Dynamics and dimensionality of information representation for higher cognitive function

Dante Francisco Wasmuht
St. John's College, Oxford

D.Phil. Thesis

Michaelmas Term, 2019

Dynamics and dimensionality of information representation for higher cognitive function

Dante Francisco Wasmuht

St. John's College Michaelmas Term, 2019

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

Abstract

Many cognitive functions are thought to rely on higher brain regions, in particular the prefrontal cortex. However, neural processing in prefrontal cortex can be notoriously complex, both in single neurons as well as at the population level. This thesis aims to investigate some of the representations and dynamics that emerge from the concerted activity of many neurons during cognitive processing. All experiments in this thesis involved the analysis of electrophysiological recordings from non-human primates taken while the animals performed complex cognitive tasks involving working memory and decision making components. In the first experiment, I analysed prefrontal cortex recordings from animals trained on a cognitively demanding dual-task that required simultaneous performance of two overlapping tasks. I found that many neurons were tuned to mixtures of features from both tasks, thus indicating possible interference on the neural level. However, despite mixed responses in single neurons, population level representations of the two tasks were nearly orthogonal effectively resolving interference between task representations. The temporal dynamics of the dual-task seemed to play a critical role in this orthogonalization process. In a second experiment I took a closer look at the temporal dynamics during cognitive processing by analysing recordings from multiple brain regions while animals performed a working memory task. I found that the heterogeneous dynamics of single neurons i.e. dynamic or stable delay coding could be predicted by their intrinsic temporal firing stability. In the last experiment I investigated how correlations between neurons influence the processing of decision related variables in single neurons. I found that decision related activity in response to a perceptually ambiguous stimulus was predicted by different timescales of interneuronal correlations. Finally, I summarize the major findings of this thesis and discuss future lines of research that aim to explore the circuit mechanisms underlying the observed task representations and how they might arise during learning.

This thesis was generously funded by a 1+3 year BBSRC studentship

Approximate word count: ~ 69 000

Acknowledgements

My utmost gratitude goes to my two supervisors Mark Stokes and Kristine Krug. I want to thank Mark for giving me the possibility and freedom to pursue all sorts of scientific endeavours over the past three years, ranging from exploring human intracranial recordings to working with Japanese macaques. I could not have wished for a more diverse and exciting ride. I would also like to thank Kristine for kindly supporting me throughout my entire DPhil and for supervising my rotation project, which made it all the way into this thesis.

I owe many thanks to the various people that I have collaborated with on the projects that form the chapters of this thesis. First, I would like to thank Kei Watanabe who has provided the data presented in chapter 2 and who is collecting the data for the learning experiments mentioned in chapter 5. I also want to thank Kei for hosting me several times in Japan and for sharing his fantastic experimental skills. I am grateful to Tim Buschman and Earl Miller for providing the data presented in chapter 3, as well as to André Bastos who collected the laminar data, mentioned in chapter 5. Special thanks go to Eelke Spaak, for brilliant supervision during the first year of my DPhil and collaboration on chapter 3 and to Nicholas Myers for fruitful scientific collaboration - culminating in what has now become chapter 2. I would like to thank the BBSRC for supporting me financially and the people at the DTC in general for their continued support.

I would also like to thank past and present members of the Stokes lab for enriching lab life, scientifically and otherwise, in no particular order: Sam Hall-McMaster, Michael Wolff, Frida Printzlau, Michał Wójcik (for being a great MSc student), Ilenia Salaris, Darinka Trübutscheck, Emilia Piwek, Sofia Deleniv, Lev Tankelevitch, MaryAnn Noonan, Janina Jochim, Wen Wen, Paul Muhle-Karbe, Andrea Bocincova and Jasper van der Meulen. I would also like to thank the welcoming people at OHBA for hosting us after the apocalyptic breakdown of our Department.

Last but not least, thank you to old and new friendships and of course to my family. Especially: my parents, my sister, my brother and my dear wife.

Table of Contents

Table of Contents 5

Chapter 1. General Introduction 7

1.1 Introductory remarks 7

1.2 Higher cognitive function and the prefrontal cortex 7

1.3 Neuronal responses in PFC 14

1.4 Information that hides in the pattern of neural activity 21

1.5 Maintenance of information 39

1.6 Synopsis 46

Chapter 2. Dynamical multiplexing of information during dual-task performance in monkey prefrontal cortex 49

2.1 Abstract 49

2.2 Introduction 50

2.3 Methods 53

2.4 Results 76

2.5 Discussion 102

Chapter 3. Intrinsic neuronal dynamics predict distinct functional roles during working memory 113

3.1 Abstract 113

3.2 Introduction 114

3.3 Methods 117

3.4 Results 127

3.5 Discussion 144

Chapter 4. Interneuronal correlations at longer timescales predict decision signals for bistable structure-from-motion perception. 154

4.1 Abstract 154

4.2 Introduction 155

4.3 Methods 159

4.4 Results 170

4.5 Discussion 186
Chapter 5. General Discussion 195
5.1 Chapter summaries and contributions 195
5.2 Limitations and outstanding questions 200
5.3 Future directions 205
References 208
Appendix 1. Supplementary Figures for Chapter 2 238
Appendix 2. Supplementary Figures for Chapter 3 244
Appendix 3. Supplementary Figures for Chapter 4 254
Appendix 4. Supplementary Figures for Chapter 5 265

Chapter 1. General Introduction

1.1 Introductory remarks

The experimental chapters of this thesis are relatively self-contained, addressing specific questions and comprising their own introductions and discussions. The aim of this general introduction (**chapter 1**) is therefore to present important concepts and frameworks that set the stage for the experimental work and offer some contextualization. I would like to start this chapter by giving a broad description of what has been conceptualized as cognitive control in Psychology and Neuroscience, and how it relates to the prefrontal cortex. Subsequently, I will elaborate on the neural signals underlying aspects of cognitive control. I will start by focusing on the types of information carried by single prefrontal neurons. I will then diverge from the single neuron perspective and give an overview of how information can be represented in distributed codes at the level of the population. This will be followed by a discussion on how to interpret the representations emerging from distributed codes, in which I try to argue that many cognitive processes are best examined at the level of neural populations. Finally, I will introduce the concepts of dynamic coding and how this relates to working memory.

1.1 Higher cognitive function and the prefrontal cortex

1.2.1 How can we understand cognition?

A reasonably safe first principle regarding the function of cognition is given in the environmental complexity thesis (ECT), which states: The function of cognition is to enable an agent to actively deal with environmental complexity (Godfrey-Smith, 2001). Hence, cognition can be understood as a collection of capacities, such as perception, memory, decision making, internal representations (i.e. models of the world), learning and production of behaviour which, in orchestration, allow agents to achieve a directed coordination between their actions and the world. In this context, 'higher' cognitive function - as used here - means an advanced instantiation of controlling behaviour in the face of environmental complexity. Let us take an example: imagine life as a jellyfish. Although you possess only a few nerve cells, you can swim up or down in response to somatosensory stimulation or a given salt concentration (Albert, 2011), and probably a few things more. This implies that you have some capacity that allows you to react to the environment and changes therein. In a sense, you possess low level cases of some of the capacities, which in more elaborate cases can constitute cognition (Godfrey-Smith, 2001). However, as a jellyfish, you cannot imagine or plan alternative scenarios i.e. actions leading to a different situation. You cannot even want alternative scenarios, and above all, you cannot remember that you could not imagine, plan or want an alternative scenario i.e. you cannot hold things in mind i.e. in your brain (since you do not have one). All of these things (among others) can be described as higher cognitive functions (Frith and Dolan, 1996), and a full description is out of the scope of this thesis. It is a difficult and probably misleading endeavour to find a clear cut boundary between simple behaviour and cognitive function. However, millions of years of evolution along the lines of ECT have endowed many animals, and above all humans, with remarkable capabilities to control their behaviour in the face of environmental complexity.

Importantly, this is often achieved flexibly i.e. without the need to go through lengthy processes of adaptation that change the neural circuits for the long term.

1.2.2 What is cognitive control?

To get a handle on how higher cognitive functions act in the brain, Psychology has coined the term cognitive control or executive function (Cohen, 2017). Broadly speaking, cognitive control acts through 'higher level' brain mechanisms that control and orchestrate 'lower level' processes towards a common goal (Miller, 2001). Let us consider a concrete example to clarify this view: First, you are not a jellyfish anymore, instead, you are standing at a busy crossroads and the light turns green. If your intention was to cross the road you know that now is the time. As you cross the road you confidently ignore any approaching cars since you assume they will stop at the intersection. Alternatively, if it's 3am in the morning and nobody is around you might cross the road ignoring all traffic lights (this is a fictional account) but paying close attention to whether a car is approaching. In this example, lower level processes include all kinds of sensorimotor processes but also memories and so on. More generally, they can include behavioural mechanisms that can be seen as reflexive or rather automatic and habitual stimulus response mappings e.g. orienting towards a salient stimulus (car) or only crossing a road if there is a green traffic light. These processes are usually fast and efficient because they are hard coded into the brain's fundamental structure through experience, and are often referred to as 'bottom up' (Miller and Cohen, 2001). In the face of such 'lower level' or 'bottom up' processing, cognitive control acts through a 'higher level' i.e. 'top down' process that is more flexible, capable of supressing, augmenting and modulating those 'other' processes in order to achieve flexible, context-

dependent behaviour that is decoupled from a direct mapping and geared towards a certain goal.

1.2.3. Where is cognitive control?

According to Miller and Cohen (2001), a system within the brain that could exert cognitive control via top down mechanisms should fulfil certain requirements. It should have access to and be able to influence a wide range of other brain regions, that is, be able to exchange information with them. This should include information about both internal (e.g. a rule) and external variables (e.g. a stimulus). As a result, representations within the system could reflect a wide range of associations signalling convergence and integration of behavioural relevance. Representations within the system should be sustained (i.e., robust to interference), but also flexible (able to rapidly switch with changing environmental demands). Author's note: Many of these assumptions still seem appropriate although some might need to incorporate new findings (e.g. the nature of sustained activity, as will be discussed later on). Miller and Cohen (2001) then put forth the prefrontal cortex as a brain region that satisfies most of the requirements to exert cognitive control.

1.2.4. (Coarse) Anatomy of the PFC

The prefrontal cortex (PFC), also called the association cortex of the frontal lobe (Fuster, 2002) forms the apex of the cortical hierarchy. Phylogenetically, PFC is one of the latest cortices to develop, with the largest relative size in humans (Jerison, 1994; Donahue et al., 2018), where it constitutes about one third of the neo-cortex (Fuster, 2002). For

comparison: <4% of total cortical volume in cats; ~10% in macaques; ~17% in apes; ~30% in humans. In humans, PFC also develops late in ontogeny (Flechsig, 1920; Huttenlocher, 1990; Sowell, 1999), reaching maturity approximately during late adolescence (Paus et al., 1999; Diamond, 2002). The PFC of humans and primates represents a densely interconnected set of sub-regions (Catani, 2012). There is little consensus about where to draw the boundaries between those subdivisions (Passingham and Wise, 2012 (book)) and despite recent efforts to delineate functionalities (e.g. Buckley et al., 2009; Hunt et al., 2018; see Szczepanski and Knight, 2015 for review), many cognitive processes recruit overlapping sub-regions of the PFC network (Duncan, 2010; Wilson et al., 2010). Nonetheless, a broad classification is helpful for the remainder of this chapter: In humans, the PFC can be roughly divided into two subdivisions: the ventromedial PFC (vmPFC; all mammals), and the lateral PFC (lPFC; only primates) (Striedter, 2004, 2005). The medial part of vmPFC can be further divided into the medial PFC (mPFC), which contains the anterior cingulate cortex (ACC) and the ventral PFC (vPFC), which contains the orbitofrontal cortex (OFC). The lPFC can be further divided into the ventro-lateral PFC (vlPFC) and dorso-lateral PFC (dlPFC). Importantly, PFC has reciprocal connections with nearly all cortical and subcortical structures (Croxson et al., 2005; Ilinsky et al., 1985; Selemon and Goldman-Rakic, 1988; Alexander, 1986). These include interconnections with virtually all of sensory neocortex and multimodal inputs from a range of association cortices (e.g. Pandya & Barnes, 1987; see Miller and Cohen (2001) or Fuster (2001) for a more extensive list of references). The dlPFC is particularly strongly connected to motor cortices (Fuster, 2001 for review); superior colliculus (Schmahmann and Pandya, 1997) and other deep structures. Furthermore, PFC sends and receives inputs from the limbic system (specifically OFC and vmPFC) either directly or indirectly (via the Thalamus); for

example, the hippocampus and the amygdala (Amaral and Price, 1984; Barbas & De Olmos 1990; Goldman-Rakic, 1984, van Hoesen et al., 1972) and the basal ganglia (Alexander, 1986). In macaques, lPFC is anatomically connected to over 80 regions as opposed to the primary visual cortex, which is connected to only 30 (Markov et al., 2014). In total, this satisfies the first requirement in the Miller and Cohen (2001) proposal and puts PFC in an excellent position to integrate many streams of information in the brain and exert top down influence on other structures.

1.2.5. Lesions to the PFC

Some of the most compelling evidence suggesting that the PFC might play a role in processes associated with cognitive control comes from studies that had patients with PFC damage perform demanding cognitive tasks in a laboratory setting. Two of the most influential are the Stroop- and the Wisconsin card sorting (WCS) tasks. In the Stroop task (Stroop, 1935), participants have to read a word or name the colour in which it is displayed. This task requires cognitive control to suppress the default behaviour (reading the word) to accurately name the ink colour. Prefrontal patients show impaired performance with this task, especially if the word and the colour conflict e.g. the word red displayed in green (Perret 1974, Vendrell et al., 1995) or when instructions vary frequently e.g. report the colour in which the word is displayed or vice versa (Cohen et al., 1999). Similarly, during the WCS task participants have to sort cards according to one of various features e.g. shape, colour etc. The relevant feature can change and participants find out through trial and error. Patients with PFC damage often learn the first rule but are unable to switch their behaviour after the rule changes (e.g., perseveration: Milner, 1963; Barcelo and Knight, 2002). Similar results were obtained

from monkeys with a lesion to dlPFC, performing analogues of the WCS task (Dias et al., 1996b; Buckley et al., 2009; Moore et al., 2009; Walker et al., 2009). Results from those classic studies seem to suggest that PFC might be important for adhering to the goals and rules of a task and in overwriting a competing, potentially more habitual response (Miller and Cohen, 2001).

Other studies showed that patients with PFC damage have problems in executing a set of actions in the right temporal order (Duncan, 1986, Grafman, 1989), particularly if they are faced with abstract or open-ended scenarios (Shallice and Burgess, 1991; Shallice, 1982) or when they have to make a move that initially takes them further away from their end goal (Goel and Grafman, 1995). The description of patients' behaviour by Shallice and Burgess (1991) makes a striking point: Patients were able to execute individual routines such as buying bread, however, they were unable to carry out multiple errands in an organized sequence and frequently entered the wrong shop. Along similar lines, Luria (1966; 1973) observed that patients with PFC damage exhibited problems in applying a suitable strategy to a novel task, resulting in disorganised attempts to solve it. Those observations are complemented by a set of studies investigating the capacity of PFC patients to organize their behaviour during performance of tasks with multiple components. Duncan et al. (1996; 1997) noted that patients neglected certain key components, inserted irrelevant ones instead or inappropriately sequenced them. Later on, Dreher et al. (2008) and Roca et al. (2011) reported that patients with more rostral PFC lesions showed multitasking deficits in similar multi-component scenarios, which could not be explained by poor memory of any of the components nor a misunderstanding of how any of the individual components was to be carried out. It thus appears that PFC patients have difficulty in organizing

various pieces of information or sub-goals into a coherent task set and coordinate their behaviour accordingly.

Many other influential studies with human patients as well as lesion studies in primates have further dissociated processes that rely on a functioning PFC network or sub-region. Some of them are: Working memory i.e. the ability to hold in mind and manipulate information over short time periods (which will be discussed in greater detail later on in this chapter; mainly in lPFC regions); Spatial attention and inhibitory control (mainly in vlPFC); Inhibitory control and decision making, especially in relation with reward but also task switching (mainly in OFC; see Szczepanski and Knight, 2014); Assigning actions to reward on a flexible basis (mainly ACC, see Kennerley et al., 2006; Rudebeck et al., 2008). Given the vast catalogue of mostly 'higher level' functions that seem to depend on the involvement of the PFC, it is worth mentioning that most patients or monkeys with damage to their PFC typically do not lose their perceptual capacities, motor skills, long-term memory or other domain specific aspects. To go back to Miller and Cohen (2001), it then seems reasonable to assume that PFC could be a potential candidate for a system sub-serving cognitive control.

At this point, I will begin to narrow down on a central question addressed in this thesis: how does PFC represent information in a format that make it suitable for higher cognitive function? For this, I will start small by looking at what individual neurons encode in PFC.

1.3 Neuronal responses in PFC

A hallmark of PFC is the sheer diversity and flexibility of its neuronal response profiles. In many 'lower' brain areas (e.g., sensory cortices), neurons are characterized by defined tuning curves. For example, in the middle temporal visual area (MT or V5) in macaques, cells respond to certain features of visual stimuli, and do so in a highly predictable way. According to Born and Bradley (2005), those include: a) retinal position; b) direction of motion; c) speed of motion; binocular disparity (see last chapter of this thesis) and stimulus size. In fact, many of the success stories in Neuroscience derive from the characterization of single cell response profiles in certain areas e.g. 1981 Nobel prize for information processing in the visual system. Often it seems as if the distinct and exclusive tuning of single neurons in some brain regions openly display that region's function. In contrast, single neuron tuning profiles in PFC typically resist clear categorizations. To underline this, in the following I will give a short overview of the diverse response profiles displayed by PFC neurons. To not digress too far from the experimental content in this thesis, I will limit this overview to what has been found in non-human primates - especially macaques.

1.3.1 Prefrontal neurons respond to… many things

Most of the studies presented here recorded from areas pertaining to the lPFC (upper/lower banks of the principal sulcus), unless otherwise noted.

Researchers have been recording from lPFC for over 50 years, and even early on it had become clear that whatever task the animal performed, a substantial proportion of prefrontal neurons would be tuned to some aspect i.e. feature of it. In general, responses in PFC were found to reflect all (task relevant) aspects of visual stimuli e.g. location

(Funahashi et al., 1989; Wilson et al., 1993), shape (Meyer et al., 2011; in FEF: Peng et al., 2008, Clark et al., 2012), orientation (Luo and Maunsell, 2018), colour (Buschman et al., 2011; Riley et al., 2017) but also features of stimuli pertaining to other modalities e.g. tones (Fuster et al., 2000), touch (Romo et al., 1999), taste (OFC, Lara et al., 2009); and even multimodal features (Vaadia et al., 1986). Prefrontal neurons also respond to the identity of objects i.e. images (Rao et al., 1997; Rainer et al., 1998; Asaad et al., 1998; Kadohisa et al., 2015) and other multidimensional stimuli e.g. faces (Ó Scalaidhe et al., 1997,1999) and so on. Many prefrontal neurons are motor related: they respond to motor events or are selective for directions of saccadic eye movements, manual responses and so on (Watanabe, 1986; DiPelligrino and Wise, 1993; Marcos et al., 2019). Furthermore, prefrontal neurons modulate their activity at the time of reward (Niki and Watanabe, 1979) and distinguish between reward identity or magnitude (Watanabe, 1996; Cavanagh, 2018). It thus seems that PFC represents many of the ingredients that constitute a task. Importantly, it also mixes the signals that it receives.

1.3.2 Mixed neurons

For example, prefrontal neurons have been shown to be selective for an object's location as well as for its identity (Rao et al., 1997; Rainer et al., 1998; Genovesio et al., 2012). That is, they display mixed selectivity, which can be seen as evidence that neurons in PFC somehow integrate information across different stimulus features, similar to multisensory neurons in heteromodal sensory areas. The nature of such mixed signals, however, allow for very different interpretations. Consider a classic study by Assad et al. (1998): They taught monkeys a task that required them to respond to two different objects (A/B) with specific saccades (e.g. left/right). After a while, contingencies were

reversed. A large proportion (~40%) of recorded neurons responded only to a given object when it was linked to one saccade direction, but not the other. In other words, those neurons showed mixed selectivity because their response depended on a mixture of object identity and saccade direction. Conversely, this means that neurons in PFC are able to associate, that is to conjunctively code for information about different task features. In fact, the recorded sample contained a higher number of associative or mixed selectivity neurons than pure selectivity neurons (i.e., neurons that responded differently to either object identity or saccade direction). Over the years, similar observations have been made for all kinds of mixed i.e. associative signals in PFC (Pan et al., 2008; Warden and Miller, 2010; Tsutsui et al., 2016). For example, in the experiments conducted by Watanabe, (1990,1992), prefrontal neurons only responded to a certain cue if it was rewarded, and this was further dissociated by whether the cue was auditory or visual. Interestingly, those exact mechanisms underlie what can be interpreted as selectivity to a task rule or task context selevtivity. This is especially clear in some of the earlier studies including the work by Assad et al. (1998) (but also see Hoshi et al. (1998) or White and Wise (1999)). There, neurons showed 'rule coding' when specific actions were associated with distinct cues, but cue-action associations changed flexibly as a function of an explicit cue. Similar 'rule tuned' neurons were also observed when the current rule had to be inferred through trial and error (Mansouri et al., 2006).

1.3.3 Abstract neurons

Only a few years after those studies the notion of a rule or contextual signal in PFC was significantly expanded by Wallis et al. (2001). In their task, monkeys saw a sequence of two objects and depending on the rule in effect, had to release a lever if the two objects

were the same i.e. match or not i.e. non-match. Crucially, the match rule was cued by either a drop of juice or a low tone, while the non-match rule was cued by no juice and a high tone. As expected, Wallis et al. (2001) found a large proportion of prefrontal neurons that displayed rule selectivity as well as mixed selectivity to rule/object pairs and so on. However, this time the interpretation of a rule-like signal was different: it actually displayed an abstracted or factorized signal that generalized over two cues from two different modalities (tone & juice). In other words, rule selective neurons compressed the information about individual rule cues, to a binary, highly relevant signal simply indicating which rule was currently active. Further studies from the same group could show that such abstracted or generalized signals were ubiquitous in PFC, for a range of tasks. Particularly influential was a study by Freedman et al. (2001). There, monkeys were trained to assign ambiguous stimuli with varying degrees of 'cat-ness' and 'dog-ness' into definite cat and dog categories. Many neurons primarily signalled the stimulus category, rather than a parametric estimate of its identity, again, suggesting that neurons in PFC are able to form an abstracted representation (in this case, an animal), through generalizing over within-category dissimilarities. Furthermore, Cromer et al. (2010) showed that category sensitive neurons, as defined on the dog vs. cat stimulus set, were equally selective for categories based on a new stimulus set (cars). This further underlines the fact that neurons in PFC present task related information in an abstracted format, entirely uncoupled from a purely feature or stimulus dependent signal. Moreover, the abstracted categorical signal gained in one task can be readily applied in a new task setting with the same structure. Following these studies, evidence for prefrontal neurons encoding abstracted information has steadily accumulated, revealing neurons coding for functional categories, as opposed to the perceptual categorizations described above (Brincat and Miller, 2015; Tsutsui et al., 2016; see also

Wallis et al., 2001); Numerosity (Nieder and Miller, 2003); Logical operations (Eiselt and Nieder, 2013; Cohen et al., 2019 (ACC)); Behavioural sequences (Shima et al., 2007); Strategy (Genovesio et al., 2005); Decision variables (Xie and Padoa-Schioppa, 2016, (OFC)) , and this list is still expanding.

1.3.4 Flexible neurons

Nothing is static, everything flows (Heraclitus).

As mentioned in previous paragraphs, literally any cognitive task that is being performed likely evokes responses in PFC that encode the currently relevant behavioural parameters. Moreover, large proportions (40% or more) of randomly recorded neurons are typically engaged in representing those parameters, even more so if the task is demanding (Duncan and Miller, 2002). At the same time, the majority of studies report that any given type of selectivity is relatively evenly distributed across the recorded sample; and although more recent studies propose gradients of selectivity types in lPFC, particularly linked to the degree of abstraction (Wutz et al., 2018, see Costantinidis and Qi, 2018 for review) as well as moderate clustering of neurons according to location preference (Markovitz et al., 2015; Masse et al., 2017; Bullock et al., 2017; Kollmorgen et al., 2019), clear structural organizations e.g. retinotopic or spatial maps are clearly absent (Costantinidis and Qi, 2018). Furthermore, neurons in PFC switch their selectivities with changing task demands (Rao et al., 1997; Roy et al., 2010), and even over the course of a single trial (Rao et al., 1997; Spaak et al., 2017). Take for example the experiment by Rao et al. (1997). There, monkeys had to remember the identity and location of a target. However, at a later phase in the trial, identity information could be

dropped; only location remained relevant. Strikingly, this was reflected in the neural coding. Cells that initially coded for target identity (or mixtures of target identity and location) dropped their identity selectivity and started to exclusively code for location instead. These results combine two important observations. First, neurons in PFC care about task relevant information i.e. location in this case, and second, prefrontal neurons can switch their selectivity i.e. they can flexibly code for different things, even within a trial. A similar conclusion can be gained from a study by Roy et al. (2010). They recorded prefrontal activity during a categorization task, similar to the ones described previously (i.e. dog vs. cat), while changing the category boundary within the same stimulus set. Many neurons that showed categorical selectivity during one boundary condition rarely continued to do so in the other boundary condition (or often stopped to encode the former when it became irrelevant). Recordings over the last 30 years have continuously contributed evidence indicating that prefrontal neurons are not hardwired to do one thing only, but instead flexibly encode task relevant variables according to current demands (Duncan, 2001; Duncan, 2010; Stokes et al., 2017). Together, this has led to formulations such as dynamic or 'adaptive coding', firstly put forward by Duncan (2001) (see also Duncan and Miller, 2002). In general, 'adaptive coding' describes PFC as a neural resource pool or workspace that can be recruited flexibly in order to represent or perform computations on currently relevant information to achieve behavioural goals (Stokes et al., 2017). While ideas such as adaptive coding might form little more than a loose framework in which to view neural responses in PFC, it resonates with modern approaches in computational neuroscience that use large recurrently connected neural networks (RNNs) trained to perform a specific task or sets of tasks to model prefrontal activity and characterize the underlying computations (e.g. see Mante et al., 2013; Enel et al., 2016; Chaisangmongkon et al., 2017; Masse et al., 2019).

