'(b'(c'd' + cd) + bc'd) + abcd'$	13		
XVIII	$a'b'c'd' + a'b'cd + abc'd' + abcd$	$(c'd' + cd)(ab + a'b')$	8		
XIX	$a'b'c'd' + a'b'cd + abc'd + abcd'$	$ab(c'd + cd') + a'b'(c'd' + cd)$	12		

Table 5. Characterisation of rule family 4[4] adapted from Feldman (2000).

Supplementary Information for Chapter 5

This appendix contains the demographic information and exclusion details for the participant cohort from Chapter 5.

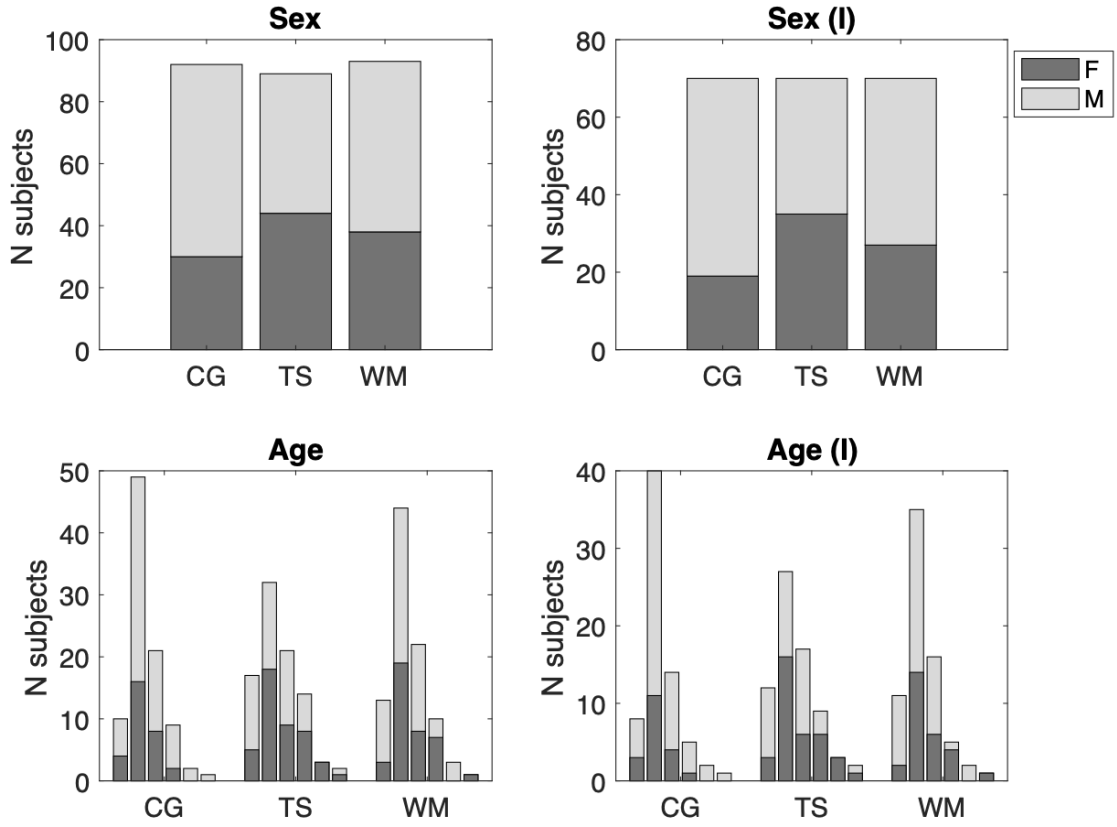


Figure S1. Demographic information about the subject cohort from Study 1. *Top left:* Total number of females (dark grey) and males (light grey) in each experimental group. *Top right:* Number of females and males included in the main analyses for each experimental group. *Bottom panels:* Number of subjects in each of the age brackets (18-20, 21-30, 31-40, 41-50, 51-60, 61+ years old). Figure conventions as in the top panels. *CG:* control group, *TS:* temporal segmentation, *WM:* working memory. *I:* subjects included in the analysis