In this section I aimed to summarize three important functional properties of prefrontal neurons. First, they represent complex mixtures of task related features, commonly referred to as mixed selectivity. Second, many prefrontal neurons show abstract responses, that is their responses are decoupled from distinct features or combination of features, but generalize commonalities across exemplars (perceptual or functional). And third, response selectivities are not static but flexible, both over time and in content. All these properties are somehow orchestrated to form a task representation that might be used to guide adaptive behaviour. However, individual tuning profiles can only reveal snippets of such an underlying representation. The following section will elaborate on how the pattern of activity on the neural population can be leveraged to gain a better understanding of how tasks in particular and information in general might be represented in PFC.

1.4 Information that hides in the pattern of neural activity

1.4.1 Local coding and grandmothers

The enormous success of single neuron electrophysiology that started during the first half of the 20th century led to the view of the brain as a hierarchical confederation of highly specialist i.e. feature detecting neurons, each dedicated to represent a particular perceptual, behavioural or cognitive variable. According to this view, information processing in cortex starts relatively simple with neurons detecting basic stimulus properties e.g. color or location in the retina. Along the hierarchy, neurons become more specialized by integrating signals from many different 'simpler' cells. At the extreme end of the hierarchy the convergence of signals culminates in so called 'grandmother'

cells i.e. cells that respond only to your grandmother and nothing else. A representation of a particular state of affairs e.g. your grandmother frying an egg thus involves some grandmother cells, a few egg cells and another set that likes frying. In a situation where your grandmother instead fries bacon, the egg cells would not respond at all but would be replaced by bacon cells. The high dependence on a distinct set of cells being active in any given situation earned this view the name 'single neuron doctrine' (Barlow, 1972; Shepherd, 1991), sometimes referred to as local coding (Churchland and Seijnowski, 1992). There is a certain appeal in the idea of local coding in that it simply extents the notion of feature detectors to incorporate complex states of affairs, hence giving a simple explanation of how we are able to rapidly distinguish between or act upon situations e.g. either the bacon cells are active or not. While there is no question that neuronal responses correlate with certain features, sensory or otherwise, a pure view of local coding quickly clashes with the types of neural responses observed in higher order brain regions. Let us take the famous example of neurons in the human medial temporal lobe that respond to particular faces (Kreiman et al., 2000), perhaps one of the strongest pieces of evidence for feature selectivity (Yuste, 2015). It is hard to imagine that the researchers were lucky enough to find the one out of a million cells that uniquely responded for one of the tested faces. This means that there must be many cells that respond to particular faces. If a given cell only codes for one face, we very quickly run out of neurons to represent any given face. A coding scheme that seems much more plausible is that information pertaining to any particular face is distributed across a large population of neurons (Chang and Tsao, 2017).

Local coding is at odds with cells switching their selectivity and doing more than one thing in more than one task, in the very areas that are most linked to the ultimate stages

of perceptual cognitive and memory processing (Eichenbaum, 2018). Instead, a distributed code seems more likely to underlie the complex responses observed in PFC, described before.

1.4.2 Distributed codes - Introduction

The perseverance of the local coding view during most of the 20th century was tightly linked to the success of single neuron recordings, micro-stimulation and the mapping of receptive fields throughout the brain (Yuste, 2015). At the same time, techniques to record from more than one neuron were limited and there was no clear breakthrough regarding the processing specificities of populations or assemblies of cells (Churchland and Sejnowski, 1992). Rather, theories regarding the distributed nature of the neural code were limited to simulation studies using artificial neural networks (MClulloch and Pitts, 1943; Rosenblatt, 1958; Marr, 1970). The basic idea of a distributed code in the brain is that an assembly of cells (Hebb, 1949, 2005) i.e. an interconnected group of neurons can represent the composite features of concepts or events, without the need for an individual neuron to specifically represent that entity. In other words, the relevant information lies in the pattern of activity between neurons, not necessarily in single neurons. Additionally, systems in which many elements are interconnected can generate emergent properties (Anderson, 1972), which are by definition not present within the individual elements. Recent years have seen a gigantic leap in the number of neurons that can be recorded simultaneously. Likewise, analysis techniques to study activity at the level of neural population have developed at a fast pace. Together, this has opened up the stage to probe how information is represented within the distributed patterns between neurons, which has already begun to further our understanding of the

processing mechanisms underlying the implementation of complex behaviour. In the following I will offer some visual explanations of what can be understood as a population representation and give some framework on how such representations are usually interpreted. I will then give a few examples of how the analysis of population representations has helped to better understand neural computations underlying behaviour, which would not have been possible by focusing on single neurons. As before, I will particularly focus on PFC but also describe findings from other higher-level brain areas.

1.4.3 Distributed codes – Interpreting representations

The basic idea behind a distributed neural code is that the activity over a set of neurons (*N*) can be represented as a vector. Each entry in this population vector is the activity of one neuron (*i*). Conversely, the population vector represents the neural population activity as a point in the space spanned by all *N* neurons, which is referred to as state space. In a sense, a point or collection of points in state space is what can be called a population response or for that matter representation of a specific feature, collection of features or even a task, depending how far one wants to go. Consider the example in Fig. 1.1. State space in this case is defined by a population of three neurons, each of which determines an axis i.e. direction. In other words, here, state space is three dimensional (for visualization purposes). A hypothetical task could be to remember the location of a stimulus varying its location on each trial and at the end of each trial to reach towards the remembered stimulus location. The population response to e.g. eight distinct trials is depicted as eight circles. The spatio-temporal dynamics of the task can be represented as trajectories in state space. In this case, the population representation

can be described as three dimensional because it explores all three dimensions of state space. Note, the maximal dimensionality of a representation is limited by the number of neurons i.e. the dimensionality of the state space (but also depends on correlations between neurons, task complexity etc. etc. see e.g. Gao and Ganguli, 2015, for a theoretical discussion). Often, a representation (or aspects thereof) is sufficiently described within a lower dimensional subspace, that is a set of new axes or directions defined within state space (See Fig. 1.1b square on lower right). Typically, those new axes are defined using dimensionality reduction techniques e.g. principal component analysis (PCA) or similar. As can be seen in Fig. 1.1, the neural representation contains information about the position of the experimental stimuli as well as time, which is due to the fact that one neuron encodes vertical position, a second encodes horizontal position and a third simply varies its activity with time. Importantly, there is no need for eight individual neurons representing one stimulus location each. Furthermore, one could easily imagine any sort of rotation of the population representation, which would lead to less interpretable neural responses i.e. mixtures of time with vertical and horizontal stimulus information, while the characteristics of the population representation would be essentially conserved. Under this view, responses of individual neurons can correlate with certain features of stimuli or tasks; but even if they do not, they might still contribute to the population representation. Thus, shifting the focus on the population representation does away with some of the interpretational difficulties regarding heterogeneous and dynamic single neuron responses in regions such as PFC.

However, what guides the characterization and interpretation of population representations is to figure out whether it can tell us something more about how a given feature or task etc. is being processed to ultimately guide behaviour. Going back to our

example in Fig. 1.1, in order to generate appropriate behaviour e.g. reach for stimulus x, an individual or set of downstream neuron(s) has to read out stimulus location from the conjoint activity of the three presented neurons (in fact, only two neurons would already be sufficient to represent the task). A conservative but widely held assumption is that the brain implements linear read-out mechanisms in the form of weighted sum operations akin to a linear discriminator, which could be visualized as a separating hyperplane in Fig. 1.1b. (Fusi et al., 2016). One can easily find such a plane that distinguishes each circle (stimulus) from all other circles, implying that the presented representation can be exploited to solve the task. However, the geometry of the representation allows for the implementation of diverse behaviours. Similar stimuli are grouped together in state space, which shows that similarity between the actual stimuli is reflected in similarity in the population responses i.e. a smaller distance between stimulus specific representations. If the task instructions were to respond differently to left stimuli or right stimuli an appropriate read-out could be easily implemented as well (Fig. 1.1c, red lines), the same goes for up versus down and so on (Fig. 1.1c). Fig. 1.1d shows a similar representation that was evoked in the PFC of monkeys trained to categorize motion stimuli (Swaminathan and Freedman, 2012; Chaisangmongkon et al., 2017), which is reflected in the bias of trajectories towards the categorical prototypes.

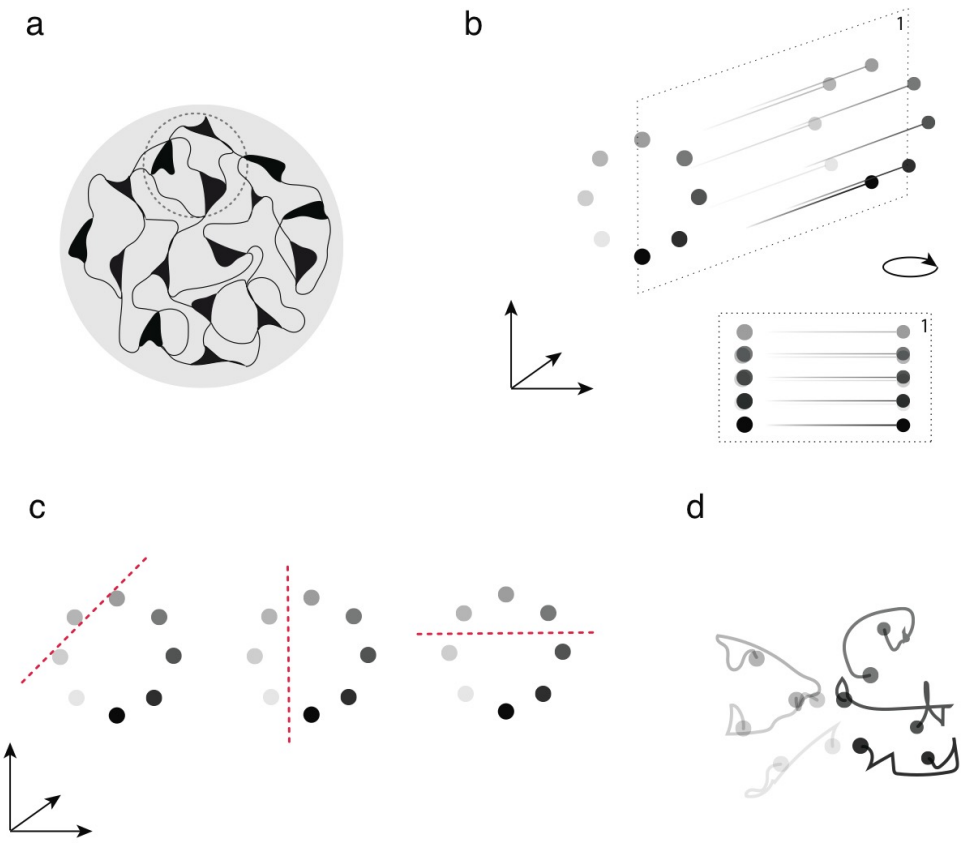


Figure 1.1: Neural state space representations, trajectories and subspaces. a) Sketch of a hypothetical neural population. Dashed ring: assembly or sub-population consisting of three neurons. **b)** A state space representation: Activity of three neurons in response to six different experimental conditions (e.g. stimulus locations). Trajectories over time are depicted by graded lines. The inset shows the same but within a subspace of the full state space. In this case, the subspace is spanned by two of the three neurons. **c)** Example of different implementations of linear read-outs (red dashed lines), depending on task demand; left: respond to each stimulus with a distinct reach; middle: only respond to stimuli on the left; right: only respond to upper stimuli. **d)** Population representation/trajectories during a motion direction classification task (PFC recordings). Start of the trial (dots in centre) until delay period (dots in the periphery). Panel **d** was redrawn from Chaisangmongkon et al., 2017, based on data from Swaminathan and Freedman, 2012.

1.4.4 Distributed codes – Emergent properties

In the previous paragraph, we considered the nature of a population representation, and under what terms it is usually interpreted. As mentioned before, distributed codes can display emergent properties, which cannot be predicted or understood by only looking at the individual cells. In a sense this is shown in Fig. 1.1, however, to make this more

explicit I will turn to some concrete examples where analysing the properties of population representations has helped to better understand how tasks are processed in order to guide successful behaviour.

An influential study by Rigotti et al. (2013) re-analysed PFC recordings from monkeys performing a complex series of tasks. Animals were presented with a sequence of two objects. In one task, they had to determine whether a following sequence matched the original sequence. In a second task they were presented with an array of objects that included the two objects from the first sequence and had to saccade to those objects in the same order. As expected, Rigotti et al. (2013) found neurons representing object identity but also neurons with all kinds of mixed selectivity to objects, task and position in the sequence. Neurons that are selective to mixtures of task variables are observed often in PFC but many times lack interpretation (Warden and Miller, 2010). Investigating the population representation of the tasks performed by the monkeys, Rigotti et al. (2013) found that precisely neurons with mixed selectivity - specifically non-linear mixed selectivity - were essential to create a representation that could be used to solve both tasks. This was because non-linear mixed selectivity increased the dimensionality of the neural representation, allowing for the implementation of an appropriate linear read-out to solve the two tasks embedded in a common representation. Figure 1.2a-c gives a visual explanation of how non-linear mixed selectivity increases the dimensionality of a representation. Let us take as an example a classic association task with an 'exclusive or' (XOR) structure. Such a task requires a different response to a given stimulus e.g. shape depending on a second factor e.g. its colour (Fig. 1.2a), in this case, respond right if you see a circle but only if it is black (otherwise respond left) and vice versa for the triangle stimulus (Fig. 1.2a). This means that any binary

classification of the four task stimuli into a right and left response category cannot be based on a single stimulus feature alone. This is also a problem for a hypothetical neural population consisting of only feature selective neurons. For example, in Fig. 1.2b one neuron is selective to colour, one is selective for shape and a third neuron's response is determined by a linear combination of both (linear mixed selectivity). Such a population would represent all stimuli on a plane i.e. the representation would be two dimensional. While this representation allows for a distinction between all stimuli and even a simple classification according to shape or colour, it is no suited for the task. That is, one cannot find a linear read-out i.e. plane that separates (or classifies for that matter) the grey triangle AND black circle from the black triangle AND grey circle. If instead the neural population includes neurons with non-linear mixed selectivity i.e. whose response to a stimulus is not a linear combination of their responses to the individual features, the dimensionality of the representation increases (Fig. 1.2c). Now, one can find a plane that separates the stimuli according to the required response. Note, increasing the dimensionality of the representation is akin to the 'kernel trick' in machine learning (Vapnick, 1998) and in many neural networks this is achieved by recurrent connections or hidden layer(s) that mix feature specific inputs and induce non-linear responses. The example given in Fig. 1.2 is analogous to the problem in the Rigotti et al. (2013) experiment, where the same input had to be reconfigured for two different behavioural protocols (an XOR type problem). To be fair in their case, 'pure' task specific neurons could solve the representational problem, as would a hypothetical non-linear read-out mechanism. However, all aspects of the task, including task identity could be decoded after removing the pure selectivity component from the neuronal population, in fact, even better than when only focusing on the pure selectivity component (see previous paragraph on why distributed codes do not rely on pure selectivities). Furthermore, the

dimensionality of the representation was higher than would be expected from pure selectivity alone and collapsed when animals made errors (Rigotti et al., 2013). Together, their results make a strong case for a highly distributed code in PFC, which contains information that is not contained in pure selectivity to task features. Furthermore, certain mixings of information can radically change the dimensionality of a neural representation, which can be leveraged to generate arbitrary and flexible input-output mappings, which speaks to the role of PFC in cognitive flexibility and control.

Oftentimes it is necessary to generalize over commonalities of various features or task variables in order to create an invariant, abstract representation. As we have seen this is the case for many prefrontal neurons that display categorical signals (see the *Abstract neurons* section). In a recent study, Bernardi et al. (2019) characterized population representations in PFC during a task that required abstract processing on various levels and revealed coding mechanisms not apparent at the level of the single neuron. A schematic of their task and main results are depicted in Fig. 1.2 d-f. Monkeys performed a serial reversal learning task that required them to switch back and forth between two contexts. Importantly, a change in context was not explicitly signalled, but had to be inferred by the animals through trial and error. Under each context, certain object response pairs were rewarded e.g. if a square was presented during the red context (note, 'red' just for visualization) a button release was rewarded with juice (+) whereas keeping the button down resulted in a time out and trial repetition. If instead a triangle was presented, monkeys had to keep pressing the button to directly proceed to the next trials (-), releasing the button would result in timeout and trial repetition (see Fig. 1.2d for all trial types). Bernardi et al. (2019) focused their analysis on the epoch before a new stimulus was presented to characterize the latent population representation that might

guide the subsequent stimulus processing. All task variables (context, action, value) were clearly distinguishable in the population representation (Fig. 1.2e) i.e. could be read-out by linear classifiers. Importantly, the geometry of the representation was such that all variables were represented simultaneously in an abstract format: objects formed parallel planes for instantiations of each variable. This is depicted in Fig. 1.2e&f where the plane spanned by all objects under the red context is parallel to the plane spanned by all objects under the grey context. There is a similar relationship for the action and reward variables (Fig. 1.2f, only one plane per variable is depicted for simplicity). The format of the variable representations can be described as abstract because in each case it generalizes over all objects and other variables, however, information about variables remains preserved (as opposed to a collapse along any of the depicted dimensions). In other words, a linear read-out for context, defined on the (+) conditions would generalize to the (-) conditions i.e. distinguishes between contexts there just as well. This also holds true for the reverse, and also for the action variables. Such a representation could be useful for the animals performing the task. For example, when a new stimulus appears, an existing representation of the described format enables monkeys to rapidly read-out which action to perform but also to switch behaviour if the context has changed (Bernardi et al., 2019). Importantly, while individual neurons with abstract responses (e.g. context selectivity) might be an indicator of the described representation, the format or geometry of the representation is a property that emerges at the population level. Furthermore, as in the previous examples, any rotation of the representation in state space would lead to less interpretable mixtures of neural responses (in comparison to those used in the examples), without affecting the geometrical properties of the representation per se. In fact, Bernardi et al. (2019) observed only few neurons that showed pure context, action or reward selectivity. As opposed to the example which

made the case for non-linear mixed selectivity in PFC, the format of the representation in the current example would be distorted by non-linear mixed selectivity (see arrow in Fig. 1.2f) and actually requires less than the maximal dimensionality, for the benefit of generalization and at the cost of arbitrary linear read-outs.

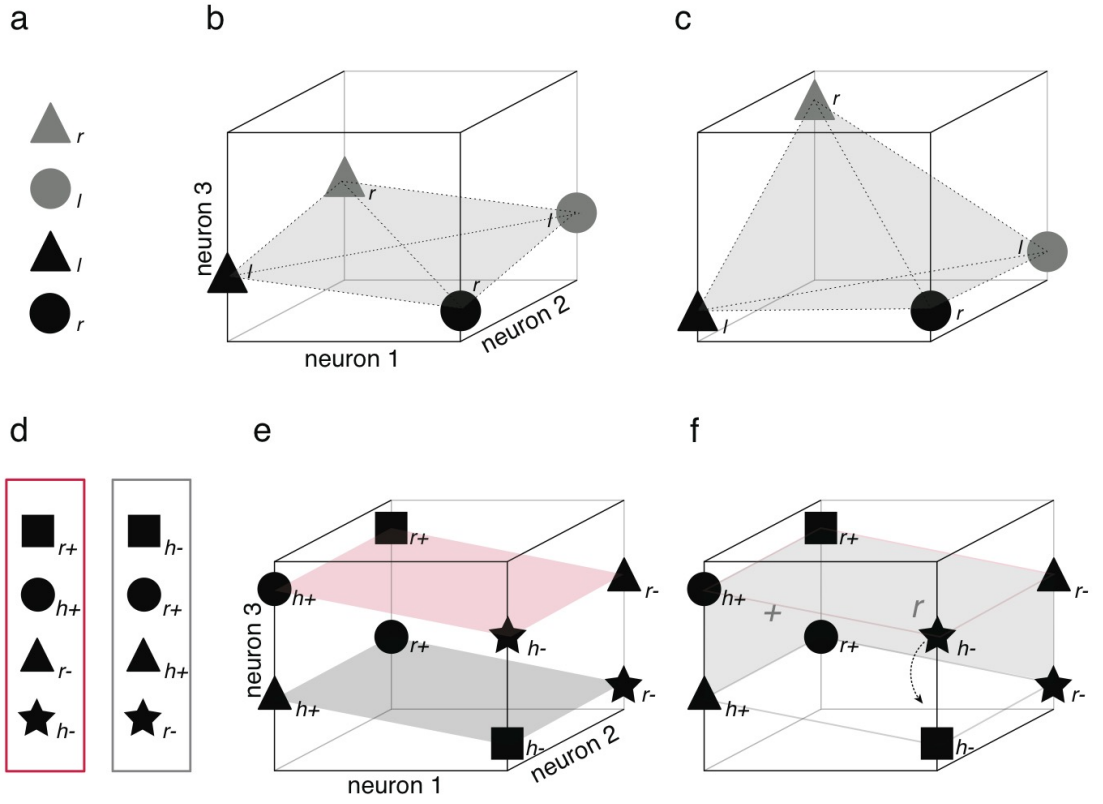


Figure 1.2: Examples of task representation in PFC. Panels in **b** and **c** were redrawn from Fusi et al., 2016; Panels in **d**,**e** and **f** were redrawn from Bernardi et al., 2019. **a)** Stimuli and associated responses (left; right) for the XOR task. **b)** Representation in the state space spanned by three hypothetical neurons. Neuron 1 responds maximally to circular shapes (pure shape selectivity). Neuron 2 prefers grey colour (pure colour selectivity). Neuron 3 prefers triangles and grey color (linear mixed selectivity). The shaded area indates that all stimuli lie on a two dimensional plance. No plane defined in this state space could separate the stimuli associated with the to r from the stimuli associated to the two l. **c)** This time, neuron 3 exhibits non-linear mixed selectivity increasing the dimensionality of the representation from two to three, which is indicated by the shaded area forming a tetrahedon. Now, a plane could be found to separate stimuli according to response. **d)** All possible context-object-response-reward combinations i.e. trial types from the experiment in Bernardi et al. (2019). Red and grey boxed indicate respective contexts. Letters indicate actions; r: release button, h: hold button. Symbols indicate reward type for correctly performed trials: +: juice, -: no time-out, no trial repetition. **e)** Representation in state space spanned by three neurons. Neuron 1 responds to reward type (+,-), neuron 2 responds to action (r,h) and neuron 3 is sensitive to both context and reward type. Note, neuronal selectivities could be very different and still produce a similar representation (imagine an arbitrary rotation of the entire representation). Red and grey shaded areas indicate (abstract) planes corresponing to each context. **f)** The same representation but now shaded areas indicate planes corresponding to the + reward type and the *r* action. One can see that planes are parallel to their respective counterparts. The dashed arrow gives an example of how non-linear mixed selectivity could affect the format of the representation.

1.4.4 Distributed codes – Dynamic coding and subspaces

So far, I have focused on the representational properties assuming complete stationarity of the distributed code. In the following I will give a short overview of how temporal information can be represented at the level of the neural population, for this, I will mainly draw on examples outside PFC. However, a more thorough introduction on the importance of temporal dynamics in PFC will be given in the section on working memory (*Maintenance of information*). Considering distributed codes over time involves analyzing trajectories in state space (see Fig. 1.1b). As we have seen before, any point that forms part of a trajectory is defined by the population activity vector at a certain time point. Thus, the pattern of activity over time is directly reflected in the dynamics (shape and speed) of the trajectory in state space. Insight into the temporal dynamics of trajectories can be gained by interpreting them as states in a dynamical system. Note, a state in a dynamical system means a position in state space, just as we have defined it before. The two most important definitions we need from the dynamical systems theory for the remainder of this chapter is that of an attracting state and that of a transient state. An attracting state (or simply attractor) is a state to which the system tends to converge over time. A good analogy is that of a bead in a bowl. No matter where you position the bead in the bowl, it will end up at the bottom. This mean that when in an attractor state, neural activity remains constant i.e. the trajectory does not move (in the simplest case). The most general definition of a transient state is simply a state through which the trajectory moves relatively quickly, typically towards an attracting state (in case there is one close by).

In the following I will focus on the advantage and implications of transient states for dynamic coding. A particularly prominent example for dynamic coding is the olfactory information processing in the locust (Laurent, 2002). Specifically, in response to odors projection neurons in the locust antennal lobes exhibit complex time-varying activity, even when stimulus input remains constant. Because of the robust temporal dynamics in the individual cells, processing of two odors is reflected in two very distinct but reproducible trajectories in state space. Activity from the projection neurons is read out by a different population of neurons, called Kenyon cells. Although information about the two odors could be read out while the system is in an attractor state i.e. no change in trajectories over time, Kenyon cells are most active while the trajectories are in 'motion' i.e. passing through transient states, which coincides with the timing the locust would need to choose between the two odors (Mazor and Laurent, 2005). These results from the locust olfactory system provide compelling evidence for the computational role of the dynamic code that can be extracted at the level of the population (see Buonomano and Maass, 2009, for a more extensive review of transient coding schemes).

Dynamical coding is particularly prominent in motor cortex and associated regions, where single neurons exhibit a great variety of multiphasic patterns (Churchland et al., 2012). On the level of the population, this is translated in trajectories that typically show output specific (e.g. reach directions) rotational dynamics, which contribute to the generation of a motor command through implementing a coordinate transformation between desired movement properties and muscle activity (Churchland et al., 2012; Kalaska, 2019). Dynamic codes are also important in generating timed responses. Remington et al. (2018) recorded from medial frontal cortex while monkeys had to produce a timed response based on their estimate of a preceding time interval. The

researchers rigorously analyzed the geometry and dynamics of the population level trajectories during the response production interval and found that trajectories underwent scaling to ensure correctly timed responses. That is, the trajectory for a late response was simply a slower version of the trajectory for a quick response, whereas the geometries between trajectories were otherwise preserved. Importantly, a subsequent study by the same group showed that the curvature of the neural trajectories, an inherently dynamic property, was a direct implementation of a bias in the time estimate of the monkeys, which was evident in the monkeys' behavior (Sohn et al., 2019).

Dynamic codes rely on transient i.e. time varying states, which means that the trajectory visits different parts of state space at different times. If those states all lie in the same direction in state space, the trajectory will be one dimensional. The number of directions a trajectory explores add to the dimensionality of the representation as a whole. This illustrated in Fig. 1.1b: The dimensionality of the representation without taking into account time would be two. Adding the dimensionality of the trajectories (i.e. the dimensionality of time) makes it three. Thus, similar to non-linear mixed selectivity, a dynamic code can increase the dimensionality of the representation. A corollary of this is that instantaneous representations can reside in entirely different regions i.e. subspaces of state space (see the *Distributed codes – Interpreting representations* section, which gives a definition of a subspace). This property of dynamic codes is extremely useful. Take for example the case of delayed reaching or saccade tasks where monkeys typically have to wait for some time before they initiate their movements. This period of waiting is used by the monkeys to prepare their responses (preparatory epoch). Interestingly, neural activity in motor cortex encodes target positions from when they are first seen, throughout the preparatory epoch and during movement execution. A

long-standing question had been how the very similar activities during preparatory and movement epochs elicited movement in one case, but not the other. Part of the answer lay in the fact that population activity while in certain subspaces (output-potent) can generate muscle activity whereas in others it cannot (output-null) (Kaufman et al., 2014). During the preparatory epoch population activity resides in an output-null subspace but after the GO cue transitions to an output potent subspace where it leads to the generation of movements. In fact, preparatory epoch and movements epoch subspaces are essentially orthogonal to each other (Elsayed et al., 2016), which means that the angle between them approximates 90 degrees (see Fig. 1.3a for a toy visualization). The big advantage of such a coding scheme is relatively simplistic: Separate subspaces allow for unique i.e. time specific read-out while minimizing interference between time specific computations (preparing versus executing a movement). The idea of dynamic coding and distinct subspaces for specific computations is not limited to motor cortex but is increasingly applied to understand computations in PFC during cognitive tasks. An important study by Mante et al. (2013) analyzed population level dynamics during a context dependent decision-making task, where monkeys had to integrate evidence and decide on either the color or the motion direction of colored moving dots. They found that trajectories for color and motion direction evolved along two distinct subspaces but fell onto a common subspace that clustered trajectories according to the decision. Crucially, integration along the decision subspace was selective to the currently relevant variable (color or motion direction) (see Fig. 1.3b). The study by Mante et al. (2013) elegantly shows the presence of separate subspaces for distinct processes in the same population of PFC neurons. Two orthogonal subspaces for color and motion direction encoding and an orthogonal decision subspace, that integrates evidence towards a

choice. Additionally, it shows how these subspaces are formed and interact through the dynamics of the population activity (see Mante et al., (2013) for further details).

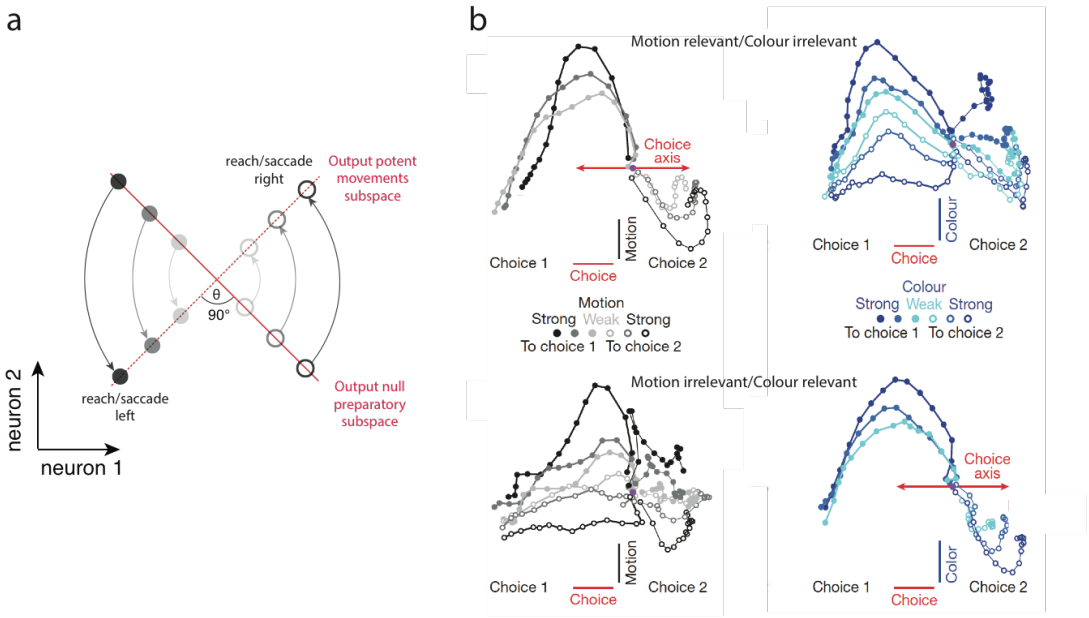


Figure 1.3: Dynamic trajectories and coding subspaces. a) Task dependent activity in two orthogonal subspaces (hypothetical sketch). Circles represent six hypothetical task conditions e.g. target locations to which the monkey has to saccade or reach. Shadings indicate the left/right eccentricities of the targets. Filled circles represent targets on the left, empty circles represent targets on the right. Solid red line represents the preparatory subspace i.e. the axis (or set of axis) in state space that explains most of the variance between task conditions during the preparatory epoch. The dashed line represents the movement epoch subspace. During the task, activity transitions from the preparatory subspace to the movement subspace (direction of trajectories indicated with arrows) as only the latter aligns with output potent dimensions that elicit movement. In other words, the movements read-out is aligned with - or trained to classify target related activity within - the movement subspace. Importantly, preparatory and movement subspaces are orthogonal to each other (θ of 90 degrees), which means that any variability in target related activity in the preparatory subspace does not affect the movement subspace. Note, mathematically, the higher the dimensionality of state space, the higher the probability that 2 randomly picked subspaces are orthogonal to each other. Conversely, the higher the dimensionality of the randomly picked subspaces (number of axis that define them), the lower the probability that they are orthogonal. **b)** Condition specific trajectories within the motion, color and choice subspaces (here denoted as axis) in the Mante et al. (2013) experiment. Trajectories first spread out along the motion and color subspaces. Trajectories for the currently relevant feature (motion upper left; color lower right) rotate towards the choice axis as evidence towards a specific choice is being integrated, and spread out along that axis accordingly until the decision is made. Note, clustering of feature values according to choice. For the currently irrelevant feature, there is no clustering according to feature values and choice along the choice axis (upper right for color; lower left for motion). Note, color and motion axis were orthogonal to each other (not depicted). Graphic adapted from Mante et al. (2013).

So far, I have focused on distinct subspaces within the same task. These included subspaces that are specific to certain task variables, as in the last example (also see Machens et al., 2010 and Keemink and Machens, 2019 for a review) as well as subspaces brought about by dynamic coding. A natural question that arises is how subspaces might change as a function of the task that has to be preformed. Preliminary results from motor cortex are somewhat ambiguous in that respect. A recent study by Gallego et al. (2018) showed that, at least superficially, subspaces defined on many individual tasks (carried out by the same monkeys) largely overlapped. However, their results are somewhat more complicated. They partitioned the general task subspaces (also called manifolds) into smaller subspaces that were specific to certain parameters such as time, target locations and so on. Most of those subspaces were shared across tasks, possibly reflecting similarities in timing and onset of events but also target locations. However, some subspaces were task specific (they only explained variance among features of certain tasks). The authors concluded that while subspaces for specific tasks may largely overlap, a certain amount of task specificity is always preserved. A more direct conclusion can be drawn from a study by Miri et al. (2017), showing that in mice, activities in a 'pull' task and 'walk' task were confined to two orthogonal subspaces. However, few studies have compared neural population dynamics and coding subspaces in cases where animals performed more than one task and the specificity, flexibility and dimensionality of task related subspaces is currently a much-debated topic.

In summary, the format i.e. geometry and dynamics of neural representations can give insight into how a specific task is being processed in a given network to generate appropriate behaviour. To contrast the prevalent view of the single neuron doctrine in Neuroscience, held during most of the 20th century, nowadays populations of neurons

are put forth as the computational units of cognition (Saxena and Cunningham., 2019). To cite Chaudhuri et al. (2019): '…*Information representation, computation and dynamics unfold at the level of the neural population and the collective states across neurons of a circuit are the natural way to understand this process*…'. Under this view, the activity of each neuron is not much more than a local window on the computational process that unfolds on the level of the population. However, a population representation critically depends on the computations that can be performed by the neurons it draws upon, no matter if as a result of network architecture or cell inherent properties. For example, high dimensional representations rely on neurons with non-linear mixed or time varying selectivity. The types of neuron, or better, tuning properties that are encountered in a certain brain region can give clear hints on how this region might represent information on the level of the population. Given the diversity of tuning profiles observed among prefrontal neurons (see section *Neuronal responses in PFC*) it thus follows that PFC is endowed with an enormous capacity to represent information on the level of the population. While population representations in PFC are increasingly being explored (Rigotti et al., 2013; Bernardi et al., 2019), it is still largely unclear how they are leveraged to sub-serve higher cognitive function.

1.5 Maintenance of information

1.5.1 Working memory and persistent activity

As we have seen at the very beginning of this chapter, a cornerstone for any system exerting cognitive control is the ability to actively and robustly maintain information until a goal is achieved (Miller and Cohen, 2001). The cognitive function that is most

strongly associated to this requirement is called working memory (WM). The sensory inputs we act upon are not constant and our behavior is not instantaneous, which requires us to store information for future use. Working memory refers to the capacity to keep in mind but also manipulate information over short periods of time. Conceptually, WM can be divided into two major processes: One involves the maintenance of information while the other ensures that information suits the current task demands (Baddeley et al., 1986). Most of this chapter has concerned the latter process i.e. how information is represented in task-oriented formats. Although this was not done explicitly in the context of WM, many of the tasks that I presented throughout were essentially WM tasks or contained an explicit WM component. Take for example the task in the Asaad et al., (1998) study discussed before, where monkeys had to associate objects with the direction of a saccade. The saccade had to be made after the object was no longer present. This means that the monkeys somehow maintained that information over a delay. In this section, I introduce and discuss the process of maintenance of information in PFC, specifically under the framework of WM.

In addition to cognitive control in general (see *Lesions to the PFC* section), lPFC has specifically been implicated in the maintenance processes pertaining to WM. Studies with prefrontal patients and lesion studies in non-human primates reported behavioral deficits during delayed response tasks (Jacobsen, 1935), and general problems in maintaining relevant information (Owen et al., 1990; Manes et al., 2002). These earlier observations were complemented by studies inflicting more precise lesions to the macaque dlPFC, which led to considerable impairments in delayed saccade tasks (Funahashi et al., 1993), and in remembering hidden food locations (Butters and Pandya, 1969; Goldman et al., 1971; Passingham, 1985). Evidence form those lesion studies

motivated some of the first electrophysiological recordings in dlPFC. Two studies that appeared in 1971 independently reported prefrontal neurons that persistently fired action potentials during a delay between stimulus presentation and action execution (Fuster and Alexander, 1971; Kubota and Niki, 1971). Subsequent studies repeated this observation while interpreting the persistent and stimulus specific delay period activity as the neural correlate of WM (Goldman-Rakic, 1987, 1995; Goldman-Rakic et al., 1992). Persistent neural firing during delayed response tasks (and variants thereof) has since been observed in many brain regions including posterior parietal cortex (PPC) (Constantinidis & Steinmetz, 1996; Gnadt & Andersen, 1988; Quintana & Fuster, 1992); PFC (Boch & Goldberg, 1989; Bruce & Goldberg, 1985; Funahashi, Bruce, & Goldman-Rakic, 1989, 1990, 1991); inferotemporal cortex (IT) (Fuster & Jervey, 1981; Naya et al., 1996); superior colliculus (Mays and Sparks, 1980) and even visual areas (Miller et al., 1993; Miller and Desimone, 1994; Chelazzi et al., 1998, 2001), specifically MT (Bisley et al., 2004), V3 (Nakamura & Colby, 2000) and V1 (Supèr, 2001); to give a few examples. Persistent activity seems to be a common signature of mnemonic computations during WM, which is by no means restricted to the PFC (Dotson et al., 2018) and has been reported for many species. Recent studies even observed persistent mnemonic activity in single neurons in various brain regions in humans (Kornblith et al., 2017; Kaminski et al., 2017), including in human hippocampus (Boran et al., 2019).

The ubiquity of persistent activity accounts throughout the brain has been met with a large effort to create theoretical models that can explain it. While some research point towards cell autonomous mechanisms leading to persistent firing (Egorov et al., 2002; Wang et al., 2015), most theories relate persistent neural activity to circuit mechanisms. A particularly successful approach consists in creating neural networks that favour

attractor states (as defined in *Distributed codes - Dynamic codes and subspaces* section) in which activity remains stable over time (Amit and Brunel, 1997; Hopfield, 1982; Compte et al., 2000). Typically, the prerequisite for achieving attractor states relies on specific architectures such as strong recurrent connectivity and fine tuning of the balance between excitation and inhibition in the network. This ensures that activity reverberates in the network without 'blowing up' or 'dying away', resulting in persistent spiking in single neurons over certain time periods after the input is removed. Persistent delay activity observed experimentally could be explained by attractor dynamics acting on the level of the population (Wimmer et al., 2014; Inagaki et al., 2019). Indeed, a recent study in mice elegantly related attractor mechanisms to single-cell persistent delay spiking, while ruling out cell autonomous activity as an alternative explanation (Inagaki et al., 2019). Apart from the described observations in single neurons attractor dynamics are also in accordance with population representations, which can be remarkably stable during WM delay periods in monkey PFC (Meyers et al., 2008; Spaak et al., 2017; Cueva et al., 2019), PPC (Sarma et al., 2016) or human medio-temporal cortex (Kaminski et al., 2017). There are several ways to test this on the level of the population. If, for example in the simplest case, the activity of all neurons in the population is relatively stable over time, trajectories in state space should display little to no dynamics but instead form points corresponding to the experimental conditions, or for that matter, discrete attractor states. Another method, which is used extensively in this thesis, relies on simply correlating the population activity vector with itself over time. In case of a stable representation, the correlation should be high at all time points.

1.5.1 Dynamic codes during working memory

As we have seen earlier in this chapter, prefrontal activity can be highly dynamic, both in the spatial as well as in the temporal domain. While those dynamics increase the computational power of PFC, they do come at a cost for stability over time. Activity during simple paradigms such as the classic delayed saccade task is usually quite stable during the delay period but is dominated by strong dynamics that guide transitions from encoding to maintenance or maintenance to saccade (Spaak et al., 2017). During more complex tasks, dynamics can even dominate the delay period (Barak et al., 2010; Brody et al., 2003; Shafi et al., 2007; Zaksas and Pasternak, 2006), for example, if additional stimuli are introduced or task demands increase (Stokes et al., 2013; Spaak et al., 2017; Parthasarathy et al., 2017; Chaisangmongkon et al., 2017). In many of these experiments, the proportion of 'purely' tuned cells, with stable activity over time was relatively small, which has led researchers to speculate that attractor dynamics may not be the dominant process underlying WM. In their simplest form, dynamic population codes are driven by cells that synchronously scale their responses e.g. ramping activity, such that trajectories remain relatively low dimensional (Cueva et al., 2019). However, in many cases cells independently switch their selectivity or become selective only at certain time-points in the trial (Spaak et al., 2017), resulting in high dimensional dynamics. Those high dimensional dynamics observed during WM directly speak to the dynamic or transient coding mechanisms that I have introduced earlier (see *Distributed codes – Dynamic coding and subspaces* section). 'Reservoir' or 'echo-state' networks (Jaeger and Haas, 2004; Maass et al., 2002) are a class of networks that promotes particularly high dimensional dynamics and representations. In contrast to attractor networks, connectivity in reservoirs is almost random leading to high proportions of mixed selective units with transient activity, which endow such networks with a high representational capacity. Reservoirs have been popular in modelling activity in PFC

because they capture the complex tuning profiles and dynamics observed in prefrontal neurons during WM (Enel et al., 2016). In such networks, information maintenance does not rely on persistently active units but rather on the dynamics unfolding at the level of the population. In a sense, artificial responses produced by attractor networks and reservoirs stand at the extremes of the spectrum of what is observed in PFC. However, theoretically as well as practically, dynamic codes can coexist with stable representations. To explain this, we have to go back to the idea that coding can be restricted to distinct subspaces in state space. A study by Murray et al. (2017) comparing delayed saccade and tactile WM tasks showed that high dimensional state space contained two distinct coding subspaces. One that captured most of the transient cue-encoding related activity and a second mnemonic subspace. Activity in this second subspace arose during the cue period and remained within that subspace until the end of the delay period. Importantly, the format of the representation within the mnemonic subspace was stable in that temporal dynamics were orthogonal to the axis that explained stimulus related variance. This is directly comparable to the sketch in Fig. 1.1b showing trajectories that are obviously dynamic since they change over time, however, any read-out trained to distinguish between stimuli would cross generalize perfectly across all time points (this is different from a static representation in attractor networks). This shows that in spite of heterogeneous and time varying tuning in single neurons, activity can be maintained stably on the level of the population. Furthermore, this does not preclude dynamic coding, but simply asserts that a stable subspace can exist in parallel. Those results are also supported through network models (Murray et al., 2017; Druckmann and Chklovskii, 2012) and were again confirmed by a recent modelling study combining a sensory layer including strong attractor dynamics with a reservoir-like layer (Bourchacourt and Buschman, 2019). Some additional ways to model WM

dynamics are particularly worth noting. Those include feedforward chain like models, where delays are bridged through sequential activations of units (Ganguli and Latham, 2009; Goldman, 2009). This type of 'tiling' dynamics has been observed in rodent frontal cortex during WM tasks (Baeg et al., 2003; Pastalkova et al., 2008; Runyan et al., 2017). Other models include rapid synaptic weight changes (Mongillo et al., 2008; Masse et al., 2019; Bourchacourt et al., 2019; Manohar et al., 2019) that could serve as yet another mechanism of maintaining information 'silently' across delays (Stokes, 2015).

Regardless of the precise underlying mechanisms, WM processes in PFC can entail dynamic as well as stable components. According to several theories that emerged over the last 20 years, PFC is viewed as being more important for the control of WM (see e.g. Curtis and d'Esposito, 2003; Postle, 2006 for a more detailed discussion), rather than pure storage. In a sense, this is consistent with the proposed role of PFC in cognitive control in general, which we have discussed at the beginning of this chapter, with WM just being another 'sub-function' within this general framework (Miller and Cohen, 2001). In PFC, transitions between task epochs are often dominated by rich dynamics that underlie transitions from sensory to mnemonic or from mnemonic to motor preparation subspaces, while activity within the individual subspaces often remains stable (Spaak et al., 2017; Parthsarathy et al., 2017; Parthasarathy et al., 2019; Tang et al., 2019). This suggests that PFC juggles between dynamical regimes to accommodate task demands, instead of simply maintaining one type of information. While mnemonic coding has been shown for a huge diversity of information in PFC (please refer to section: *Neuronal responses in PFC*), it is particularly pronounced for task relevant information, specifically if that information is somehow abstracted e.g. rules, categories, strategies and so on. Furthermore, mnemonic coding in PFC can be particularly robust

to distraction (Miller et al., 1996; Suzuki and Gottlieb, 2013), incorporate more than one item (Warden and Miller, 2010; Buschman et al., 2011) and be biased towards items in the focus of attention (Lebedev et al., 2004; Lara and Wallis, 2014). This versatility probably requires some form of switching between dynamical regimes, which potentially even act in parallel. It might be that specific information is held online in attractor like states e.g. rules and contexts or task states (Durstewitz et al., 2010; Balaguer-Ballester et al., 2011; Mante et al., 2013) while other information might be more dynamically integrated or discarded dependent on current needs.

1.6 Synopsis

I wish to present the three experimental chapters of this thesis in reverse chronological order. That is, I will start with the chapter that has been completed last and will end with the chapter that has been completed first. While this might seem slightly unconventional there is a good reason. Mainly, what is now **chapter 2** is the most central to the theoretical topic of this thesis and directly speaks to many points that have been raised in the general introduction (i.e. **chapter 1**). It is also the largest and densest of the three experimental chapters. **Chapter 3** is conceptually related to **chapter 2** and can be easily understood through it. **Chapter 4** concerned a separate analysis project, which is not covered by the conceptual framework set up during the General introduction and is not as directly related to the other experimental chapters. However, it extends some of the analysis tools that I used in **chapter 3**, which are also of importance during the general discussion (i.e. **chapter 5**). All three experimental chapters were concerned with the analysis of neural recordings from non-human primates. Two of the datasets pertaining to the analysis in **chapters 3 & 4** had been published previously, while the data from

chapter 2 has not been published before. In the general discussion, I also describe some other incomplete projects, and how I would follow them up.

Chapter 2. *Dynamical multiplexing of information during dual-task performance in monkey prefrontal cortex*

In this chapter, I describe the results of a study where we analyzed prefrontal recordings from monkeys performing a challenging WM task containing various components. A large part of the dataset had been collected by our collaborator Kei Watanabe in Osaka. The project also entailed various trips to Osaka to assist in data collection. Our main question in this study was how the prefrontal cortex processes and maintains information about two simultaneously performed tasks. We found that information about items from the two tasks was intermixed in single neurons. However, on the level of the population task representations were contained in two distinct subspaces, reducing interference during simultaneous performance. This chapter makes a contribution towards understanding how the representational capacity of PFC can be leveraged to process multiple streams of information simultaneously.

Chapter 3. *Intrinsic neuronal dynamics predict distinct functional roles during working memory*

In this chapter, I describe a study in which we analyzed recordings from three brain regions of the macaque cortex: lPFC; frontal eye fields (FEF) and lateral intraparietal area (LIP). We were interested in the spatial and temporal dynamics of information maintenance during WM in each of these areas. We found that the ability of cells to maintain information in a stable format was predicted by their intrinsic temporal profiles. The intrinsic temporal profiles also predicted a cell's capacity to encode

information dynamically. This chapter makes a contribution towards the debate on how information is represented during WM in multiple brain regions, highlighting the inherent capacity of cells to contribute to either stable or dynamic coding regimes.

Chapter 4. *Interneuronal correlations at longer timescales predict decision signals for bistable structure-from-motion perception.*

In this chapter I take a bit of a detour, and describe a study which was carried out under the supervision of Kristine Krug. We analyzed neural recordings from visual cortex, specifically area V5/MT. In this study we wanted to understand how interneuronal correlations influence decision signals during the perception of an ambiguous stimulus. We found that interneuronal correlations evoked by the ambiguous stimulus were higher than those evoked by a range of other stimuli. Furthermore, decision signals about the inherently ambiguous stimulus were predicted by correlations at large timescales. This chapter shows that decision signals in single neurons are linked to correlations between neurons. Furthermore, it provides moderate evidence of a signature of top-down processing on sensory cortex.

Chapter 5. General Discussion

In this chapter I will summarize and discuss the results from the three experimental chapters. I will specifically focus on some of the open questions that are provoked by each chapter. Finally, I will give an overview of future directions that arose from the work presented in this thesis including some preliminary findings and ideas.

Chapter 2. Dynamical multiplexing of information during dual task performance in monkey prefrontal cortex

2.1 Abstract

Cognitive multitasking is limited by interference between component tasks. The lateral prefrontal cortex (lPFC) processes information in a dynamic, context dependent manner, possibly implementing neural computations underlying the concurrent performance of multiple tasks. We recorded from neurons in lPFC while monkeys performed a dual task comprised of a spatial attention component task and a spatial working memory (WM) component task. Monkeys learned to perform the dual task with a high success rate and over various levels of experimentally induced inter-task interference. As expected, neuronal activity was tuned to unique task features as well as to complex mixtures of features from both component tasks and time. Within this response heterogeneity, population activity transitioned from a common coding scheme for spatial location during stimulus processing to orthogonal subspaces for coding the separate component tasks. Coding for each subtask also generalised across variation in the other concurrent task, consistent with a stable low-dimensional readout of the behaviourally relevant task variables. Together, our data suggest that lPFC multiplexes information across tasks as well as time and minimizes interference across task representations while ensuring a stable representation within tasks.

2.2 Introduction

Both humans and animals have the remarkable ability to perform multiple tasks in close spatial and temporal proximity; e.g., you can make a sandwich while talking on the phone (about how to make a different kind of sandwich). Such multitasking requires the ability to store and manipulate different pieces of information for distinct goals, while avoiding interference between component tasks. A major challenge in this process is the optimal allocation of shared and limited processing resources (Kahneman, 1973; Duncan, 1980; Wickens, 1980, 2002; Baddeley et al, 1992), specifically if input or output dimension overlap (Nijborg et al., 2014; Alvarez and Cavanagh, 2015).

The central region attributed to cognitive multi-tasking is the lateral prefrontal cortex (lPFC) (Shallice and Burgess, 1991, D'Esposito et al., 1995; Passingham, 1996; Baddeley et al., 1997, Klingberg & Roland, 1997; Miller and Cohen, 2001; Watanabe and Funahashi, 2014). Neuroimaging studies have repeatedly shown that lPFC activity increases during performance of multiple tasks as opposed to when the individual tasks are performed on their own (D'Esposito et al., 1995, Koechlin et al., 1999). Furthermore, patients with frontal lesions exhibit performance deficits in multi-task conditions, but not in the component single task conditions (Baddeley et al., 1997). In accordance with the hypothesis that lPFC draws upon limited processing resources (i.e. the limited resource hypothesis; Kahneman et al., 1973; Wickens, 1980, 2002), human lPFC activity shows an increased spatial overlap in response to component tasks when they are performed concurrently (Klingberg & Roland, 1997; Nijborg et al., 2014). These observations are supported by neural recordings from monkeys performing a spatial attention and concurrent spatial working memory (WM) task (Watanabe and Funahashi 2011, 2014). Both tasks activated a largely overlapping population of neurons in lPFC.