Group	Total N	N Excluded	% Excluded

CG	92	22	23.9%
TS	89	19	21.3%
WM	93	23	24.7%

Table S1. Participant exclusion details for Study 1. Subjects were excluded from analysis if they failed to reach the criterion as specified in the Methods section. *N* - number of subjects

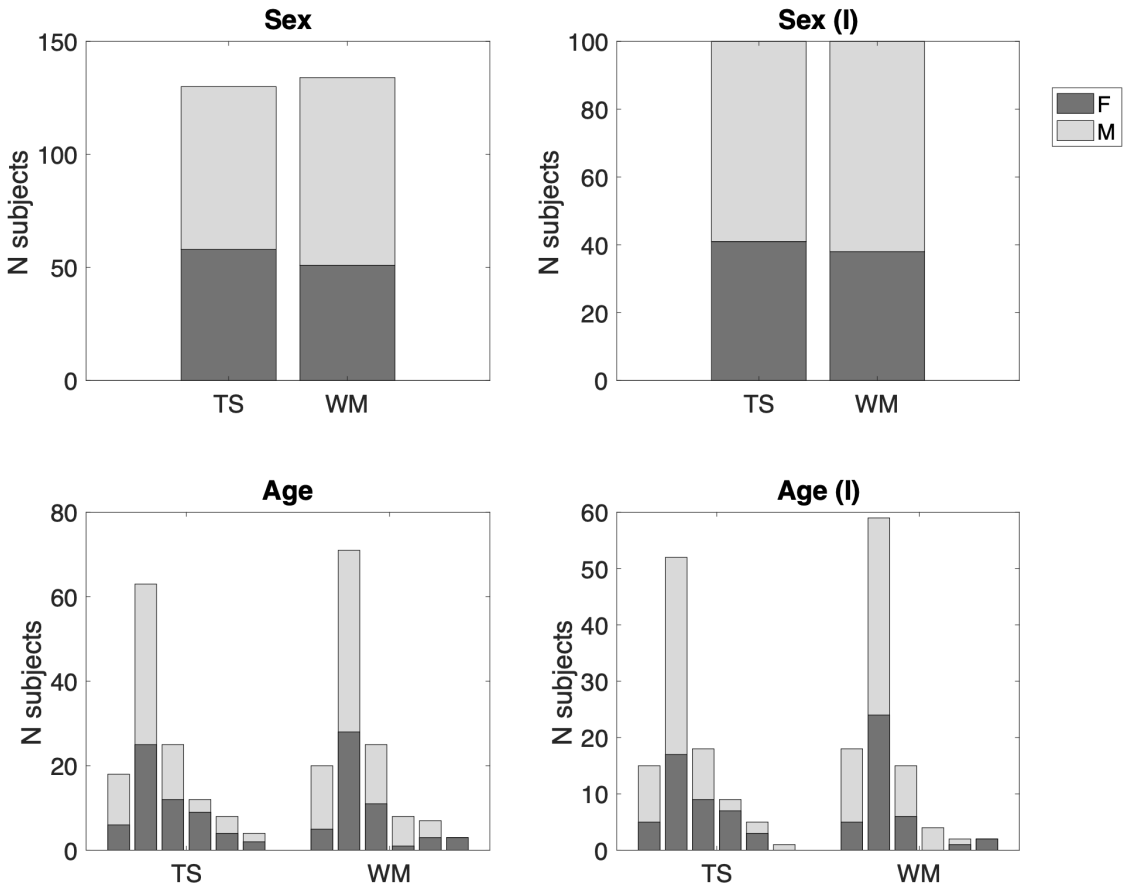


Figure S2. Demographic information about the subject cohort from Study 2. All conventions as in Fig S1.

Group	Total N	Excluded N (TD)	Total N (-TD)	Excluded N (C)	% Excluded (C)

TS	130	0	130	30	23.1%
WM	139	5	134	34	25.4%

Table S2. Participant exclusion details for Study 2. Subjects were excluded from analysis due to technical difficulties (TD) and for not reaching the criterion (C). *Total (-TD)* - total number of subjects adjusted for technical difficulties; *% Excluded (C)* - percentage of subjects excluded for not reaching the criterion relative to the TD-adjusted total.