Indeed, most electrophysiological recordings from animals performing two tasks concurrently (dual-task) report neurons with dual-task specific activity (Pang et al., 2001 frontal cortex in rats; Lebedev et al., 2004; Messinger et al., 2009 lPFC in monkeys). That is, a considerable proportion of neurons is tuned to linear and non-linear mixtures of features from both tasks.

Tuning to mixtures of task variables (mixed selectivity) is commonly observed in lPFC (Rao et al. 1997; Assad et al., 1998; Rigotti et al., 2013) and other higher order brain regions (Pagan et al., 2013; Meister et al., 2013) during performance of cognitive tasks. Such mixed selectivity is typically interpreted as advantageous because it increases the representational dimensionality of the neural population code (Rigotti et al., 2013) enabling flexible, context dependent input to output mapping (Barak et al., 2013; Lindsay et al., 2017). However, a high dimensional neural response may not be desirable in a dual task like scenario, e.g. when making context-invariant judgements about objects or factoring out distracting information (DiCarlo et al., 2012; Mante et al., 2013; Jakob & Nieder, 2014, Mack et al., 2018; Bernardi et al., 2019). Given the prevalence of mixed selectivity during dual tasking, our central aim in this study was to establish how task information is maintained without interference or confusion within a mixed selectivity population.

Mixed selectivity can also refer to interactions between task features and time (Parthasarathy et al., 2017; Cueva et al., 2019). More specifically, neural population activity during WM tasks shows dynamic transitions between different coding schemes across task epochs (Meyers et al., 2008; Stokes et al., 2013; Murray et al., 2017; Spaak et al., 2017; Parthasarathy et al., 2017; Wasmuht et al., 2018; Cavanagh et al., 2018).

For example, an early dynamic encoding period is typically followed by a stable delay period. Importantly, the pattern of activity during the delay period is qualitatively different from that of the encoding period because cells dynamically change their selectivity or become active at different time points (Spaak et al., 2017), effectively creating distinct coding subspaces on the level of the population (Murray et al., 2017). This form of mixed selectivity could be useful in reducing dual task interference by transferring information from an encoding pattern to an orthogonal delay pattern, thus allowing new information to be processed within the separate encoding subspace. Therefore, we specifically hypothesized a role of time-based mixed selectivity (dynamic coding) in separating out concurrently held but sequentially encoded task variables.

We predicted that the degree and type of neural overlap between distinct task variables and epochs should play a critical role in dual tasking. We recorded from large populations of neurons in monkeys' lPFC while they performed an established dual task paradigm involving a concurrent spatial attention (attention task) and spatial WM (WM task) component (Watanabe and Funahashi, 2011, 2014). We found that information about features from both tasks was present in the lPFC population at all times during the trial, with neurons displaying both linear and non-linear mixed selectivity. Consistent with previous evidence for dynamic coding, population task codes transitioned between distinct formats across encoding and delay epochs. As a result of these transitions, component task codes were restricted to orthogonal subspaces during the dual task period. That is, variability in the attention task code did not affect variability in the WM task code and vice versa, indicating a coding scheme that offered protection from interference. Additionally, attention and WM task codes were surprisingly stable over variation within the attention task (and vice versa). That is, the WM code in one trial

generalized to trials with differing attention conditions and demands and vice versa, which was however compromised by non-linear mixed selectivity.

2.3 Methods

Subjects and apparatus

We used two female macaques (monkey T, *Macaca mulatta*, 5.0 kg; monkey K, *Macaca fuscata*, 8.0 kg). During both training and neural recording sessions, we restrained the monkey's head movement non-invasively using a thermoplastic head cap (MT-APU, 3.2mm thick, CIVCO Radiotherapy, IA., USA). Visual stimuli were presented on a 17-inch TFT monitor placed 50 cm from the monkey's eyes. The lever (customized microswitch) was attached to the front wall of the chair. Eye movements were sampled at 120 Hz using an infrared eye tracking system (ETL-200, ISCAN, MA.). Eye fixation was controlled within a 6.5 ° invisible square window. TEMPO software (Reflective Computing, WA.) was used to control behavioural tasks. All experimental procedures were approved by the Animal Research Committee at the Graduate School of Frontier Biosciences, Osaka University and were in full compliance with the guidelines of the National BioResource Project "Japanese Macaques".

Behavioural paradigm and training schedule

Monkeys simultaneously performed a spatial attention (attention task) and spatial working memory task (WM task): i.e. a dual task. The two tasks depended on the same input modality (independent visual cues) but differed in their output modality, a lever

release in the attention task and a saccade in the WM task. Note, monkeys also performed each of the two tasks separately, in which case we refer to them as the single attention or the single WM task.

The single attention task: The single attention task (Fig. 2.1a, in the results) started with a lever press and fixation on a central unfilled white ring (fixation period: 1s). Subsequently, we presented two peripheral unfilled white rings (peri rings; up and down). Peri rings were presented in contralateral visual hemifield relative to the recording hemisphere (the left and the right visual hemifield in monkey T and K, respectively). After 0.6-1s the animal was cued at either one of the peri rings or the central fixation ring, which would temporarily turn red (attention cue: 0.4s). Following a variable delay (2-6.5s) the cued ring would temporarily turn red again (attention target: 0.25s) and the monkeys had to release the lever within 0.85s to receive a juice reward. However, on 40% of the trials (catch trials), during the delay prior to the attention target, one of the un-cued rings turned red for 0.25s (randomly chosen), which the monkeys had to ignore. These catch trials were included to ensure that attention was selectively allocated to the cued location (Watanabe & Funahashi, 2014).

The single WM task: The lever apparatus was removed for the single WM task. As in the single attention task, monkeys were required to fixate on a central ring (1s), after which the peri rings were presented for 2-6.5s. This was followed by presentation of the memory cue (0.25s; white square) in one of six pre-determined far-peripheral locations (see Fig. 2.1b&d). After a variable delay (1.4-6s), all three rings (central and peri rings) were replaced by placeholders at the six possible memory cue locations (memory target). The animal was required to make a saccade quickly (within 0.5s) to the placeholder

where the memory cue had been presented and keep gazing at it for 0.6s to receive a juice reward.

The dual task: On dual task trials (Fig. 2.1c), monkeys started with the attention task (initial events and timings matched those from the single tasks, except where noted otherwise). In the delay period after the attention cue (1.2-5s), the memory cue was presented. After a second delay (0.4-2.5s) the attention target was presented and the animal had to release the lever. Premature lever release or failure to release the lever resulted in abortion of the trial and time-out (~8s). Following lever release the animal had to continue fixation for another 0.5-1s until the memory target appeared and a saccade had to be executed. Approximately 40% of the dual task trials were catch trials as described for the single attention task. Onset of the catch targets were scheduled to occur within the delay either before (30% of catch trials) or after the memory cue (70% of catch trials) (at least 0.5s before or after memory cue onset). Catch targets were scheduled at least 1.8s after attention cue offset and at least 1.4s before attention target onset.

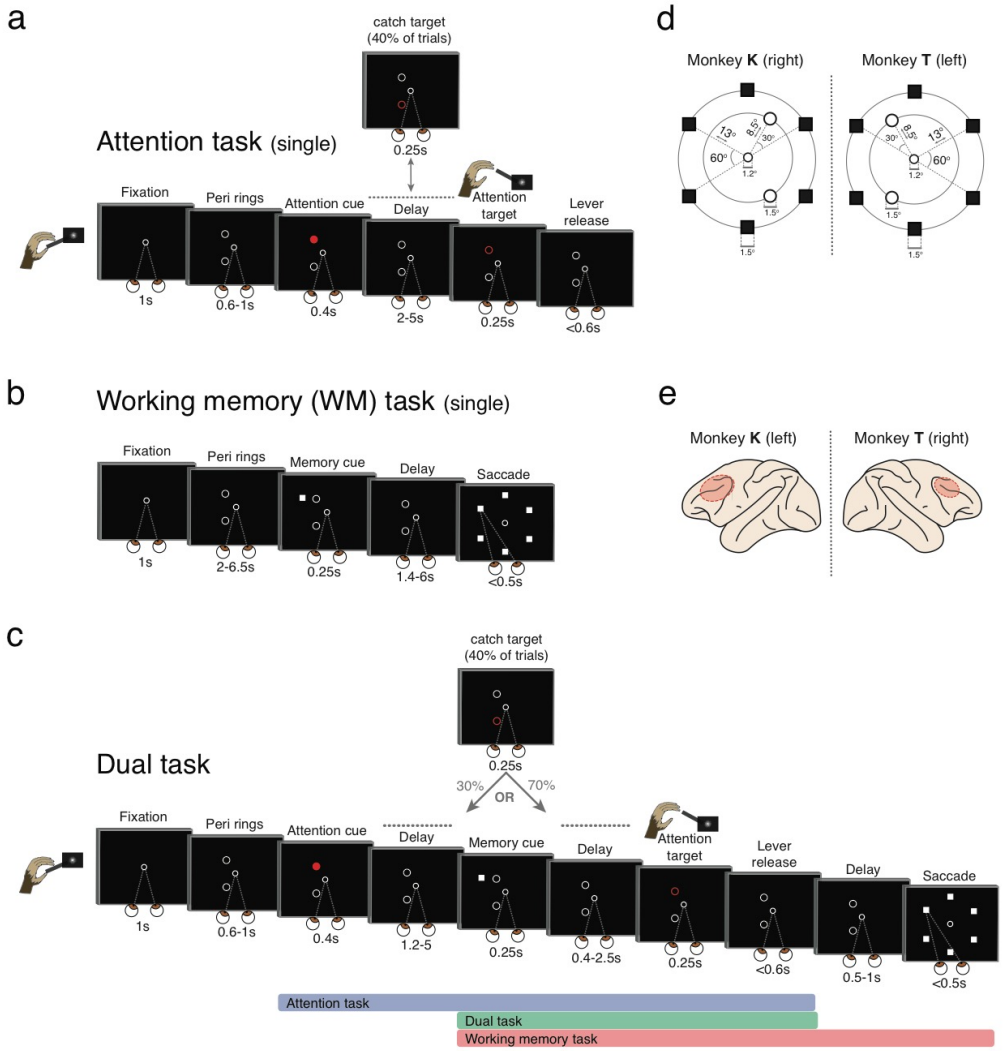


Figure 2.1: Experimental paradigms. a) The single attention task. Monkeys pressed down a lever and fixated on a central ring to initiate a trial. Two peripheral rings (peri rings) were added to the screen and the attention cue i.e. one of the three rings (peri and central) turned red. After a variable delay, monkeys were required to release the lever as soon as the cued ring turned red again (attention target). On 40% of the trials, one of the un-cued rings would turn red during the delay (catch trials). On such trials monkeys were required to maintain the lever press regardless, until the attention cue turned red. **b)** The single working memory task (single WM task). Trials were initiated by fixating on a central ring. Subsequently, two peri rings appeared which were followed by a memory cue at one out of six possible locations. After a variable delay six placeholders appeared and monkeys had to make a saccade to the cued location. **c)** The dual task. Monkeys started with the attention task component by pressing down a lever and fixating on the central ring. During the delay following the attention cue, a memory cue was presented initiating the WM task component. After a second delay the cued attention ring would turn red and the attention task component was terminated upon lever release. Following a variable post lever release delay, the WM task component was terminated by a saccade to the cued memory location. On 40% of trials, a catch target appeared in either the delay before or after the memory cue. Coloured bars indicate the component tasks and their overlap during the trial. **d)** Geometrical arrangement of the memory and attention cues and their possible locations. **e)** Sketch of approximate recording locations in lPFC.

In a typical session, monkeys first performed a block of single WM task trials (100-150 correct trials). These were followed by a block of dual task trials (200-350 correct trials). Importantly, during the dual task block a memory cue was only presented in around 70% of all trials (randomly chosen), which made those trials dual task trials. The remaining 30% were effectively single attention task trials. This was followed by another block of single WM task trials (100-150 correct trials). Note, working memory is also required during the attention task. Nonetheless, in accordance with a previous study reporting a similar task design (Watanabe and Funahashi, 2014) and to avoid confusion between the two tasks we chose to call this task attention task.

We first trained monkeys in the single attention task and then in the single WM task. Training for the dual task was only started after monkeys performed above a 90% correct rate in three consecutive sessions. For the dual task we trained monkeys until they achieved an 85% correct rate in three consecutive sessions. To ensure that the monkeys could switch readily between the single WM and dual task blocks in a single session, we further continued the training for about a month.

Recordings

After training was completed, we conducted an aseptic surgery under general anesthesia. We stereotypically implanted a plastic recording chamber on the lateral surface of the prefrontal cortex, under the guidance of structural MRI images. In monkey T, we implanted a cylindrical chamber (RC-T-S-P, internal diameter 12.7mm, Gray Matter Research, MT.) in the right hemisphere (AP = 31, ML = 15.5; AP, anterior-posterior; ML, medio-lateral). A 32-channel semi-chronic microdrive system (SC-32, Gray Matter

Research) was mounted inside this chamber. In monkey K, we implanted a cuboid chamber (width 12 mm, depth 16 mm, height, 15 mm, S-company ltd., Tokyo, Japan) in the left hemisphere. A craniotomy was made under each chamber. During task performance, both monkeys exclusively used the hand contralateral to the recording hemisphere for lever manipulation.

We collected neural data in a total of 42 sessions (16 in monkey T; 26 in monkey K). In monkey T, we used the 32-ch microdrive (SC-32) that housed 32 single-contact tungsten electrodes with inter-electrode spacing of 1.5 mm. In monkey K, we used a 32-ch linear microelectrode array (Plexon U-Probe, Plexon, TX) with an interelectrode spacing of 150 μm. The array was advanced by a custom-made hydraulic microdrive (S-company ltd.). Raw extracellular neural signals were amplified and recorded in reference to a titanium bone screw at the vertex (in monkey T) or the shaft of the linear array (monkey K) using a neural signal amplifier RZ2Bioamp Processor (Tucker-Davis Technologies, Fl.). For analysis of spiking activity, raw neural signal was filtered (300 Hz to 6kHz) for offline sorting (Offline Sorter, Plexon). In monkey T, approximately three hours before each recording session, we took the monkey to the testing room and advanced each electrode in the SC-32 by a minimum of 62.5 μm in order to ensure recording of new neurons. In monkey K, we inserted the U-probe in steps of 500 μm, while monitoring electrocardiogram (pulsatory fluctuation) on superficial electrodes to identify the point of cortical entry. Following the protocol reported in Ninomiya et al., (2015), we usually left 3 - 5 superficial channels outside the cortex. After insertion, we waited 2 - 2.5 hours until the recorded single-unit and multiunit activities indicated that the electrode was stably positioned in cortex. While waiting, the monkey watched nature and animal video clips.

To determine location of the frontal eye field (FEF), intracortical microstimulations (22 biphasic pulses, 0.2ms duration at 333 Hz, ≤150 μA) were applied through microelectrodes. When eye movements were elicited below 50 μA, the site was considered to be in the low-threshold FEF. In monkey T, our recording area did not include any sites within low-threshold FEF. Monkey K has not been tested yet.

Data preprocessing

Single unit spike trains were convolved with a Gaussian kernel (s.d. = 50ms) to produce firing rates in spikes per second (sps/s), which were down-sampled from 1000Hz to 100Hz, allowing for more efficient computing, while maintaining a reasonably high temporal resolution. Because timings between events in a trial were randomized we aligned firing rates to the onset of events of interest and chose different cut-off values around those time points: -0.5s to 1.1s (attention cue); -0.5s to 1.1s (memory cue); -0.4s to 0.7s (attention target); -0.7s to 0.5s (memory target). We then spliced those snippets to obtain a single firing rate trace. Neural data analysis focused on dual task trials. We only included correct trials in our analysis and excluded trials were the attention target was scheduled <0.8s after the memory cue onset. We did not exclude catch trials from our analysis. Furthermore, we required at least 5 trials per condition (3 attention conditions x 6 WM conditions = 18 conditions). Neurons were not screened for tuning properties but had to have a non-zero firing rate evaluated on the spliced traces. Those steps led to the exclusion of three sessions; leaving us with 508 neurons from 15 sessions in monkey T and 379 neurons from 24 sessions in monkey K. All population level analysis, that is all analysis except those using the PEV metric (see below), were

preceded by a soft normalization step. For each time point we divided the single trial firing rates by the square root of the firing rate range estimated over all trials and time points. This was done to mitigate the influence of high firing rate/variance neurons (see Elsayed et al., 2016; Remington et al., 2018 for a similar procedure). Because trial numbers varied across conditions, all analysis within this study included a stratification step. That is, we randomly sub-selected trials from each condition so that all conditions contained as many trials as the condition with the lowest trial number. This was normally done within session except for the angular decoding approach were this had to be done across sessions. All analyses were performed on stratified data and final results present the average over 100 repetitions for different sub-selections of trials. All data analysis was implemented in Python using custom written code, unless noted otherwise.

Behavioural analysis

Behavioural analysis was restricted to sessions and trials that were used in the neural data analysis. For the attention task (dual and single) we compared the percentage of correct rates and lever release RTs using a series of pairwise Wilcoxon signed-rank tests or Wilcoxon rank-sum test (where applicable), which were corrected for multiple comparisons using Holm's SRB procedure. For percentage correct rate calculations, errors trials were premature lever release or failure to initiate lever release within 0.85s of attention target onset. Trials on which monkeys broke fixation were simply discarded (not counted as either error or correct trials). For the single and dual WM tasks we applied the same non-parametric testing and correcting procedures as in the attention tasks. For percentage correct rate calculations, errors trials were failure of correct saccadic target acquisition within 0.5s after memory target onset or failure to

maintaining fixation on correct memory target for at least 0.6s. Trials on which monkeys broke fixation before memory target onset were simply discarded (not counted as either error or correct trials).

Percentage of explained variance metric

To quantify the information content of single neurons we applied the bias corrected percentage of explained variance (PEV; ω^2) metric (Olejnik and Algina, 2003), which can be extended for multifactorial designs using a partial PEV i.e. partial ω^2 (ω_p^2) index (Kline, 2004). Specifically, we were interested in the amount of information each neuron carried about spatial locations i.e. conditions in the WM task (6 possible conditions) and attention task (3 possible conditions) as well as their combinations (18 possible combinations). To this end, we performed a 2-way ANOVA including WM task and attention task conditions as main factors. The resulting parameters were used to calculate the partial PEV (which is referred to as PEV in the results):

$$\omega_p^2 = \frac{(df_{effect} - 1) * (F_{effect} - 1)}{(df_{effect} - 1) * (F_{effect} - 1) + N} \qquad (2.1)$$

where *df* is the degree of freedom and *F* is the F-value of the factor under consideration. *N* is the number of samples. For each neuron and time point we calculated the partial PEV for the WM task conditions, for the attention task conditions and for their combinations i.e. interactions.

Statistical testing: We used a cluster-based permutation test (Maris and Oostenveld, 2007) to evaluate the significance of the observed partial PEVs for each neuron and time

point. For this, we shuffled trial labels for each neuron and recalculated the partial PEV. We repeated this procedure 1000 times to create a null distribution of partial PEV values. Contiguous time points at which the observed PEV exceeded the 95th percentile of the null distribution were deemed candidate clusters. To correct for multiple comparisons over time, we compared the maximum observed summed cluster to a new null distribution of maximum summed clusters derived from the first null distribution. If the size of the observed cluster exceeded the 95th percentile of the new null distribution it was deemed significant. To test for a significant partial PEV (greater than zero) at the level of the population we applied a similar procedure, which differed only in the creation of the first null distribution. Specifically, we flipped the sign of the partial PEVs of half of the neurons (randomly selected) and averaged over all neurons. We then obtained the null distribution by repeating this process 1000 times.

Assigning neurons to sub-populations: Neurons were assigned to different subpopulations based on their PEV. Pure selectivity cells: A significant PEV for one main factor at any time point starting from attention cue onset and no significant PEV for the interaction (at any time point). LMS cells: A significant PEV for both main factors at any time point from memory cue onset and no significant PEV for the interaction (at any time point). NLMS cells: a significant PEV for the interaction between main factors at any time point from memory cue onset.

Cross temporal discriminability analysis

Cross temporal discriminability of location information on the level of the neural population was assessed using the analysis described in the studies by Stokes et al.

(2013) and Spaak et al. (2017). Within each task, we randomly assigned each trial to one of two independent data splits. We then computed the mean firing rate over trials per neuron per independent split and per condition and took the pairwise differences between all conditions (3 for the attention task and 15 for the WM task) for each neuron within each independent split. For each pairwise condition difference we then calculated the Pearson correlation for each time point across neurons between independent splits and averaged the resulting correlation coefficients using Fisher's z-transformation to obtain a single time resolved discriminability measure (discriminability score). This procedure is analogous to a decoder being trained on split one and time point one t_i and being tested on t_i' in split two. It is straightforward to extend this basic decoder to cross temporal decoding by computing the correlation between each time point (t_i) in split one and all (same and other) time points t_j' in split two. The result is a two dimensional matrix (time by time) representing the discriminability score at all time points (on diagonal) as well as generalisability across time points (off diagonals). Significance of discriminability was assessed by cluster-based permutation testing with randomly shuffled condition labels (1000 permutations)

Significant off-diagonal reduction (Dynamic coding): We tested for dynamic coding over time by evaluating whether the off-diagonal discriminability $r(t_i,t_j')$ was significantly reduced relative to the corresponding on-diagonal discriminabilities $r(t_i,t_i')$ and $r(t_j,t_j')$ (Spaak et al., 2017). More specifically, we computed $r(t_i,t_i')$ - $r(t_i,t_j')$ and $r(t_j,t_j')$ - $r(t_i,t_j')$ and determined whether both of these values were significantly greater than zero using a cluster based permutation test, where the null distribution was constructed by randomly shuffling off-diagonal versus on-diagonal points a 1000 times.

Principal component analysis

Many analyses within this study involve the use of principal component analysis (PCA). In general, PCA was performed as follows: First, for each neuron and time point of interest we computed the mean firing rates across trials per condition yielding an initial data matrix $D \in {}^{N x K x T}$, where N is the number of neurons, K is the number of conditions and T is the number of time points of interest. We then mean-centered D by subtracting the column means and concatenated the condition and time axis to obtain the final data matrix $D \in \mathbb{R}^{N \mathrm{x} KT}$. We then performed PCA on D to obtain Eigenvalues and principal components (PCs; or Eigenvectors). Any arbitrary set of PCs defines a subspace which can be written as a matrix $Q \in \mathbb{R}^{P \mathrm{x} N}$, where P is the number of PCs i.e. the dimensionality of the subspace. To obtain visualization as in Fig. 2.5a&c, Fig. 2.7c&d & 2.8a and Supp. Fig 2.2a&c, we typically projected certain time-points or time averages of data matrices (D) onto the top PCs (1-3) of subspaces (Q) defined in different splits and/or different time points of the data. Note, visualizations in Fig. 2.5a&c and Supp. Fig. 2.2a&c contain exceptions where the to-be-projected data D was defined on the same data as the subspace Q that was used for projections.

Alignment index

We used the alignment index (Elsayed et al., 2016) to quantify the overlap between subspaces defined in different parts (x,y) of the data. Depending on the analysis, different parts of the data can refer to different time points during the trial, different train/test splits of the data or data from different tasks. Therefore, we express the AI in following generic formula:

$$AI = \frac{Tr(Q_x^T C_y Q_x)}{\sum_{i=1}^{P} \sigma_{y(i)}} \quad (2.2)$$

where Q_x is the *P* dimensional subspace obtained from the data matrix D_x in part *x* of the data. C_y is the covariance matrix of the data matrix D_y from part *y* of the data. σ_y are the singular values of C_y and *Tr()* is the matrix trace. The numerator reflects the amount of variance in part y explained by the P dimensional subspace of part x. The denominator reflects the variance in part y explained by the P dimensional subspace also defined in part *y,* and hence the maximal amount of variance explainable by the top *P* dimensions of part *y*. Thus, the AI gives an unbiased measure of subspace overlap, which is important because different tasks and/or task epochs substantially differ in variance and dimensionality of neural activity. An AI of 1 indicates perfect overlap between subspaces whereas an AI of 0 indicates perfectly non-overlapping (orthogonal subspaces). Subspace dimensionalities to compute AIs were usually set to 5, unless otherwise mentioned. This was because 5 dimensional subspaces usually explained > 80% of the variance in our data (>85% in the attention task). However, we tested also other dimensionalities (1-10), which did not change any of our main results.

Cross-validation procedure: Except in Fig. 2.5 and Supp. Fig. 2.2 AIs were cross-validated. We randomly assigned each trial to one of two independent data splits (split x and split y) such that both splits had equal numbers of trials per condition. For each neuron, trials within splits were averaged according to conditions of interest, for each time point of interest and data matrices were created (D_x and D_y). AIs were computed according to Eq. 2.2, where x now represents split x while y represents split y. This was

also done in reverse and the average of the two AIs was taken as the final cross-validated AI.

Across task comparisons: To compare AIs across tasks (Fig. 2.7), in one split of the data we averaged all trials with the same attention condition, irrespective of the WM condition, whereas in the other split we averaged trials according to WM conditions, irrespective of attention conditions. AIs were computed according to Eq. 2.2, and following our general procedure.

Within task comparisons: For within task comparisons (Fig. 2.6 and Fig. 2.7a), the procedure was identical except that in both splits data was averaged according to WM conditions or attention conditions.

Time resolved AI measures: In most cases (Fig. 2.6; Fig. 2.7a&b; Fig. 2.8c; Fig. 2.9; Fig. 2.10) we calculated AIs in a sliding time window (0.4s; 0.1s step-size) across the whole trial, which ensured a sufficient temporal resolution. To extend this procedure across time we followed the same protocol as described in our cross-temporal discriminability analysis. That is, the x in Eq. 2.2 now represents time window x in split 1, while y represents data from time window y in split 2. Doing this for all possible pairs of time windows yields the matrices of across-time AIs (Fig. 2.6b&c; Fig. 2.7b). The within-time AIs are represented on the diagonal of those matrices. Note, Fig. 2.5d and Supp. Fig. 2.2d contain exceptions where the across-epoch AIs where not cross-validated.

Statistical testing: We wondered in how far subspace overlap was determined by the dimensionality of the sampled subspaces, and the dimensionality and covariance of the data. If, for example, the dimensionality of the data is high (as determined through the number of neurons and the covariance between them), and the dimensionality of the sampled subspaces is low, by chance one would expect a low AI and vice versa. To test whether AIs were different from what would be expected by chance given the data covariance and chosen subspace dimensionalities we adopted a recently developed testing procedure (see Elsayed et al., 2016). First, we estimated the covariance matrix C of the neural data over all conditions and all available time points of the trial. We then initialized a random matrix (drawn from a Gaussian distribution) from which we drew pairs of subspaces of the same dimensionality as our observed subspaces (which were usually five dimensional). In a final step, we normalized the subspaces to unity and biased them towards C (see Elsayed et al., 2016, Supplementary Material). We created 1000 pairs of subspaces and computed the AI between them, thus constructing a null distribution of AIs, which was used to determine whether the observed AIs were significantly smaller ($<2.5^{th}$ percentile) or larger ($>97.5^{th}$ percentile) than expected by chance.

Subspace optimization

To find maximally orthogonal subspaces between task epochs (Fig. 2.5b&c) or tasks (Fig. 2.7d&e) we applied a recently developed optimization method (Elsayed et al., 2016), which is particularly robust to different dimensionalities or amounts of variance during different task epochs or tasks, while explicitly orthogonalizing subspaces. This is not always guaranteed in similar Methods (e.g. Kobak et al., 2016). The optimization

method finds two orthogonal subspaces of arbitrary dimensionality, one defined in part *x* of the data and the other defined in part *y* of the data. Importantly, the variance explained within each subspace is kept as high as possible while ensuring orthogonality between subspaces, which is done by optimizing the following objective function:

$$[\hat{Q}_x, \hat{Q}_y] = argmax_{Q_x,Q_y} \frac{1}{2}\left(\frac{Tr(Q_x^T C_x Q_x)}{\sum_{i=1}^{P_x} \sigma_x(i)} + \frac{Tr\left(Q_y^T C_y Q_y\right)}{\sum_{i=1}^{P_y} \sigma_y(i)}\right) \qquad (2.3)$$

$$subject\ to: \quad Q_x^T Q_y = 0, \quad Q_x^T Q_x = I, \quad Q_y^T Q_y = I$$

where variable names and subscripts are the same as in Eq. 2.2. $\hat{Q}_x$ and $\hat{Q}_y$ are the optimized subspaces of dimensionality P_x and P_y, respectively, which do not have to be the same but were set to be 5, unless noted otherwise. Optimization was performed over the Stiefel manifold using the Pymanopt toolbox (Townsend et al., 2016). Note, Eq. 2.3 can be easily extended to optimize three or more subspaces, as was done for the analyses pertaining to Fig. 2.5 and Supp. Fig. 2.2.

Statistical testing: We tested whether optimized subspaces explained significant amounts of variance using a bootstrapping procedure. That is, we sampled 1000 subspaces - of the same dimensionality as the optimized subspaces - from the respective part of the data and quantified the variance explained. If the variance explained in the optimized subspaces exceeded the 95th percentile of the bootstrapped distribution it was deemed significantly higher than expected by chance. We also compared the difference between the AIs of the optimized subspaces with the AIs of the original subspaces defined in the same part of the data, using the method for creating a null distribution of

AIs described in the *Alignment index* section. By taking the differences between the 'random' AIs, we can construct a distribution of random AI differences, which can be used to evaluate the significance level of the observed difference between the AI for the optimized and original subspaces ($p<0.025$).

Angular decoding

We aimed to test whether a decoder trained to identify locations in the WM task could be applied to identify locations in the attention task. Since different screen locations were used as cues in the two tasks, we translated WM and attention locations into a common reference frame by calculating, for each location, the visual angle of the location with respect to a horizontal line extending rightward from fixation. We then applied an angular decoding approach (see Supp. Fig 2.3a) (similar to Trübutschek et al., 2019). For this, we randomly assigned each trial to one of two independent data splits (split x and split y, with equal amounts of trials per condition). We then fitted a combination of two ridge regression models to the data in split x. One model contained the sine (*sin(x)*) and the other model contained the cosine (*cos(x)*) of the WM locations' angles. We then used the weights from our fitting procedures to predict the sine and cosine of the attention conditions (*sin(y)* and *cos(y)*) in data of split y. The arctangent of the individual predictions gives the estimated angle $\hat{y} = arctan2(\hat{y}_{sin}, \hat{y}_{cos})$ for each trial in the data of split y. We then calculated the mean absolute difference between the actual (y) and predicted ($\hat{y}$) angles across trials and computed a deviation from chance score (which we subtracted from π) with the following range: $-\pi/2$ to $\pi/2$ (chance = 0). Note, logically for this procedure we had to discard trials on which the attention cue was presented on the fixation ring (since it had an eccentricity of 0, its angle was not defined).

We used cluster-based permutation tests based on shuffling trial labels to assess statistical significance.

Across-task discriminability analysis

As in the cross-temporal discriminability analysis (see Methods), we randomly assigned each trial to one of two independent data splits (split x and split y; with equal amounts of trials per condition). We then computed the mean firing rate over trials per neuron per split and per condition. Importantly in split x we averaged according to the 6 WM conditions while in split y we averaged according to the three attention conditions. In split x we then constructed 3 'artificial' attention conditions through a weighted average of the 6 WM conditions, as determined by the true Euclidean distances between the actual attention and WM cue locations. We then took the pairwise differences between the 3 'artificial' attention conditions in split x and the 3 'real' attention conditions in split y and calculated the Pearson correlation as described before.

Statistical testing: To test for a significant difference from chance we used cluster-based permutation tests as described before. However, instead of randomly shuffling trial labels, to construct the null distribution we assigned all trials from one condition the label of another random condition and so on.

Correlation of loading weights/PEV

Loadings: To test whether same or different cells contribute to a given set of subspaces we computed the correlation between subspace loadings. Loadings are defined as the

product between the subspace (Q) and the square root of the eigenvalues obtained from our PCA analysis. Since the result is a matrix: $L \in \mathbb{R}^{NxP}$, we calculated the norm for each row, resulting in a vector where each entry represents the 'average' load or contribution of neuron on the subspace. Loadings of the same neurons onto different subspaces, defined in different splits of the data, were correlated using Spearman's r. Loadings were computed in a sliding 0.4s window (0.1s step-size), but also across time (as describe before).

Statistical testing: To test whether correlation between subspaces was different from chance we leveraged the same Method used to determine significance of the AI (see *Alignment index* section) and constructed a null distribution by computing the loading correlation between pairs of randomly sampled subspaces of the same dimensionality (1000 samples). An observed loading correlation was deemed significant if it was below the 2.5^{th} percentile or above the 97.5^{th} percentile of the null distribution.

Cross-temporal PEV: To expand on the question of different versus same sub-populations, we also correlated the PEV between splits of the data as well as within and across time. For this, we randomly assigned each trial to one of two independent data splits (split x and split y) such that both splits had equal numbers of trials per condition. We than computed the PEV (see *Percentage of explained variance metric* section) for the WM and attention task conditions in each of the splits. When then computed the Spearman correlation between the WM PEV in split x with the attention PEV in split y (as well as vice versa) and averaged the resulting correlation values. This was done within time points of different splits as well as across time points of different splits, as described before for various other metrics. Statistical testing was performed using a

permutation test where the null distribution was created by shuffling trial labels, and observed values that exceeded the 95th percentile of the null distribution were deemed significant.

Across instance analyses

For the across instance analyses we reduced the six original WM conditions to three in the following way: We pooled trials from (1) the two locations contralateral to the attention locations; (2) the two upper locations ipsilateral to the attention locations and (3) the two lower locations ipsilateral to the attention locations. Pooling was done to ensure >5 trials per condition in the across instance analyses.

Across instance discriminability: For a given task e.g. the attention task, we calculated the average response to all nine (3*3) combinations of attention and WM conditions in two independent splits of the data. We then computed the pairwise differences between the three attention conditions, for each WM location and in each split of the data. Our aim was to test whether pairwise differences (i.e. the discriminable pattern) for attention conditions was conserved across different WM conditions (and vice versa). To this end we calculated the Pearson correlation between the attention pairwise differences for one WM condition and the corresponding pairwise differences (in the other half of the data) for the same WM condition (within-instance discriminability) or a different WM condition (across-instance discriminability). We did the same for the WM task but vice versa. The within-instance discriminability gives an upper bound of attention task or WM task condition decoding. The difference between the within- and across-instance discriminability scores reflects the cost of generalizing the task specific activity patterns

across conditions of the other task. Other decoding methods (linear discriminant analysis, not presented here) led to comparable results.

Across instance AI: For the within and across-instance AI analysis the data was ordered in the same way as for the within- and across-instance discriminability analysis and the AI for a given task was calculated within or across conditions of the other task. AIs were computed in a sliding 0.4s window (0.1s step-size)

Statistical testing: We used two types of cluster-based permutation tests to evaluate whether effects in the discriminability analysis were significant. First, we checked whether within- and across-instance discriminability scores were different from chance by comparing the observed scores with that of a null distribution obtained by randomly shuffling trial labels, while keeping all other aspects of the analysis constant (1000 permutations). In a second test we wanted to determine whether the difference between the within- and across-instance discriminability scores or AIs were significantly larger than expected by chance. To this end, for a given task, we only shuffled trial labels pertaining to the other task and computed the difference between the within- and across-task discriminability scores or AIs. We repeated this procedure 1000 times to create a null distribution of differences to which the observed differences were compared (differences $<2.5^{th}$ percentile or differences $>97.5^{th}$ percentile, were deemed significant).

Simulations

We simulated a simplified version of the task used in the experiment, which nevertheless retained all critical features. Each trial began with a 50ms baseline period, followed by

a 250ms period in which the attention cue was presented, and a 250ms attention-only delay. Next, the WM cue was presented for 250ms, followed by another 250ms delay. The simulated task used three possible attention and WM locations. In contrast to the experimental task, the simulation used the same locations for the attention and WM tasks. This made it easier to simulate the effect of task-invariant coding (i.e., cells that responded to the same location irrespective of whether it was cued for the attention or the WM task).

Selectivity types: Simulated neurons could be tuned only to an attention location or a WM location, or to locations for both. We separately simulated the activity of six different populations: 1) Pure selectivity neurons. Each neuron was tuned exclusively to either an attention location or a WM location, i.e., there was no mixed selectivity. Half of the population was tuned to an attention location, while the other half was tuned to a WM location. 2) Common coding neurons. Each neuron was tuned to the same location for the attention and the WM task, i.e. a form of linear mixed selectivity (LMS). 3) Common coding with saturation. This population was the same as population 2, but activity saturated: the firing rate of tuned cells was the same on trials where the preferred location was cued for both the attention and WM tasks as on trials where the preferred location was only cued for one of the two tasks. This is a form of nonlinear mixed selectivity (NLMS). 4) LMS neurons. The cells in this population were all tuned to an attention and a WM location. However, spatial tuning was independent across tasks. That is, cells tuned to location 1 in the attention task could be tuned to locations 1, 2, or 3 in the WM task. The selectivity was purely additive (i.e., no saturation), such that this population exhibited LMS. 5) NLMS. These cells were tuned to both attention and WM locations with no consistency across tasks. In other words, a cell could be tuned to WM

location 1 on trials where attention location 2 was presented, but to WM location 3 on trials where attention location 1 was presented. This resulted in NLMS. 6) Empirical population. Finally, we simulated an estimate of the empirical population, consisting of untuned cells (37.4%, always at baseline firing rate), pure WM cells (29.6%), pure attention cells (14.3%), LMS (8.5%), and NLMS cells (10.2%).

Simulation protocol: Each simulation used a population of 100 cells. For each cell and trial, we simulated a Poisson spike train at 1000Hz. The baseline firing rate was set to 10Hz. During the attention and WM encoding periods, cells tuned to the presented locations increased their firing rate to 60Hz (while untuned cells or cells tuned to other locations maintained the baseline firing rate). During the subsequent delay periods (which, for the attention feature, extended to the end of the trial and thus overlapped with the WM encoding and delay periods), the firing rate went down to 35Hz (i.e., remained above baseline). Spikes were simulated by drawing from a uniform distribution. Whenever the random draw fell below the instantaneous firing probability (current firing rate * time bin width), a spike was recorded. Resulting spike trains were smoothed with a 50ms Gaussian kernel. For each population, we simulated 900 trials. We repeated each simulation 30 times.

Analysis performed on simulations: We performed two main analyses on each simulated population. Firstly, we examined subspace overlap (via the AI metric) between attention and WM subspaces. Subspace projections for attention and WM were calculated on the data from their respective encoding periods. However, given that tuning did not change over the course of the trial, the chosen time window should not influence the calculated subspace. We also performed the within- and across-instance discriminability analysis

on the simulated populations. This analysis was performed exactly the same way as on the real data. Other decoding methods (linear discriminant analysis, not presented here) led to comparable results.

2.4 Results

Behavioural performance

Monkeys performed a dual task consisting of a spatial attention component task and a spatial working memory (WM) component task. For a full description of the behavioural paradigm refer to Methods and Fig. 2.1 therein. Monkeys performed the dual task with a high success rate. The percentage of correct trials (accuracy) for the attention task was around 89 % for monkey K (ranging from 83-97% across sessions) and around 81% (ranging from 70-89% across sessions) for monkey T. Accuracy for the WM task was even higher; around 98% in both monkeys (ranging from 93-100% across sessions). In fact, overall, accuracy for the dually performed tasks did not differ from the accuracy for the corresponding single tasks (Fig. 2.2). This was true even when splitting the data into the different conditions.

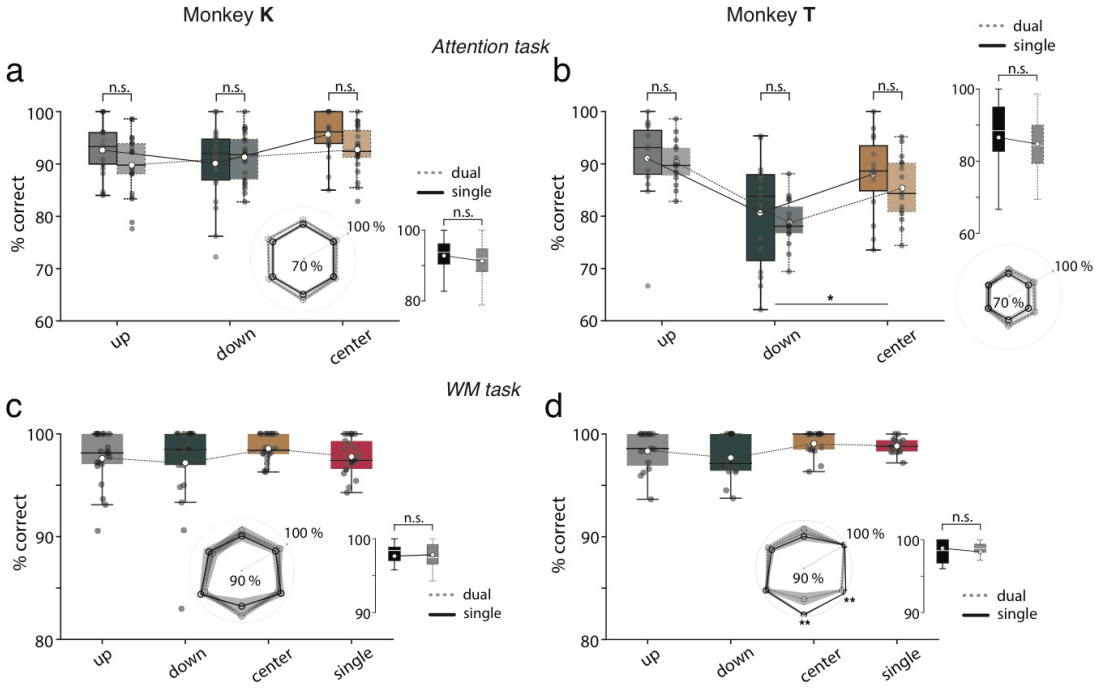


Figure 2.2: Behavioural performance - Accuracy. Boxplots showing the monkeys' behavioural performance via percentage of correctly performed trials (% correct; accuracy). Throughout the figure dots represent unique sessions (24 for monkey K; 15 for monkey T). Whiskers extend to the upper and lower quartiles of the respective distributions. Black or white lines within boxes indicate the median and white dots indicate the mean of the respective distributions. * $p<0.05$; ** $p<0.01$; *** $p<0.001$ (*Wilcoxon signed-rank test*; corrected for multiple comparisons by *Holm's sequentially rejective Bonferroni procedure*). **a) & b)** Accuracy for the dual attention task (light colours & dashed line) and single attention task (dark colours & solid line) as a function of attention cue location (up, down & center). Box plot insets compare overall accuracies for the dual (grey; dashed) and single (black; solid) attention tasks, irrespective of cue locations. Hexagonal plot insets compare accuracies for the dual attention task as a function of WM cue location (dashed lines). Note, solid lines indicate the accuracy for the single attention task, which is constant since there was no concurrent WM component, but displayed for comparison (shaded areas: s.e.m.). **c & d)** Accuracy for the dual WM task as a function of the concurrent attention cue conditions (up, down & center) and when performed on its own (single). Box plot insets compare overall accuracy for the dual and single WM tasks, irrespective of cue locations. Hexagonal plot insets compare accuracies for the dual (grey; dashed) and single WM tasks as a function of WM cue location (black; solid) (shaded area: s.e.m.).

For example, accuracy for the dual attention task did not differ from that for the single attention task at any of the three possible attention cue locations or six possible WM cue locations (Fig. 2.2a,b). Likewise, accuracy for the dual and single WM tasks did no differ at any attention cue location or WM cue location (Fig. 2.2c,d), except for monkey K, which showed a lower accuracy for the dual WM task at two WM locations (Fig. 2.2d)

($p<0.01$; *Wilcoxon signed rank*). For a similar breakdown regarding the monkeys' reaction times (RTs) refer to the Appendix of this chapter (Supp. Fig. 2.1). Together, these results indicate that monkeys are able to perform two tasks simultaneously while maintaining their high performance on each of the component tasks, indicating little to no task interference on the behavioural level.

Single neurons dynamically represent a mixture of task variables

We recorded single neurons from lateral prefrontal cortex (lPFC) of the two monkeys (K: n=379; T: n=508). Our main interest in this study was how information about two relevant, temporally overlapping information streams (attention task and WM task) is represented in lPFC. Therefore, neural analyses presented here focus exclusively on dual task trials, unless otherwise noted. Furthermore, in the remainder of this study we refer to different cue or target locations within tasks as features.

First, we quantified the amount of information (i.e. selectivity) each cell carried about features from both component tasks (see Methods). Neurons were selective for and maintained feature specific information from both the attention and WM tasks (Fig. 2.3a). Furthermore, there was a significant interaction effect between attention and WM task features, indicating an overlap in processing resources between the two tasks. Additionally, as observed in previous studies (e.g., Spaak et al., 2017) many neurons seemed to only transiently encode features within tasks and some even seemed to switch their selectivity preference over time (Fig. 2.3b). Note, number and significance of switches were not explicitly analyzed in the present study. When restricting our analysis to the memory cue and subsequent delay epoch we observed that a majority of selective

cells exclusively reflected information pertaining to features of either the attention or WM task. Throughout this study such cells were termed pure selectivity cells. However, a considerable proportion of selective cells showed mixed selectivity. More specifically, some cells were selective to features from both the attention and WM task and their response to any combination of features across tasks could be explained by a linear combination of their responses to features within tasks i.e. they displayed linear mixed selectivity (LMS). Conversely, other cells were selective to specific combinations of features from both tasks, which could not be explained by a linear combination of their responses to features within tasks i.e. they displayed non-linear mixed selectivity (NLMS) (Rigotti et al., 2013). Different proportions of cell types, as estimated around the time of WM cue presentation, are displayed in the scatter plot in Fig. 2.3c. The scatter plot also indicates that cells with selectivity for attention task features also tended to be selective for features in the WM task. In total, our univariate analysis revealed that cells in lPFC encode and simultaneously maintain information about both the attention and WM tasks. However, a proportion of neurons mixes information from both tasks. Further note, during the epoch analyzed in Fig. 2.3c, cells with a high selectivity in one task tended to also be selective in the other task (Fig. 2.3c).

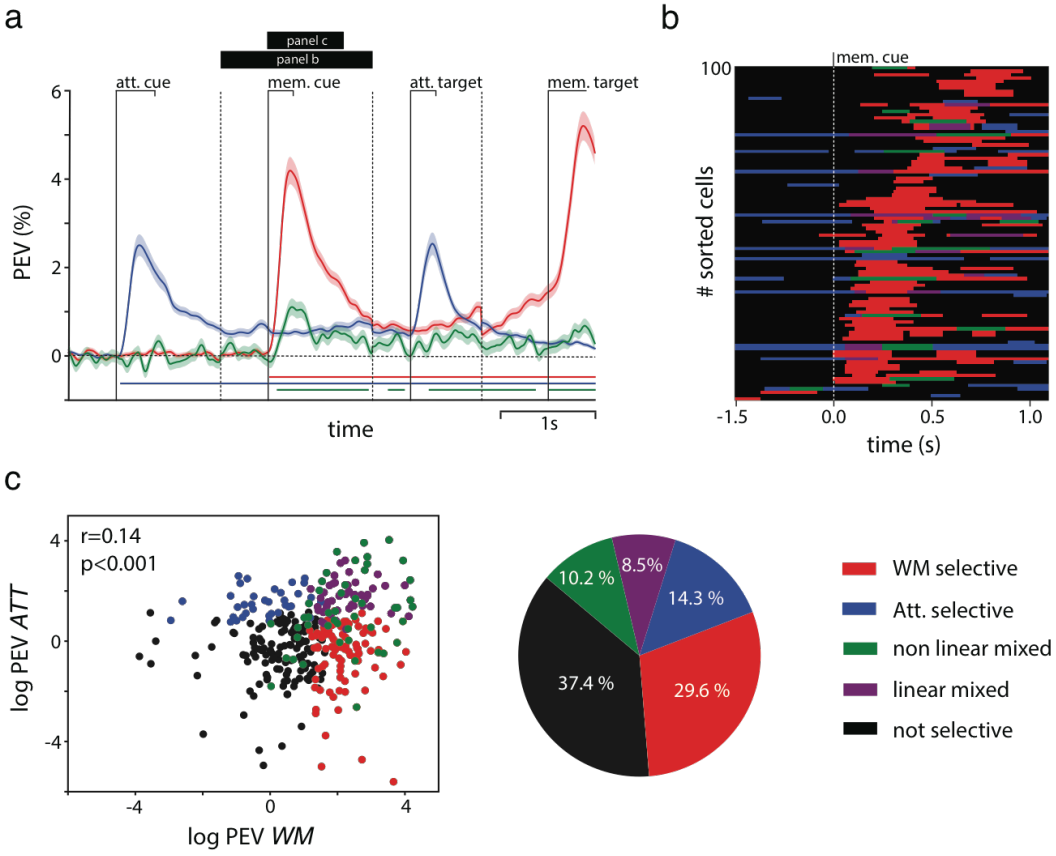


Figure 2.3: Task related & temporal tuning in single neurons. a) Percentage of explained variance (PEV) obtained from a per-neuron (cluster-corrected) 3x6 ANOVA for the attention (blue; three levels) and WM (red; six levels) task features as well as their interaction (green). Results are averaged over all neurons (n=887). Shaded areas indicate s.e.m. Solid straight lines indicate significant PEV on the level of the population ($p<0.05$, *cluster based permutation test*). Black boxes indicate the time windows of analysis in panels **b** and **c**, respectively. Black vertical lines mark the onset of relevant task events (solid) and splicing boundaries (dashed) (see Methods). **b**) Random selection of 100 cells that showed a significant PEV i.e. selectivity for attention or WM task features or their interaction. Significant effects are colour coded and cells are sorted according to the time point of their peak selectivity to the WM task features. **c**) On the left: Log transformed PEV values for the attention task are plotted against the log transformed PEV values for the WM task. Each dot represents a cell and the colour indicates its selectivity. Note, the significant positive correlation ($r=0.14$; $p<0.001$; Spearman correlation). On the right: Pie chart indicating the relative percentages of single cell selectivities across the whole population for that time window.

The population code radically changes across task epochs

Given that information about both component tasks was present throughout the trial, we wanted to explore the stability and format of the population code over time. To this end, we applied a cross-temporal decoding approach (Spaak et al., 2017) (see Methods); also

described in **chapter 3**. In brief, for a given component task, we examined how well the feature-specific neuronal activity patterns generalized from one time point to any other time point in the trial. The resulting pattern discriminability scores are plotted in Fig. 2.4. Within-time point discriminability scores showed a similar time course as was revealed by our single neuron analysis (Fig. 2.3a; and along the diagonal of the matrices in 2.4b, c). In general, the population encoded information about features within both tasks throughout all epochs of the trial. Importantly, information was not stable across all epochs of the trial, i.e., the discriminative patterns changed across time (Fig. 2.4b,c). In the attention task, the pattern of activity during the attention cue encoding showed a reduced cross generalization to that of the following delay epoch (first attention delay) but again robustly cross generalized to when the attention target was presented (Fig. 2.4b). Also, during the presentation of the memory cue, attention cue information was present but did not cross generalize well to any other time point in the trial. Changes in cross-temporal discriminability were even more distinct during the WM task. Here, discriminability was reduced between the memory cue and subsequent delay period (first WM delay). However, the pattern of activity during the first delay, albeit temporarily stable, again did not cross-generalize well to the activity pattern emerging after the attention target presentation (second WM delay; similar to Parthasarathy et al., 2017).

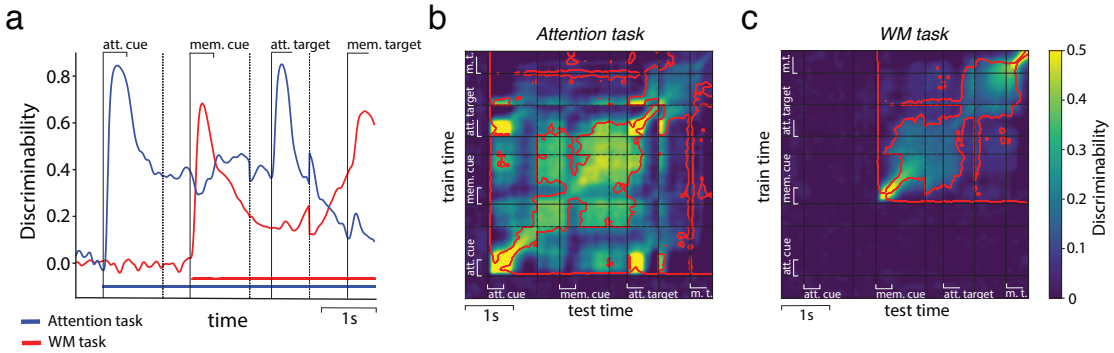


Figure 2.4: Population decoding. a) Within-time multivariate discriminability scores for the attention task (blue) and WM task (red) when performed concurrently. Solid lines indicate a significant discriminability ($p<0.05$; *cluster based permutation test*). Throughout the figure, black vertical lines mark the onset of relevant task events (solid) and splicing boundaries (dashed) (see Methods). **b) & c)** Across time discriminability score matrices for the WM and attention tasks. Colour scale represents magnitude of discriminability scores i.e. decoding performance. Red contours indicate a significant reduction of the across-time (off diagonal) discriminability scores compared to the corresponding within time (on diagonal) scores (see Methods) ($p<0.05$; *cluster based permutation test*).

In the following, we aimed to more explicitly understand changes in the feature specific patterns occurring across distinct task epochs. The population activity pattern at any time point can be thought of as a point within a high dimensional space spanned by all neurons i.e. state space. However, the activity patterns within a distinct task epoch – a collection of points in state space - is likely to be well described by a lower dimensional subspace (a set of directions in state space). If we are able to cross generalize feature specific information from time point x to time point y, this means that the activity patterns lie in the same, or overlapping subspaces. However, if the we fail to cross generalize this can mean that either the pattern has moved to a different subspace, or that it still resides in the same subspace but relations between conditions have changed and thus are no longer generalizable.

We illustrated this approach using the WM task and further elaborate on its implications in Fig. 2.5. In general, we used principal component analysis (PCA) to compute three

different subspaces for the WM task (see Methods): One for the memory cue (cue); one for the first memory delay (delay 1) and one for the second memory delay (delay 2) (Fig. 2.5). Each subspace consisted of the top ten principal components (PCs) defined in each task period, which explained >90% of the variance among conditions. Projecting the activity (averaged by feature condition) of a given task epoch onto its respective subspace faithfully recovered the ring-like structure of the original memory cues (Fig. 2.5a). By eye, it also becomes clear that the geometrical relation between features was largely preserved across task epochs (the same formation emerges in each time window). However, when projecting feature-related activity of one task epoch into the subspace of another, the representational structure collapsed. In other words, epoch-specific subspaces defined at one time window did not capture relevant variance among features at other task epochs.

To quantify this more specifically we computed the alignment index (AI) between subspaces of different task epochs (see Methods). In brief, the AI is the ratio of the variance in epoch x explained in the subspace defined at epoch y, and the variance in epoch x explained in the subspace defined at epoch x. An alignment index of one (AI=1) indicates that subspaces perfectly align, whereas an alignment index of zero (AI=0) indicates that subspaces do not align at all, i.e., that they are orthogonal to each other. AIs within and across task epochs are displayed in Fig. 2.5b. Note, unless cross-validated, the AI of a given subspace with itself is one, by definition, but is displayed here for illustrative purposes. Indeed, the AI between task epochs was close to zero, confirming that the reduced cross-temporal generalization in our decoding approach (Fig. 2.4a,b) could be attributed to a change in the task-relevant subspace. In fact, epoch-specific subspaces were more orthogonal than expected by chance (Fig. 2.5b).

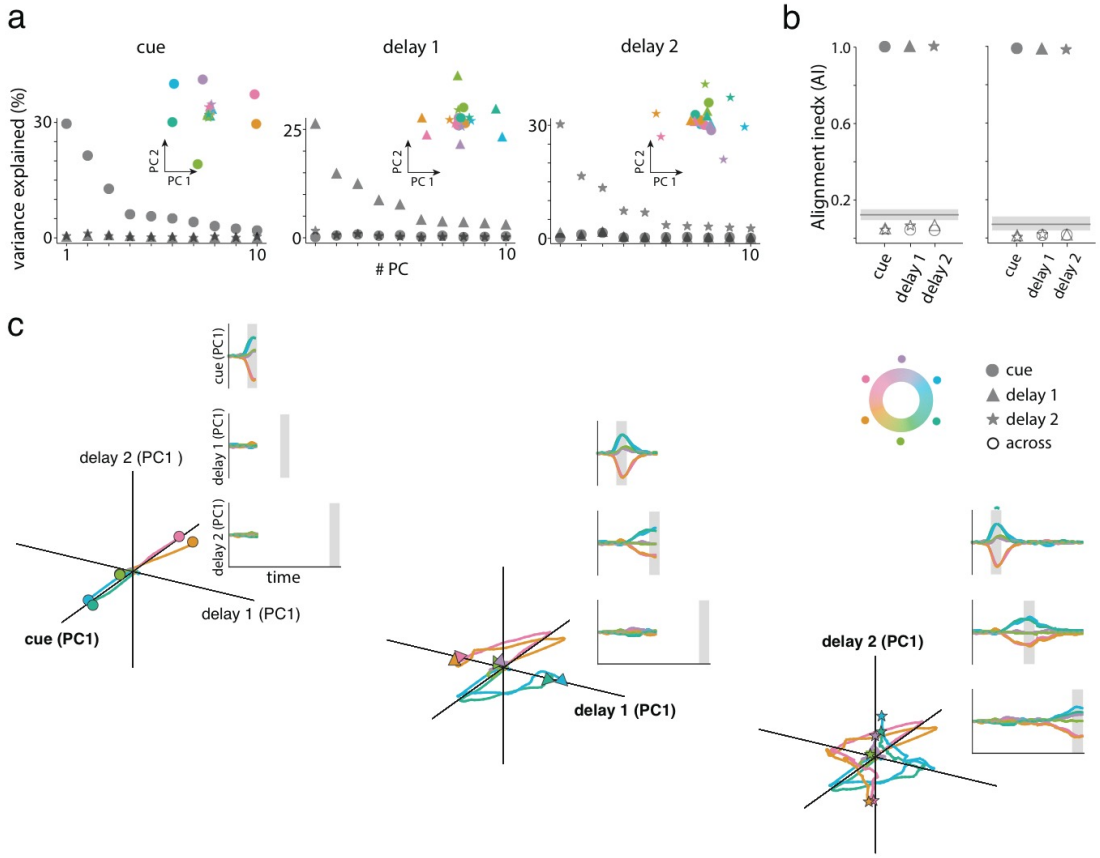


Figure 2.5: Subspace analysis – WM task. a) Condition averaged activity from three distinct task epochs was projected onto the top 10 principal components (PCs) i.e. 10 dimensional subspace computed at each task epoch. The variance in epoch-specific activity explained by each PC is displayed. Cue epoch (0.1s-0.4s after memory cue presentation; circle); delay 1 epoch (0.8s-1.1s after memory cue presentation; triangle); delay 2 epoch (-0.4s to 0.1s before memory target onset; star). Insets show condition averaged activity from each task epoch projected onto the top 2 PCs of each task epoch. **b)** The alignment index (AI) (see Methods) quantifies the overlap between epoch-specific subspaces. Symbols indicate the AI between the subspace defined at one epoch with that defined at any other epoch (empty symbols) or the same epoch (solid symbols). The grey solid line and shaded area indicate the mean and 95% confidence interval of the null distribution of the AIs computed between randomly sampled subspaces of the same dimensions (see Methods). The left panel shows the AI for the originally defined 10 dimensional subspaces. The right panel shows the AI for the same set of subspaces but reduced to 5 dimensions, and optimized to be explicitly orthogonal (see Methods) **c)** Condition averaged trajectories in the three dimensional subspace defined by the optimized first PCs for each epoch. Different panels show trajectories up to the cue (left), delay 1 (middle) and delay 2 (right) epochs. Insets show trajectories on the first PC alone for each epoch.

We then leveraged the inherent near-perfect orthogonality between the epoch subspaces to apply a recently developed dimensionality reduction method that allowed us to identify three completely orthogonal sets of orthonormal bases (one for each task epoch) (Elsayed et al., 2016) (see Methods). Each basis set can be seen as a new subspace

optimized to explain the maximum amount of variance among features in a given task epoch, while being explicitly orthogonal to the other basis sets. The inherent near-orthogonality between epochs made it almost trivial to find such optimized subspaces. In fact, the optimized subspaces explained as much variance as the original subspaces (>80% in the 5 dimensional original subspaces and >80% in 5 dimensional optimized subspaces), and their AIs were close to zero (Fig. 2.5d). However, finding a perfectly orthogonal set of subspaces more explicitly allows us to separate out computations performed on the level of the neural population and simplifies the interpretation of the population projections (Elsayed et al., 2016). Specifically, we can visualize the transitions of the feature specific populations activity between task epochs. We projected the activity across the whole trial onto the three dimensional space spanned by the first PCs of each of the optimized task epoch subspaces (Fig. 2.5c). As expected, population trajectories unfolded along a distinct axis during each epoch and rotated towards another axis at transitions between epochs. Transitions were orderly, as the relative distance between features was always preserved (not statistically evaluated). We observed a similar behaviour for the attention task (Supp. Fig. 2.2), although here, non-optimized epoch subspaces seemed less orthogonal, especially between delay 1 and delay 2.

Finally, we expanded our analysis to all times in the trial by computing the AI between subspaces defined in a sliding 400ms window as well as across time windows. Note, in contrast to the data presented in Fig. 2.5, this was now done in a cross-validated manner (see Methods). The resulting within- and across-time-window AIs are displayed in Fig. 2.6 for both tasks. Note the striking resemblance with our decoding results in Fig. 2.4, both within and across time. As mentioned before, in both tasks, the population activity transitioned between nearly orthogonal subspaces for different task epochs. Importantly,

the AI complements our previous decoding approach since it is agnostic to the relative position between features in state space or a given subspace. Rather, it only takes into account the axis that explain most of the variance among features - which also de-noises the data.

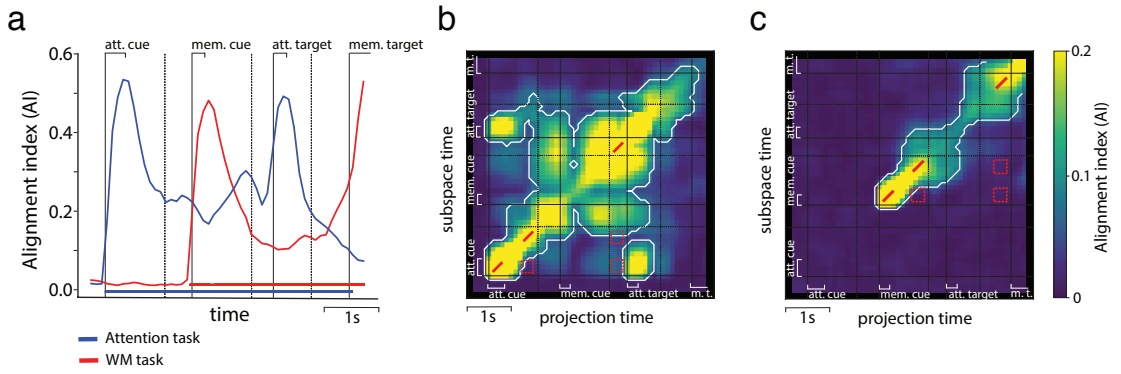


Figure 2.6: Alignment index. a) Within time alignment indices (AIs) for the attention task (blue) and WM task (red) when performed concurrently. Solid lines indicate significant AIs (see Fig. 2.5 & Methods) ($p<0.05$). **b)** & **c)** Across time AI matrices for the WM and Attention tasks. Colours encode the AI magnitude. White contours indicate above chance AI ($p<0.05$). Red bars indicate the time windows used to define epoch specific subspaces in Fig. 2.5 and red dashed contours indicate the corresponding across epoch AIs.

Orthogonal coding spaces between component tasks

Transitions in coding subspaces between the WM cue and delay epochs have been described in lPFC during a range of WM tasks (Meyers et al., 2008; Stokes et al., 2013, Murray et al., 2017; Spaak et al., 2017; Parthasarathy et al., 2017; Wasmuht et al., 2018; Cavanagh et al., 2018). While it is not clear why lPFC displays such radical changes between encoding and mnemonic epochs, a logical consequence of this rotation could be to protect memory representations from incoming sensory information.

The present dual task allowed us to directly test this hypothesis since at various stages, information about one item needs to be encoded, while information about the other item needs to be maintained. For this purpose, we leveraged the AI method described above. In addition to computing the AI within each component task, we also computed the AI between the two tasks, as well as across time (see Methods) (Fig. 2.7a,b). In brief, within a sliding 400ms window we defined a five dimensional subspace for the WM task, and projected the condition averaged activity from the attention task - from the same time windows - into those subspaces and calculated the AI. We then reversed this process by defining subspaces on the attention task and projected activity from the WM task onto them to compute a second set of AIs. Lastly, we averaged across the two sets to obtain the final across-task AI. Interestingly, we found that the across-task AI was close to zero during the entire trial (Fig. 2.7a). That is, there was no variance related to the attention task conditions in the WM task subspace, and vice versa, because they were contained within orthogonal subspaces. This was true when information related to one task was being encoded while the information related to the other task was being maintained, e.g. during memory cue presentation, but also was the case when information about both tasks was being held in memory.

We extended our AI analysis across time (see Methods) to see whether the alignment between attention and WM task subspaces was always close to zero. We found that when information about both tasks was being encoded, subspaces significantly aligned (see box 1 in Fig. 2.7b). That is, the memory cue subspace explained some of the variance between attention conditions during the attention cue period and vice versa. This result clearly supports our initial hypothesis that orthogonal cue and delay subspaces could serve to protect mnemonic information from incoming sensory information.

To further illustrate those results, we projected the condition-averaged activity from the cue periods of both tasks onto the first two PCs of the cue subspaces, defined within the respective tasks (Fig. 2.7c; panel 1). As expected, during encoding, conditions pertaining to one task could clearly be separated in the subspace of the other task. Furthermore, those visualizations suggest that the actual geometry of the task conditions is partly preserved i.e. upper cue locations in the working memory task seemed to group with the upper cue locations of the attention task and vice versa. In contrast, projections onto the first two PCs of the memory cue and the concurrent attention delay subspaces show no obvious relation between location representations across tasks when processed simultaneously (Fig. 2.7c; panel 2).

To further highlight the orthogonality between the task specific location representations, we applied the same dimensionality reduction method described in the previous paragraph (see Methods). The optimized attention and WM task subspaces, which were chosen to be 5 dimensional each, explained as much variance within the respective tasks as the original subspaces (Fig. 2.7e). Figure 2.7d shows the projections onto the space spanned by the first PC from the optimized attention task subspace and the top two PCs from the optimized WM task subspace. Task specific location representations within this reduced space clearly do not interfere with each other. Rotating this space so that the second WM task PC lies perpendicular to the plane of illustration exemplifies this as attention and WM projections spread out along orthogonal axis of that common space.

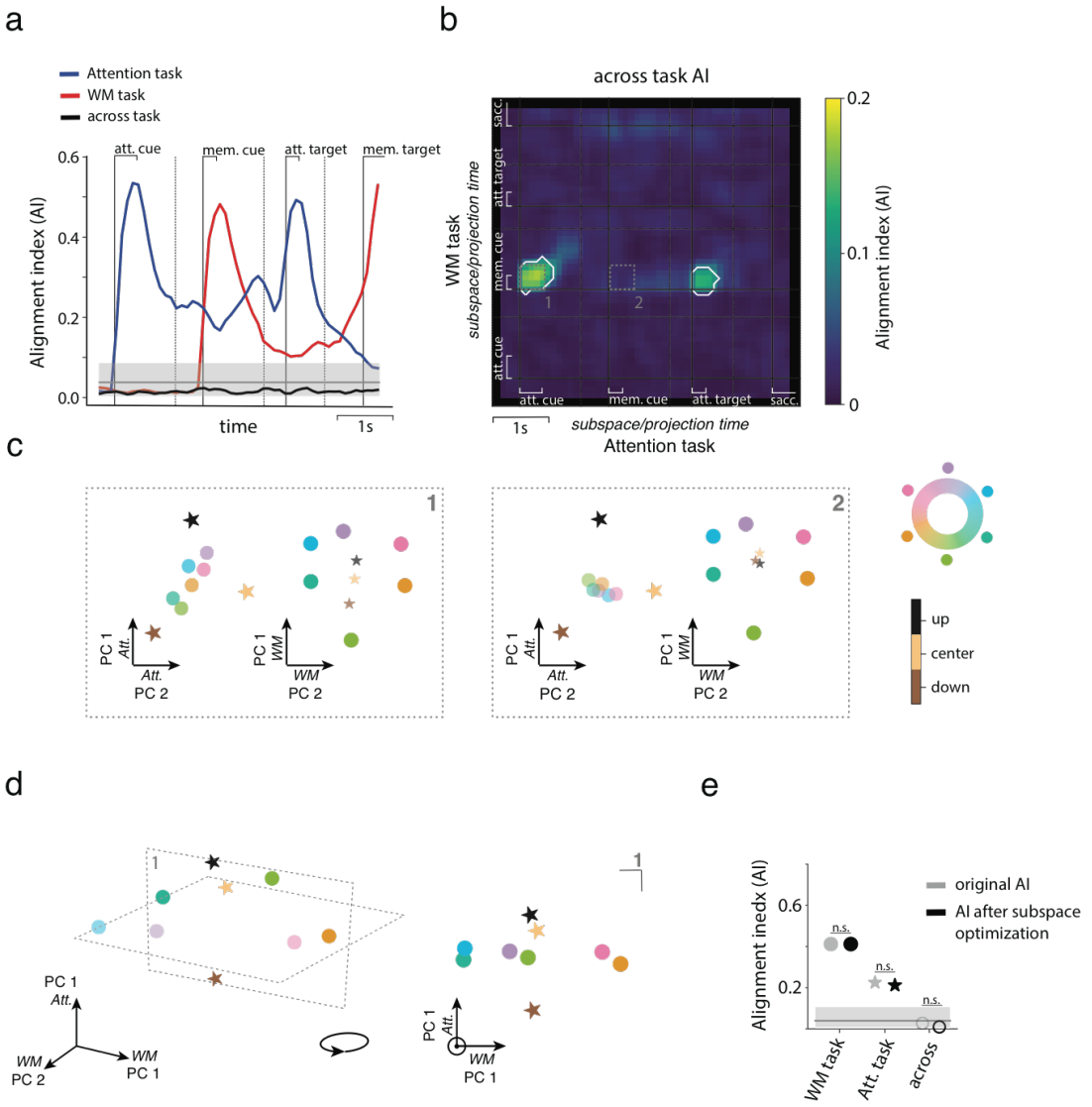


Figure 2.7: Across task alignment index and subspace projections. a) Within time AI for the attention task (blue), the WM task (red) and across tasks (black). The grey solid line and shaded area indicate the mean and 95% confidence interval of the null distribution of the AIs computed between randomly sampled subspaces of the same dimensionality (five) as the subspaces defined within tasks (see Methods). **b)** The across-task AI computed across time. Colour scale indicates the AI magnitude. White contours indicate AIs exceeding the null distribution depicted in **a** ($p<0.05$). Grey, numbered contours refer to the time windows used to compute the projections in **c**. **c)** Condition averaged activity from both tasks (from the respective time window marked in **b**) projected onto the top two PCs of the task specific subspaces (defined on the respective time window marked in **b**). **d**) Projections onto the orthogonal PCs found through the optimization procedure (see Methods). **e)** AI within and between the optimized subspaces compared to the analogous AIs associated with the original subspaces. The grey solid line and shaded area indicate the mean and 95% confidence interval of the null distribution of the AIs computed between randomly sampled subspaces of the same dimensionality as the optimized and original subspaces (see Methods).

All results in Fig. 2.7 were obtained using five dimensional subspaces since those captured >80% of the variance among location conditions in any 0.4s time window from cue onset of the respective tasks. However, it is possible that attention and WM task representations might overlap in less informative dimensions (that is, in the dimensions explaining the remaining ~20% of variance). To test this, we applied two decoding approaches that exploited the full-dimensional neural state space. In the first approach, we trained a parametric classifier on the angles between the WM task conditions and tested how well it could predict the angles between the attention task conditions (see Methods). Within- and across-temporal versions of this classification approach gave similar results to the AI based analysis in Fig. 2.7a,b. Briefly, angles between attention locations could be significantly predicted by the angles between WM locations, but only across time when both attention and memory cues were encoded, or the attention target was presented (Supp. Fig. 2.3a).

Secondly, we also applied an across-task decoding approach that more closely resembled the pattern discriminability score metric used previously for within-task decoding (see Fig. 2.4). Based on the 2D Euclidean distance between locations presented in both tasks, we constructed attention location responses using a weighted average of the WM location responses. Next, we computed the discriminability score between the constructed and real attention location responses (see Methods). Again, within- and across-temporal discriminability scores matched our results from the angular decoding and AI metrics (Supp. Fig 2.3b). Taken together, decoding based on the entire state space confirmed that attention and WM task location representations were close to orthogonal when processed simultaneously (i.e. within time). Furthermore, these approaches statistically validate our observation made in Fig. 2.7c: when subspaces overlapped -

e.g. during attention and memory cue encoding - the original geometrical relation between the location stimuli was preserved. Importantly, the AI metric and decoding approaches complement each other, as a failure to decode would not have ruled out an overlap between task subspaces, and thus, a possibility for interference on the neural level. On the other hand, an overlap in subspaces does not guarantee preserved geometrical relations and comparisons between subspaces do not take into account less informative dimensions in state space.

One reason for orthogonal task subspaces and failure to decode location information across tasks could be due to two distinct subpopulations of neurons, each dedicated to processing locations in one task only. Although such an interpretation seems unlikely given the presence of mixed selective neurons, we tested this possibility explicitly. First, we computed the correlation between the loadings (i.e. weights) on the five dimensional attention and WM subspaces (see Methods). The loading quantifies the contribution of each neuron onto the respective subspace. If the correlation between loadings of the same neurons on the two different subspaces is positive, this means that the same neurons contribute to both subspaces. If the correlation is negative, this provides evidence that task specific computations are carried out by two distinct subpopulations. Correlations between subspace loadings were high even from the beginning of the trial, suggesting that many neurons, by chance, contribute to any two randomly selected subspaces (Supp. Fig. 2.4b). However, correlations increased, displaying a significant positive correlation during memory cue presentation and after attention target offset. Across time, correlations between loadings were highest during cue presentations in both task.

Next, we sought to validate our observations considering responses within the entire state space. To this end, we separately computed the percentage of variance explained (PEV; see Methods and Fig. 2.3) for each neuron in response to attention locations and WM locations, resulting in two PEVs (one for each task) per neuron, per time point. For each pair of time points, we then computed the correlation between the attention and WM PEVs, across neurons (Supp. Fig. 2.4b). Similar to the loading correlations, PEV correlations increased after memory cue presentation, when monkeys entered the dual task, indicating that neurons coding for attention locations also tended to code for WM locations at a given point in time. As did loading correlations, PEV correlations peaked during cue presentations in both task.

Together, those results suggest that rather than two distinct subpopulations, each dedicated to one task at a time, neurons in lPFC - to some extent - participate in both tasks. However, despite the presence of such a mixed signal, task representations do not seem to interfere with each other, but rather reside within orthogonal subspaces within neural state space. populations.

Invariant representations of component tasks

During the dual task, monkeys need to generalize their behaviour in one task across conditions of the other task. For example, they should behave exactly the same way in the WM task, regardless of where the attention cue had appeared in that trial. For simplicity, we describe this as generalization across instances of a task, an instance being a condition in the other task.

To visualize how neural activity patterns for a given task behaved across instances, we plotted the condition specific trajectories of the neural population activity (Fig. 2.8a). More specifically, we focused on the memory cue epoch (0ms – 0.4ms relative to the memory cue), where we had a strong signal for both tasks, and computed a set of five dimensional subspaces, one for each task (optimized for orthogonality, as described before). We then averaged the neural activity for each attention condition for each time point, but as a function of the WM task conditions (WM conditions were reduced to three to improve statistical power; see Methods), resulting in nine conditions per time point, which could be sorted either according to attention or WM. We then projected activity from different time windows onto different combinations of axes from our subspaces to gain some intuition of the trajectories (Fig. 2.8a). Directions of trajectories on the top two WM PCs were clearly determined by WM conditions (denoted by colour in the figure). The reverse was true for the projections onto the top two attention PCs, where trajectories differed only by attention condition (denoted by shape in the figure), but different WM conditions with the same attention condition followed the same trajectory. Projections onto a common space (first WM and first attention PC) complete the picture: trajectories enter this space already slightly separated by attention conditions, which sets the initial conditions from where trajectories unfold, both separating out by attention and WM conditions but without losing their relationship across instances. Note, since task subspaces drastically change across time (Figs. 2.4,2.5 & 2.6), we would gain little in visualizing trajectories for the whole trial.

Based on these observations we sketched out a hypothetical low dimensional representation of the dual task (Fig. 2.8b). Each condition is represented (i.e. separable), but embedded within a parallel structure that allows generalisation. This would allow

the pattern of WM task activity (triangles) to perfectly cross generalize across attention instances, and correspondingly also for the attention task patterns (lines). This also implies that subspaces for both tasks should perfectly overlap across instances. To test the validity of the proposed geometry, we determined five dimensional subspaces for the attention and WM tasks separately for each instance of the other task (in a sliding 400ms window) and computed the alignment between them, yielding an across-instance AI. As an upper bound, we also computed the AI within instances (see Methods) (Fig. 2.8c). The across-instance AI was significantly smaller than within-instance AI, at various time points during the trial and for both tasks ($p<0.05$; *cluster based permutation test*). However, despite those differences, magnitudes of the across- and within-instance AI were similar, suggesting that a large part of the variability among conditions in a given task was preserved across instances. To complement the AI analysis and include information present in full state space we also computed the discriminability scores within and across instances for each task (see Methods) (Fig. 2.8d). This analysis yielded very similar results. The pattern among conditions in the WM task generalized well across conditions of the attention task and vice versa. In fact, for each task, cross-instance discriminability was higher than expected by chance at any time point in the trial and was similar in magnitude to the within-instance discriminability. However, as in the AI-based analysis, during memory cue presentation cross-instance discriminability significantly dropped.

Together, our results largely confirm the proposed geometrical structure underlying the population responses to the dual task, i.e. orthogonal subspaces between tasks and generalization within tasks. However, while we observed a slight cost of cross-task generalization, i.e. dual task interference, this was not reflected in our subspace analysis.

For illustration, consider the schemata in Supp. Fig. 2.5. Adding certain nonlinearities can drastically change the WM pattern generalization across attention instances (or the attention pattern generalization across WM instances) while variation between WM and attention conditions could still be restricted to orthogonal subspaces (Supp. Fig. 2.5). Different kinds of perturbations to the abstracted representations in Fig. 2.8b and Supp. Fig 2.5 could lead to poor generalization between instances as well as to an overlap between component task subspaces.

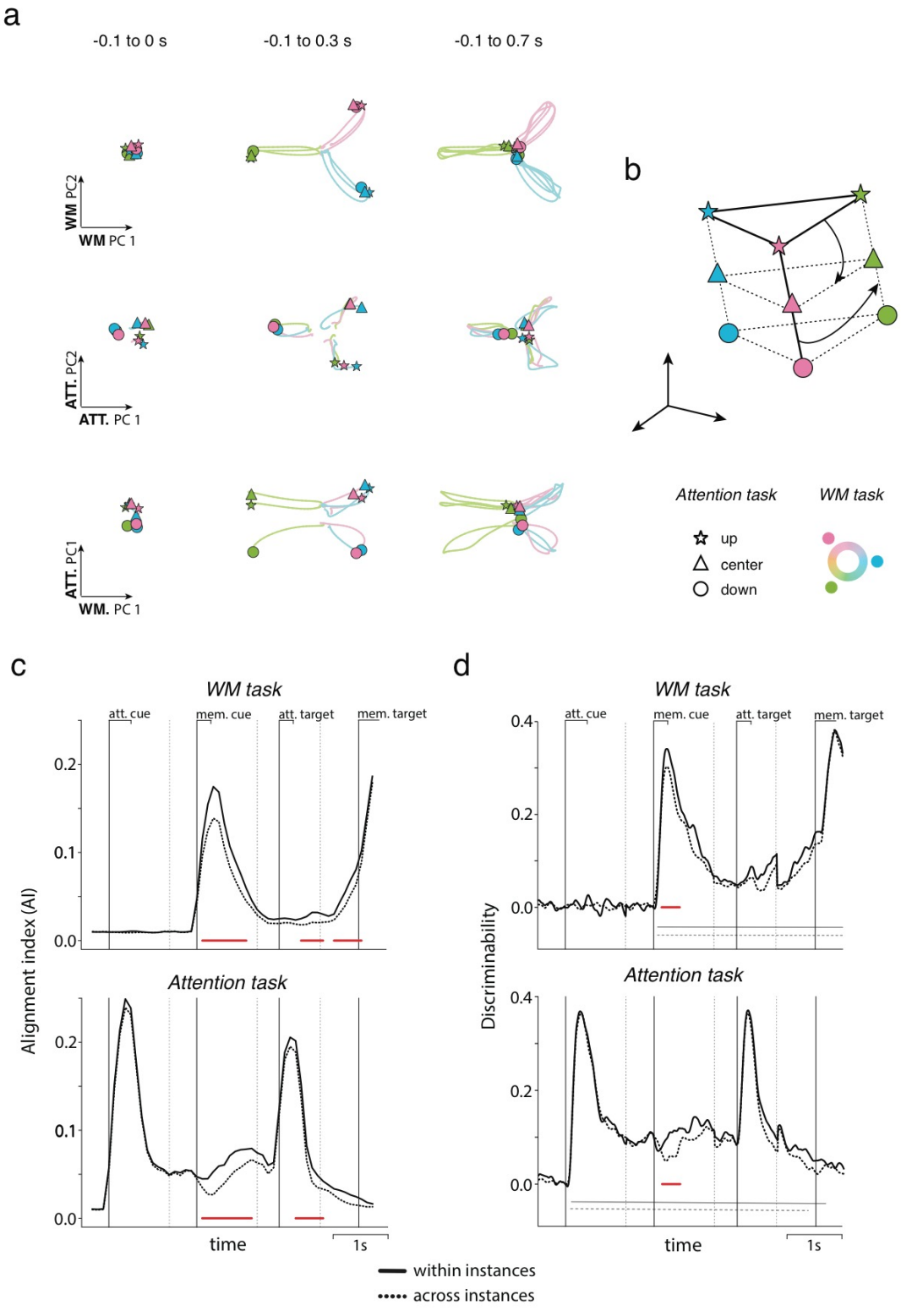

a
-0.1 to 0 s
-0.1 to 0.3 s
-0.1 to 0.7 s
WM PC2
WM PC 1
ATT. PC2
ATT. PC 1
ATT. PC1
WM. PC 1
b
Attention task
WM task
up
center
down
c
WM task
att. cue
mem. cue
att. target
mem. target
Alignment index (AI)
0.2
0.1
0.0
Attention task
time
1s
d
WM task
Discriminability
0.4
0.2
0.0
Attention task
within instances
across instances

Figure 2.8: Cross-instance generalization. a) Trajectories in PCA subspaces. Optimized PCA subspaces were computed in the interval (0-0.4s) aligned to the WM cue onset. Column titles refer to the times of the projected time snippets, relative to the WM cue. Throughout the figure, colours refer to individual WM conditions and shapes to individual attention conditions. The first row shows data averaged across individual WM and attention conditions and projected onto the first two PCs of the optimized WM subspace. The second row shows the same data projected onto the first two PCs of the optimized attention subspace. The third row show the same data projected onto the orthogonal first WM and first attention PCs. **b)** Low dimensional scheme of the proposed population response geometry underlying the dual task. Triangles indicate the WM subspaces as a function of attention conditions (instances) and lines indicate the the attention subspaces as a function of WM conditions (instances). Arrows indicate the directions of generalization. **c)** Within- and across-instance AIs (black solid and dashed lines) computed in a sliding 400ms window for the WM task (upper panel) and attention task (lower panel). Red lines indicate a significant difference between the black solid and dashed lines ($p<0.05$; *cluster based permutation test*). **d)** Discriminability scores within and across instances (black solid and dashed lines) for the WM task (upper panel) and attention (lower panel) task. Horizontal red lines indicate a significant difference between within- and across-instance scores ($p<0.05$; *cluster based permutation test*). Horizontal solid and dashed grey lines indicate significant discriminability scores ($p<0.05$; *cluster based permutation test*).

Simulated populations

We used simulated data based on idealized neural populations to gain a clearer intuition into how pure, linear and nonlinear mixed selectivity affect subspace overlap across tasks as well as the cross-instance generalization within tasks. To this end, we separately simulated the activity of six different populations during a simplified version of the dual task (see Methods): 1) Pure selectivity neurons. 2) Common coding neurons (tuned to same location in both tasks). 3) Common coding neurons with saturation, leading to non-linear responses. 4) Neurons with LMS, but random preference. 5) Neurons with NLMS. 6) Empirical population based on the estimate in Fig. 2.3c.

As expected, the pure selectivity population exhibited orthogonal subspaces i.e. an AI at effectively zero for attention and WM tasks. This can be seen in Fig. 2.9a, where the attention subspace explained (by definition) the highest possible amount of attention-related variance, but none of the memory-related variance (and vice versa). This is relatively straightforward: attention and WM locations were encoded in separate neural populations. A similar structure can also be observed for mixed selectivity cells,

including linear and non-linear, except if there was any correspondence/overlap in coding preference between component tasks within individual cells. Indeed, the common coding population exhibited perfect subspace overlap, reflecting the fact that all cells coded for the same location in the attention and WM tasks. This was true even when cells saturated their firing (i.e. firing rates were not additive when the same location was encoded for both tasks, a form of NLMS). As expected, orthogonal subspaces could be seen in the empirical population but also the nonlinear mixed population.

We subsequently analyzed the WM cross-instance discriminability of the different populations (Fig. 2.9b). In general, cross-instance discriminability was maximal in simulated sub-populations of all cell types, except in the presence of NLMS cells. Specifically, there was no cost in cross-instance generalization for pure selectivity cells, or LMS cells or the common coding population. By contrast, the NLMS population showed no cross-instance discriminability at all, since here tuning for WM was independently determined for each attention condition. The common coding population with saturation also showed a cost of cross-instance discriminability, although it remained well above chance. Similarly, across-instance discriminability in the empirical distribution was lower than within-instance discriminability, but above chance.

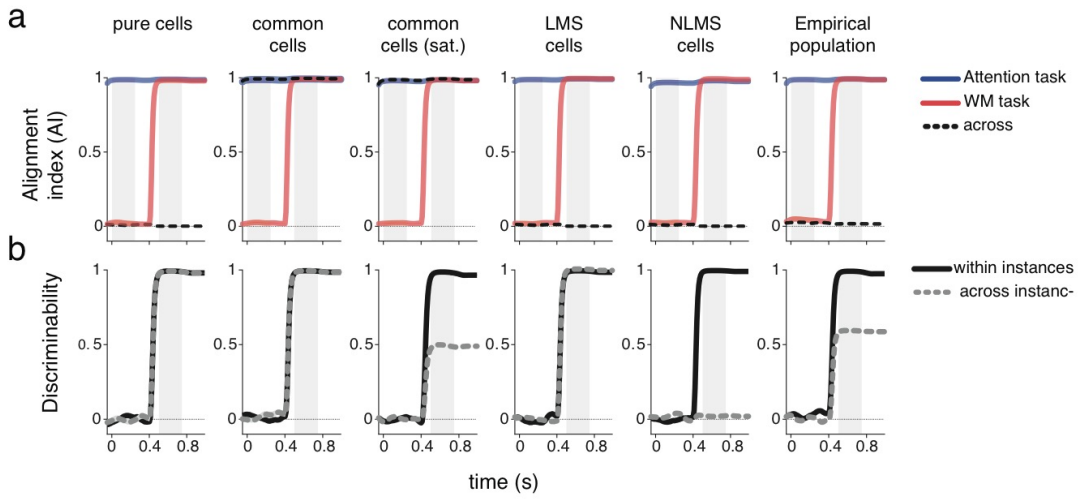
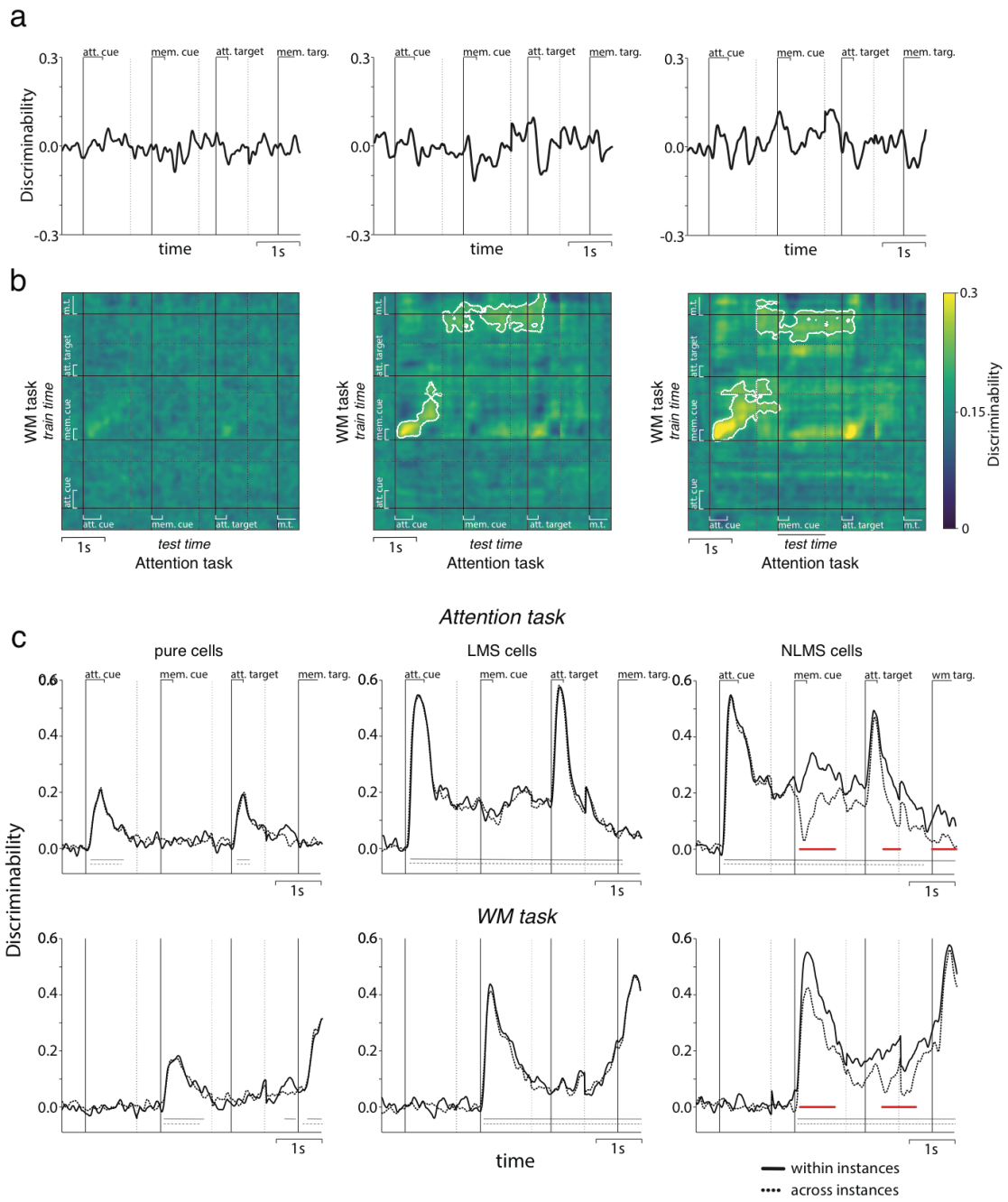


Figure 2.9: Simulated populations during simplified dual task. Throughout the figure, shaded bars indicate the presentation of the attention and memory cues **a)** The AI within the attention task (blue), within the WM task (red) and across tasks (black, dashed), for each of the simulated populations. **b)** Within- and cross instance discriminability (black solid and grey dashed lines) for the WM task in all six simulated populations.

The simulations broadly confirmed our results and singled out neurons with NLMS as a possible cause for the observed cost in cross-instance generalization. Importantly, the simulations also confirmed that NLMS does not cause the attention and WM task subspaces to overlap. Finally, we grouped the real neural population into sub-populations roughly matching those from the simulations. Note, we did not have sufficient statistical power to distinguish between NLMS caused by saturation or other factors and did not separate out common coding from the rest of LMS cells. First, we tried to decode attention locations from the WM locations using the across-task discriminability score metric (as in Supp. Fig. 2.3b; the AI metric gave similar results). In accordance with the simulations, none of the sub-populations showed across-task discriminability at any time point in the trial (Fig. 2.10a). Both LMS and NLMS cells showed across-task discriminability when comparing the attention and WM encoding periods (Fig. 2.10b). Also within those cells, WM coding during the delay before the

attention target weakly generalized to the attention delay coding pattern. Next, we repeated the cross-instance discriminability analysis on all available subpopulations (pure, LMS & NLMS) (Fig. 2.10c). As expected, pure and LMS populations did not show any cost of across-instance generalization. NLMS cells showed a substantial cost of cross-instance discriminability, as was suggested by the simulations. In general, pure selectivity cells showed very weak discriminability scores in both tasks, while mixed selectivity cells carried most of the feature related information in both tasks.



Supplementary Figure 2.10: Across-instance and across-task discriminability for different subpopulations. Pure selectivity subpopulation (n=268); LMS subpopulation (n=235); NLMS subpopulation (n=198) **a)** Within- and cross instance discriminability (black solid and grey dashed lines) for the WM and attention task in three subpopulations. Horizontal red lines indicate a significant difference between within- and across-instance scores ($p<0.05$; *cluster based permutation test*). Horizontal solid and dashed grey lines indicate significant discriminability scores, respectively ($p<0.05$; *cluster based permutation test*). **b)** Across-task discriminability scores for each time point in the trial. **c)** Across-task discriminability scores for each pair of time points i.e. across time. Colours indicate score magnitude. White contours indicate a significant discriminability score ($p<0.05$; *cluster based permutation test*).

2.5 Discussion

In this study, monkeys learned to actively process two simultaneous streams of information i.e. to dual task. They achieved this to a very high standard, showing almost no interference between the two tasks on the behavioural level. There was no apparent behavioural cost associated with performing the component tasks together, relative to single-task performance. At the neural level, individual neurons showed different types of mixed selectivity to features from both task, clearly indicating the potential for interference. Despite mixed selectivity, we found that coding sub-spaces for component tasks were orthogonal to each other. Hence, variation along the axis representing one task did not project onto the axis representing the other task, consistent with lack of dual task interference in the monkeys' behaviour. Furthermore, task specific representations generalized well across contexts, that is, over the conditions of the concurrent other task.

Context dependent coding in lPFC

Neural representations in lPFC are known to be highly context dependent (Fuster, 2008; Miller & Cohen, 2001; Rigotti et al., 2013; Mante et al., 2013; Stokes et al., 2013). For example, Mante et al. (2013) analysed data from a decision making task using coloured moving dots. Depending on the context, monkeys had to base their response on either the colour or the motion direction of the dots. The researchers found that when, for example, motion was relevant, population responses to motion direction aligned with a decision related subspace, but not when it was irrelevant (but vice versa when colour was relevant) (see also Donahue et al., 2015). The orthogonal subspaces for the attention and WM tasks are an instantiation of such context dependent dynamics in lPFC, because

here, monkeys had to decide whether to release a lever or where to make a saccade. In contrast to the study from Mante et al. (2013) here location was relevant in both contexts and could not be dismissed for one task or the other. In that sense, our findings expand on those in Mante et al. (2013) by showing that context dependent processing in lPFC extends to situations where all inputs are currently relevant. Interestingly, in both studies location (task A vs. task B) and motion direction (relevant vs. irrelevant) representations initially spread out within a common subspace, possibly reflecting a less contextualized sensory signal within lPFC (Mendoza-Haliday, 2014), which we will discuss in more depth below.

The orthogonality between task subspaces in lPFC is compelling for several reasons. First, if task subspaces would overlap, the attention cue representation could bias the memory cue representation and vice versa, which could lead to cross-talk at the behavioural level. In a more extreme case, attention target presentation could lead to a saccade, or memory cue presentation to a lever release. This of course depends on how the activity in lPFC is read out, for what purpose and by which region. A recent study in mice (Miri et al., 2017) found that two behaviourally distinct tasks (pulling and walking) were associated to orthogonal activity patterns in motor cortex leading to a distinct engagement of downstream pathways i.e. read-out circuits between behaviours. Given the connectivity and information flow between more frontal regions and motor cortex (Fuster, 2001, Yeterian et al., 2012; Markov et al., 2014), a fundamental property of lPFC might be to generate such separable representations and shape them for flexible readout (Mante et al., 2013). Indeed, a recent study imaging human fronto-parietal networks found that component task representations became more dissimilar with

training, reflecting a better performance during multi-tasking (Garner and Dux, 2015; Duncan and Mitchell, 2015).

A reason for the difference in neural codes for the two tasks could be their distinct demands on types of memory systems. During the delayed saccade task (WM task), the required motor response is known and thus an action might be planned and inhibited during the delay. Therefore, it is likely that premotor signals contribute to the delay activity pertaining to this subtask (see Tang et al., 2019 for a similar task and a dissociation of sensory WM and motor preparation signals that strikingly match the coding dynamics presented here). Although we termed the delayed saccade task the WM task for the sake of simplicity and because of its predominance in WM related studies it is important to mention that motor preparation is at least a very special case of WM and a proposed system for WM needs to 'hold in mind' information without motor preparation. In contrast, in the lever release task (attention task), a sensory representation has to be maintained on top of any motor related activity in order to correctly detect a change in stimulus location. Together, these two particular tasks tap into mostly non-overlapping cognitive functions served by possibly very distinct underlying mechanisms for motor planning and attention/sensory detection, which will be discussed further in the general discussion of this thesis. In the current study, it is hard to disentangle whether orthogonal task representations were caused by distinct task demands or types, employment of attentional resource, an interaction with an abstract state variable or simply differences in timings. In the following we will discuss the latter.

Changing code across epochs

The design of the dual task caused component tasks, although performed simultaneously, to be in different stages at any given time point. That is, when the memory cue was encoded, attention information was maintained and vice versa when the attention target was presented. Crucially, we showed that the location coding patterns for each task transitioned between different formats across time, and they were explicitly orthogonal across epochs, possibly reflecting different computations at different time points. Transitions in coding subspaces between the WM cue and delay epochs have been described in lPFC during a range of WM tasks (Sigala et al., 2008; Meyers et al., 2012; Stokes et al., 2013, Murray et al., 2017; Spaak et al., 2017; Parthasarathy et al., 2017; Wasmuht et al., 2018; Cavanagh et al., 2018). Here, we complemented those observations by extending the common framework in which such data are analysed and specifically highlight the orthogonality between coding patterns. A similar phenomenon has been described in motor cortex (Kaufman et al., 2014; Elsayed et al., 2016), where the population patterns between preparatory and movement activity lie in orthogonal subspaces, such that preparatory can optimally set the stage for movement execution, without already initiating it (Churchland et al., 2010). Similarly, activity in rodent posterior parietal cortex (PPC) transitions between orthogonal subspaces for decision and movement phases (Raposo et al., 2014). However, it is less clear why lPFC displays such radical changes between encoding and mnemonic epochs during WM. Here, we show that if new information reaches lPFC, it enters into an encoding subspace, which was nearly orthogonal to the mnemonic subspace and so, did not directly interfere with the currently maintained information. Additionally, during memory cue presentation, the active mnemonic subspace for the attention task seemed to further rotate so that it was close to orthogonal from all other subspaces at all time points. Our observation provides an alternative to the proposed mechanisms of distractor suppression during

WM in lPFC, which can be ambiguous in how far a distractor is processed during WM (Sakai et al., 2002 (in humans); Qi et al., 2010; Lennert and Martinez-Trujillo, 2011; Suzuki and Gottlieb, 2013, Mante et al., 2013, Jakob and Nieder, 2015, Parthasarathy et al., 2017; Cavanagh et al., 2018). If incoming information lies in a different subspace than the delay activity, distracting input might be allowed into PFC for some time without creating interference. Potentially, such a mechanism could provide a necessary time window for lPFC to decide whether to subsequently discard (Jakob and Nieder, 2015; Parthasarathy et al., 2017), or how to appropriately integrate the new input (Mante et al., 2013).

In our task, attention and WM subspaces were also orthogonal when both tasks were being maintained (after the memory cue offset), and therefore - according to our previous rationale - increasing the possibility for interference. However, this account is complicated by the fact that when both tasks had to be maintained, the WM task was in its first delay i.e. before attention target onset while the attention task was already in its second delay i.e. after memory cue onset. In fact, Supp. Figs. 2.3b show that when both task were in their first delay period, right after cue presentation, there was weak but significant across-task decoding using the discriminability metric. Also, the pattern for the attention task seemed to undergo a change between the early part of its first delay and the later part where it more strongly resembled the cue epoch pattern (see e.g. Fig. 2.4). In total, we cannot rule out a continued non-linear interaction with time (i.e. rotation in state space) that caused the two task subspaces to be orthogonal whenever they concurred. Additionally, during the respective cue epochs, where signal strength for both tasks was highest, across-task AI was about 50% of the within task AI.

Therefore, when within task AI decreased during the delay period, it is possible that any across-task AI would fall to chance level, simply because of changes in magnitude.

Mixed codes

Intuitively, an easy way to solve the problem of overlapping representation and therefore neural interference during the dual task would be to have two distinct sub-populations, each representing only one task. Here we found little evidence that this was the case. While there was a large proportion of cells that were exclusively selective to one of the tasks, a considerable amount of cells coded for linear and non-linear mixtures of the task variables. Furthermore, cells that contributed location information in one task also tended to do so in the other. Few electrophysiological studies could show non-overlapping neural sub-populations for different component tasks (Marcos et al., 2017), and to our knowledge, none have involved performing them concurrently. In general, our findings are in accordance with what is expected in lPFC and other high level cortices, mainly, neurons that respond to complex mixtures of many behaviourally relevant variables (Singh & Eliasmith, 2006; Mante et al., 2013; Pagan et al., 2013; Meister et al., 2013; Park et al., 2014; Rigotti et al., 2013), which often defies functional categorizations based on low dimensional tuning curves (Raposo et al., 2014).

Studies investigating dual task scenarios in rats (Pang et al., 2001) and monkeys (Lebedev et al., 2004; Messinger et al., 2009; Watanabe and Funahashi et al., 2014) made observations very similar to ours. Specifically, in a task that involved simultaneously maintaining and attending to two spatial locations, Lebedev et al. (2004) and Messinger et al., 2009 reported proportions of pure selectivity and mixed selectivity

cells (linear and non-linear) that closely matched the ones found here. Furthermore, Messinger et al. (2009) showed that first, LMS cells (termed multitasking cells) had a stronger tuning to both the remembered and attended locations and second, represented the task variables more efficiently (i.e. fewer neurons were needed to confer the same amount of information). The former observation aligns with our result, showing that LMS cells clearly carried more location information in both attention and WM tasks than pure selectivity cells (Fig. 2.10c). Note, we did not explicitly test the latter observation, and furthermore, our estimation of neural sub-populations was relatively coarser (a cell was assigned a group membership if it displayed a given characteristic anywhere during the trial; see Methods).

Recent theoretical work proposed a related role for NLMS, which was shown to increase coding reliability and efficiency (Johnston et al., 2019) as well as the number of discrete stimuli that could be reliably represented (Litwin-Kumar et al., 2017; Alemi et al., 2017). Although we did not test those predictions explicitly, we found that cells in our NLMS sub-population carried the most location information in both tasks, of all the tested sub-populations. However, the estimation of NLMS in an experimental setting might be complicated through low signal to noise ratios (SNRs) and low trial numbers, which might lead to an underestimation of the relative proportion of NLMS neurons. Conversely, the detected NLMS neurons might therefore also carry more information (higher SNR). Nonetheless, our results and those from the studies discussed above show that mixed selectivity cells are commonly activated by dual tasking and are not easily dismissed. On the contrary, information content is highest in this sub-population. Importantly, here we show that despite the prevalence of a mixed signal on the level of the individual neurons task representations are orthogonal on the level of the population.

Using simulations, we emphasize that such orthogonality is achieved quite easily in LMS and NLMS populations as long as preference for each sub condition (e.g., spatial position) are assigned randomly for each task. This is good news, because it allows parallel representation of multiple variables within the same population, which emphasizes read-out mechanisms that are specific to a variable and do not rely on selective readout from a specific sub-population (Knutsen and Ahissar, 2009).

General codes

We found that the pattern of location information in a given task generalized across conditions of the other concurrent task. Such a generalization across instances might be beneficial for efficiently solving the dual task since it implies that the monkeys could rely on a fixed read-out mechanism for a given task: e.g., for the WM task, the subject does not have to learn separate readout patterns for each sub condition of the attention cue location. However, generalisation was not perfect. Indeed, the cost of cross-generalisation (in NLMS cells) was the only signature of dual-task interference that we observed on the level of the neural population. In addition to its benefits in coding reliability and efficiency described above, NLMS has been shown to increase the dimensionality of neural representations in lPFC, thereby facilitating a linear readout of arbitrary response combinations (Rigotti et al., 2013). In other words, a high dimensionality can be important to diversify behavioural output in response to the same stimuli and thus enable context dependent behaviour (Rigotti et al., 2013; Lindsay et al., 2017). Here we suggest that NLMS is detrimental if instead the goal is to form a representation that can be generalized across contexts (describes as instances in our results). Note, this does not necessarily imply information about context is lost; a high-

dimensional representation can be maintained along parallel planes of abstraction (see Fig. 2.8b). This is consistent with recent evidence for high dimensional representations that support abstract generalisation in lPFC (Bernardi et al., 2018).

Across-instance generalization was not perfect which raises the important question: is NLMS beneficial in this task (Johnston et al., 2019, Finkelstein et al., 2018; see preceding paragraph) or can it simply not be avoided (Enel et al., 2016)? A recent study by Cavanagh et al. (2018) reported complete absence of NLMS between the locations of a spatial memory cue and the identity of a concurrent reward cue (always presented at the centre of the screen), which suggests that NLMS is not an unavoidable property of lPFC, but note the statistical problem in estimating NLMS experimentally (discussed above). However, in the dual task, cues from both tasks were spatial and importantly, both were strictly relevant. Neurons in lPFC seem to be predominantly activated by spatial stimuli (Funahashi et al., 1989; Constantinidis et al., 2001; Lara et al., 2014) and spatial receptive fields that can cover large parts of the visual field (Lennert and Martinez-Trujillo, 2011; Bullock et al., 2017). Additionally, a recent study showed that spatial receptive fields in lPFC exhibit a somewhat preserved anatomical and task general structure (Kollmorgen et al., 2019). Together, this suggests that if two spatial stimuli are presented together (as in our task) some form of response saturation/normalization (Carandini and Heeger, 2012; Duong et al., 2019) might be hard-coded into the lPFC network, inevitably leading to NLMS (as shown in our simulations). Taken together, our observation and those from other studies (e.g. Rigotti et al., 2013; Bernardi et al., 2018) suggest that population activity in lPFC might be the result of a task specific trade-off between high dimensional and generalizable representations.

Capacity limits and mutual attenuation

A recent study, using an almost identical dual task to the one described in the present study, reported that WM delay activity was attenuated by the concurrent attention task and depended on its difficulty (Watanabe and Funahashi, 2014). Similar neural attenuation effects between multiple presented stimuli have been observed in tasks used to investigate WM capacity limitations (Buschman 2011, Matsushima and Tanaka, 2014; Tang et al., 2019). In this experiment, we did not observe any dual task cost at the behavioural level, and therefore did not focus on signal attenuation. Instead, our present question focused on the streams of information that were detectable within the network and how they interfered with each other.

Conclusion

In conclusion, we showed that monkeys can perform a dual-task without showing obvious interference between the component tasks. This lack of interference was reflected in the population activity as orthogonal subspaces for the attention and WM task. While we cannot say for certain what caused this orthogonality, we suggest that it could be further supported by the dynamic nature of encoding in lPFC - that is changing codes over time - which on many occasion permit parallel processes to unfold in the same network. Additionally, we found that task-specific representations generalized well across conditions of the concurrent task. This led us to propose an ideal neural representation of the dual task that exploits the orthogonality between the two tasks while ensuring parallel and therefore generalizable representations across instances.

However, any causality between the neural representations and the animals' behaviour remains to be uncovered and will be discussed more thoroughly in the general discussion of this thesis.

Chapter 3. Intrinsic neuronal dynamics predict distinct functional roles during working memory*

3.1 Abstract

Working memory (WM) is characterized by the ability to maintain stable representations over time; however, neural activity associated with WM maintenance can be highly dynamic. We explore whether complex population coding dynamics during WM relate to the intrinsic temporal properties of single neurons in lateral prefrontal cortex (lPFC), the frontal eye fields (FEF) and lateral intraparietal cortex (LIP) of two monkeys (*Macaca mulatta*). We found that cells with short timescales carried memory information relatively early during memory encoding in lPFC; whereas long timescale cells played a greater role later during processing, dominating coding in the delay period. We also observed a link between functional connectivity at rest and intrinsic timescale in FEF and LIP. Our results indicate that individual differences in the temporal processing capacity predicts complex neuronal dynamics during WM; ranging from rapid dynamic encoding of stimuli to slower, but stable, maintenance of mnemonic information.

*The experiment in this chapter has been published as: Wasmuht, D. F., Spaak, E., Buschman, T. J., Miller, E. K., Stokes, M. G. Intrinsic neuronal dyanmics predict distinct functional roles during working memory. *Nat.Comms.* **9**, 3499 (2018)

3.2 Introduction

Single neuron dynamics in higher cortical areas are heterogeneous, complicating interpretation of their functional roles during cognitive tasks (Rigotti et al., 2013; Raposo et al., 2014). A prominent example illustrating this principle is working memory (WM), a mental process strongly associated with the prefrontal cortex (PFC) (Funahashi et al., 1989; Miller et al., 1996; Fuster, 1997). During WM, information about a transiently encoded stimulus needs to be stored and kept available over a short delay before a response can be made (Fuster and Alexander, 1971; Goldman-Rakic, 1995). The neural mechanisms by which this kind of cognitive stability is achieved remain a matter of scientific debate (Barak and Tsodyks, 2014; Stokes, 2015; Zylberberg and Strowbridge, 2017).

A large body of experimental studies emphasize a coding scheme whereby information is maintained through persistent firing of single neurons (Funahashi et al., 1989; Chafee and Goldman-Rakic, 1998; Rainer et al., 1998; Romo et al., 1999). This view is supported by theoretical models demonstrating that persistent activity can emerge from intrinsic cell properties (Loewenstein and Sompolinsky, 2003; Fransén et al., 2006) and/or reverberations in recurrently connected populations of selective neurons (Amit and Brunel, 1997; Compte et al., 2000; Wang, 2001; Machens et al., 2005; Wong and Wang, 2006). However, other studies highlight heterogeneous temporal tuning profiles in a majority of recorded neurons (Brody et al., 2013; Shafi et al., 2007; Enel et al., 2016) or distinct bursts of activity (Lundqvist et al., 2016), resulting in a highly dynamic population code underlying WM (Meyers et al., 2008; Stokes et al., 2013; Spaak et al., 2017). Supported by theoretical models (Maass et al., 2002; Mongillo et al., 2008;

Druckmann and Chklovskii, 2012), those observations have led to the view that maintenance of stable mental representations in WM is possible in the absence of persistent activity in single cells, either through coordinated transient dynamics (Druckmann and Chklovskii, 2012) or rapid connectivity changes (Mongillo et al., 2008; Rose et al., 2016).

Given the frequent observations of both persistent and dynamic coding at both the population and single neuron level, it is unlikely that these ideas are mutually exclusive (Barak and Tsodyks, 2014; Barak et al., 2013). In fact, recent studies have shown that stable population coding can coexist with heterogeneous neuronal dynamics (Spaak et al., 2017; Murray et al., 2017). While those studies stress stable population coding despite overall neuronal dynamics, critically, both regimes interact during categorizations performed by monkeys during WM tasks, relying on both persistent and dynamic neurons (Chaisangmongkon et al., 2017). Both transient and persistent activity in single neurons seems to be important for WM. However, the source and function of the neuronal tuning and temporal variability underlying WM population dynamics remain poorly understood.

One of the most striking principles observed across the mammalian cortex is the hierarchical organization of receptive field size. While this phenomenon has been well studied in the visual system, where spatial receptive field size increases with functional hierarchy (Hubel and Wiesel, 1962), a similar principle is currently being uncovered for the temporal domain (Chen et al., 2014). Murray et al. (2014) estimated decay time-constants for neurons' spiking autocorrelation during baseline 'resting' activity, coined 'intrinsic timescale', across areas in monkey cortex. They reported that intrinsic

timescales increased along the cortical hierarchy. This observation was further supported by evidence from previous studies in monkeys and mice (Ogawa and Komatsu, 2010; Runyan et al., 2017) and results from human imaging studies Hasson et al., 2008; Honey et al., 2012). Importantly, a recent modelling study demonstrated that the observed gradient of intrinsic timescales arises naturally in a large scale brain model through the balance of intra- and inter-area connection densities (Chaudhuri et al., 2015).

Intrinsic timescales can be interpreted as the duration over which cells integrate information (Goldman et al., 2008). According to this view, shorter timescales in sensory areas promote rapid detection of dynamic stimuli (Buračas et al., 1998; Salinas et al., 2000), increasing the temporal dimensionality of neural coding. Conversely, longer timescales in prefrontal areas could support integration of information over longer periods of time and could improve signal to noise ratio needed for decision making or WM (Goldman et al., 2008; Wang, 2013), but at the cost of dimensionality in the temporal domain. Consequently, heterogeneity in intrinsic timescales found within a cortical area might reflect functional specializations of individual neurons according to temporal processing demands. In line with this, a recent study by Scott et al., (2017) showed that the fronto-parietal network encodes accumulated evidence with a diversity of neuronal timescales. A similar observation was made by Bernacchia et al. (2013), who found a reservoir of timescales for reward integration across cortical neurons. Furthermore, orbitofrontal cortex (OFC) neurons associated with a relatively long intrinsic timescale where shown to carry chosen value signals over longer time periods than their short timescale counterparts (Cavanagh et al., 2016). Similarly, Nishida et al. (2014) showed that a more stable baseline activity was associated with

higher firing rates during the delay period of a WM task in lateral intraparietal cortex (LIP).

Here, we explore how the temporal stability of individual neurons at rest (i.e., pre-trial fixation period) relates to the heterogeneity in population dynamics observed during the subsequent WM trial. We tested whether a neuron's intrinsic timescale determines its functional role during WM in three brain regions: the lateral PFC (lPFC), the frontal eye fields (FEF) and the LIP. We found that neurons associated with a relatively long intrinsic timescale carried more information about task relevant features than short-timescale neurons. Importantly, in prefrontal areas, long-timescale cells also carried information in a more stable way. This was observed for each task epoch, and was especially striking during the delay period, for both individual cells and at the population level. Additionally, lPFC cells with shorter intrinsic timescales signalled item information earlier during memory encoding and showed richer dynamics during the delay period, suggesting specific functions for different cells i.e. their weighting during individual epochs of a WM task. Lastly, we found that functional connectivity at rest was correlated with intrinsic timescales in FEF and LIP but not in PFC, potentially implicating different mechanisms determining a neuron's temporal processing characteristics.

2.3 Methods

Experimental paradigm and recordings

Physiological recordings have previously been reported in detail (Buschman et al., 2011). In brief, two adult rhesus monkeys (*Macaca mulatta*) were trained to perform a delayed change localization task (Fig. 3.1a). After a short fixation period (500ms) an array of colored squares was presented for 800ms (the sample period), followed by an 800-1000ms memory delay (delay period). Finally, a test array was presented to the animal. The test array was identical to the sample array except that one randomly chosen item changed color. To receive reward, monkeys had to make a saccade to the changed item. On each trial, the total number of items on the screen varied between two and five. There were six possible item locations and two possible color values per location (changing each session). Over several sessions, simultaneous recordings were taken from single neurons in the lateral prefrontal cortex (lPFC; 584 cells total); the frontal eye fields (FEF; 325 cells) and the lateral intraparietal cortex (LIP; 284 cells).

Preprocessing

We excluded cells with fewer than 100 spikes per session from all analysis, hence reducing total cell counts per area to: 583 cells for lPFC, 323 cells for FEF and 281 cells for LIP. There were at least 20 trials per relevant stimulus condition. Unless otherwise noted, binary spike trains were convolved with a Gaussian kernel (s.d. = 20 ms) to produce firing rates in sp/s. For the crosstemporal discriminability analysis, firing rates were down-sampled from 1000 to 100 Hz, allowing for more efficient computing, while maintaining sufficiently high temporal resolution. All data analysis was implemented in Python using custom-written code.

Statistical testing

Throughout this study, we used a non-parametric cluster-based permutation test (Maris and Oostenveld, 2007). In brief, the method compares some observed test statistic with a constructed null distribution while controlling for multiple comparisons across time. Null distributions were constructed in three ways: (1) By randomly permuting condition labels within cells 1000 times (ωPEV and across time discriminability analysis); (2) By randomly permuting intrinsic timescales across cells 1000 times for comparing median split intrinsic timescales (i.e. t-test or raw differences) and correlation coefficients. (3) By randomly computing correct and error labels between trials for the correct versus error trial analysis. In general, the relevant test statistic or raw difference was computed for the observed data as well as for each of the 1000 permutations. This was done for each individual time point or pair of time points (across time discrimination analysis). For (1) and (3), points were classified by comparing the observed effect to the 95^{th} percentile of the null distribution. Contiguous points exceeding the 95^{th} percentile were deemed candidate clusters. For (2), cluster candidates were directly derived from the primary test statistic (contiguous analytical p-values < 0.05 from t-test or correlation analysis). To correct for multiple comparisons, we compared the maximum summed cluster test statistic of the observed data to a new null distribution of maximum summed clusters derived from our permutations. If the size of the observed cluster exceeded the 95^{th} percentile of the new null distribution it was deemed significant.

Estimation of intrinsic timescales

Intrinsic timescales (τ) for single cells were estimated from the 500ms fixation period (Murray et al., 2014). To calculate the temporal autocorrelation, spikes were counted in

ten 50ms successive, independent time bins. This procedure created a trial by time-bin matrix. The temporal autocorrelation is the Pearson correlation across trials, between spike counts from each specific time-bin; i.e. we computed the correlation of each column of the matrix with each other. We fitted with an exponential decay function to the resulting autocorrelation time-course using non-linear least squares, as implemented by Levenberg-Marquardt algorithm within ScPy's optimize_curve_fit function (Murray et al., 2014) (Fig. 3.1b; Eq. 3.1).

$$R(k\Delta) = A\left[exp\left(-\frac{k\Delta}{\tau}\right) + B\right] \quad (3.1)$$

where τ is the intrinsic timescale, A is the amplitude and B is the offset parameter. The $k\Delta$ parameter refers to the relative time lag between time bins (50 – 450ms). Similar to previous studies (Murray et al., 2014; Cavanagh et al., 2016), a fraction of cells showed relatively low autocorrelation values at short time lags, possibly due to refractory period or negative adaptation. To accommodate for that feature, fitting started at the first reduction in autocorrelation.

Exclusion criteria: Cells were not assigned an intrinsic timescale if they had: (1) less than 50 trials; (2) a fixation period firing rate lower than 1 spikes/s; (3) no spikes within any time bin across all trials or (4) first reduction in autocorrelation later than a time-lag of 150ms. Additionally, cells that were clearly not fit well by an exponential function, as determined by blinded visual inspection, were also not assigned a τ value. The remaining count of cells that were assigned a τ value and therefore were available for subsequent analysis was: 265 cells for lPFC, 168 cells for FEF and 136 cells for LIP.

Information content of single cells

We based our estimation of single cell information content on the percentage of variance explained (ωPEV) statistic (Buschman et al., 2011; Olejnik and Algina, 2003; also used in **Chapter 3**, albeit differently). Specifically, we were interested in single cell selectivity to location and/or color information. First, we created a dummy coded design matrix where each column represented a stimulus condition, i.e. location one (1/6) and color one (1/2). For each trial, a one denoted that that specific condition was met whereas a zero denoted the opposite. We then performed a multiple linear regression with the above-defined conditions as independent factors and the firing rate as a dependent variable. Finally, we calculated ωPEV to quantify the modulation of individual firing rates accounted for by the combination of color and location of the presented stimulus items.

$$\omega^2 = \frac{SS_{Between\ groups} - df * MSE}{SS_{Total} + MSE} \qquad (3.2)$$

where SS_{Total} denotes the total sum of squares i.e. total variance, $SS_{Between\ Groups}$ denotes the variance between groups, *df* are the degrees of freedom and *MSE* is the mean squared error of the model. Statistical significance of single cell ωPEVs was validated using cluster based permutation testing (see: *Statistical testing* in Methods) with randomly shuffled stimulus conditions.

Cross temporal discriminability analysis

General decoding approach: Cross temporal discriminability of stimulus information on the level of the neural population was assessed using the analysis described in the studies by Stokes et al. (2013) and Spaak et al. (2017) as well as in **Chapter 2**. First, we randomly assigned each trial to one of two independent data splits. We then computed the mean firing rate over trials per neuron per independent split and per condition (locations * colors = 6 x 2 = 12) and took the pairwise differences between all conditions (66) for each neuron within each independent split. For each pairwise condition difference we then calculated the Pearson correlation for each time point across neurons between independent splits and averaged the resulting correlation coefficients using Fisher's z-transformation to obtain a single time resolved discriminability measure. It is straightforward to extend this basic decoder to cross temporal decoding (see Methods **Chapter 2**). Significance of discriminability was assessed by cluster-based permutation testing with randomly shuffled condition labels.

To decode location information only, we defined conditions with respect to the stimulus location, regardless of color (number of conditions = 6; pairwise differences = 15). To decode color information only, we focused on the difference between color conditions within each location (number of conditions = 12; within location differences = 6).

Target only decoding: To decode target only information i.e. location and color information about the stimulus that was going to change its identity after delay period, we performed the same discriminability analysis with one minor change. In brief, mean firing rates were computed over trials per neuron per independent split per condition, with a condition not referring to any stimulus in location x and color y (as before) but to the target only.

Correct versus error trial analysis: For the correct versus error trial analysis, we applied our target decoding approach as described above. Correct and erroneous trials refer to trials on which the monkey either succeeded or failed in identifying the correct location of the stimulus that had changed color (i.e. the target). First, we counterbalanced correct and error trials while controlling for the overall load (i.e. the number of total stimuli on display) as well as for ipsilateral load (i.e. the number of stimuli on the same visual hemifield as the target stimulus) since this was shown to most strongly affect the monkeys' behavior as well as the neuronal signals in the present data (Buschman et al., 2011). We then applied our target decoding approach to correct and error trials separately a 100 times (for 100 counterbalanced sub-selections of correct and error trials).

In general, due to the exclusion of trials when counterbalancing between correct and error trials as well as generally fewer trials per condition for the target stimulus only (we required >8 trials per condition to achieve reliable decoding), a large proportion of cells was excluded that didn't meet the trial number criterion in addition to any aforementioned criteria applied to our data. This left us with 60 cells in lPFC; 50 cells in FEF and 35 cells in LIP for the correct versus error trial analysis. Few cells, few trials and generally weaker target decoding resulted in statistically unreliable observations in LIP, which could not be interpreted using the presented approach.

Estimation of temporal coding dimensionality

Principal component analysis (PCA) was used to estimate the temporal coding dimensionality of single cells. Commonly, PCA is used to explain the variance of

population activity in neural state space in response to task features (Cunningham and Yu, 2014). Here, we sought to apply the same idea to single neurons over time, thereby quantifying the temporal variability of task dependent firing. First, we constructed a data matrix *Y*: For each trial, we counted raw spikes falling into independent 50 ms time windows (spanning the 2000 ms trial). We then averaged binned spike counts by task condition (locations * colors = 6 x 2 = 12). Columns in *Y* correspond to 10 adjacent, independent time windows (spanning 500ms), (i.e. dimensions), and rows to the 12 task condition averages (i.e. samples). PCA is then performed on *Y* to quantify and arrange the variance along orthogonal axes i.e. principal components in the 'space' spanned by the ten independent time bins. The eigenvalues associated with each principal component give us the means to quantify the effective dimensionality (N_{eff}) (Abbott et al., 2010) of our temporal state space.

$$N_{eff} = \frac{(\sum \lambda)^2}{\sum \lambda^2} \tag{3.3}$$

where $\boldsymbol{\lambda}$ represent the eigenvalues. The N_{eff} measure penalizes small eigenvalues which could arise due to noise. A high N_{eff} (the maximum equals ten in our example) suggests a high independence across time, i.e. a high temporal coding dimensionality. In other words, a cell with a high temporal dimensionality has either an unstable temporal condition selectivity (i.e., noise) or switches its condition selectivity over time.

Temporal discriminability analysis

To decode task information in single cells by leveraging variability over time, we applied a temporal discriminability analysis described in the following. For each neuron,

we randomly assigned each trial to one of two independent data splits. We then computed the mean firing rate over trials per independent split and per condition (locations * colors = 6 x 2 = 12) and took the pairwise differences between all conditions (66) within each independent split. For each pairwise difference, we computed the Pearson correlation between time-points from a 200ms window in split A of the data with the same time points in split B of the data. This was done for all pairwise condition differences and the resulting correlation coefficients were averaged using using Fisher's z-transformation. By sliding the 200ms window along the trials (1 ms increments) and repeating the analysis we obtained a temporal discriminability score, which captures the information present in the temporal structure of the signal. Temporal variability that is purely driven by noise will result in a low temporal variability score, as will condition specific but stable firing within the 200ms. Conversely, dynamic but specific firing will result in a comparably high temporal discriminability score.

Computing spike count correlations

Spike count correlations (rSC) (also known as noise correlations) (Cohen and Kohn, 2011) were calculated from the 500ms fixation period. We derived rSC values by comparing single cell responses recorded form the same electrode on the same session. This criterion limited our analysis to electrodes on which more than one cell was recorded. Furthermore, not all cells recorded by an electrode had the same number of trials, hence, we were limited to trials common to all cells recorded on one electrode. Here, we set the minimum inclusion criterion to 20 trials. Lastly, we only included cells that had an average firing rate greater than 2 sp/s during the fixation period. Together

with our sub-selection of cells that were assigned a τ value we were left with: 126 cells for lPFC; 74 cells for FEF and 55 cells for LIP.

For each trial, we counted all spikes falling into the 500ms fixation period window. We then normalized spike rates on each trial for both mean firing and slow drifts in neural excitation by computing a z-score for each neuron's firing rate on each trial using a sliding window of 10 trials before and after the current trial (Cohen and Newsome, 2008).

$$z_i(k) = \frac{r_i(k) - \mu_i}{\sigma_i} \tag{3.4}$$

where r is the firing rate at trial k and μ and σ are the mean and standard deviation of the i^{th} neuron's firing rate estimated from the 21 trials centered on k. The Pearson correlation of a neuron's z-scores with those of any other neuron give the rSC. Ultimately, our goal was to correlate rSC with τ. Therefore, in addition to its τ value, each cell was assigned the Fisher-transformed average rSC of all its pairings on the same contact. If only two cells were recorded on one electrode, we reassigned both cells the mean of their respective log transformed τ values. Furthermore, to evaluate whether a possible correlation between τ and rSC was confounded by firing rates (de la Rocha et al., 2007), we also took the mean of the log transformed baseline firing rates making up each averaged rSC value.

3.4 Results

Spikes were recorded from lateral prefrontal cortex (lPFC, n=583), frontal eye fields (FEF, n=323) and lateral intraparietal cortex (LIP, n=281) (Fig. 3.1a, brain schematic) while monkeys performed a delayed change detection WM task (Fig. 3.1a, left panel). To initiate a trial monkeys had to fixate a red circle in the middle of a black screen (fixation period: 500ms). Upon successful fixation monkeys were presented with a sample array for 800ms (sample period). A sample array could contain two to five colored squares (items) distributed over six locations (three in each visual hemifield). After a memory delay of 800-1000ms (delay period) a test array appeared on the screen. The test array was equal to the sample array except for one of the items (the target), changed color. Monkeys had to indicate the location of the target with a saccade. For details regarding the task structure refer to Methods and to Buschman et al. (2011).

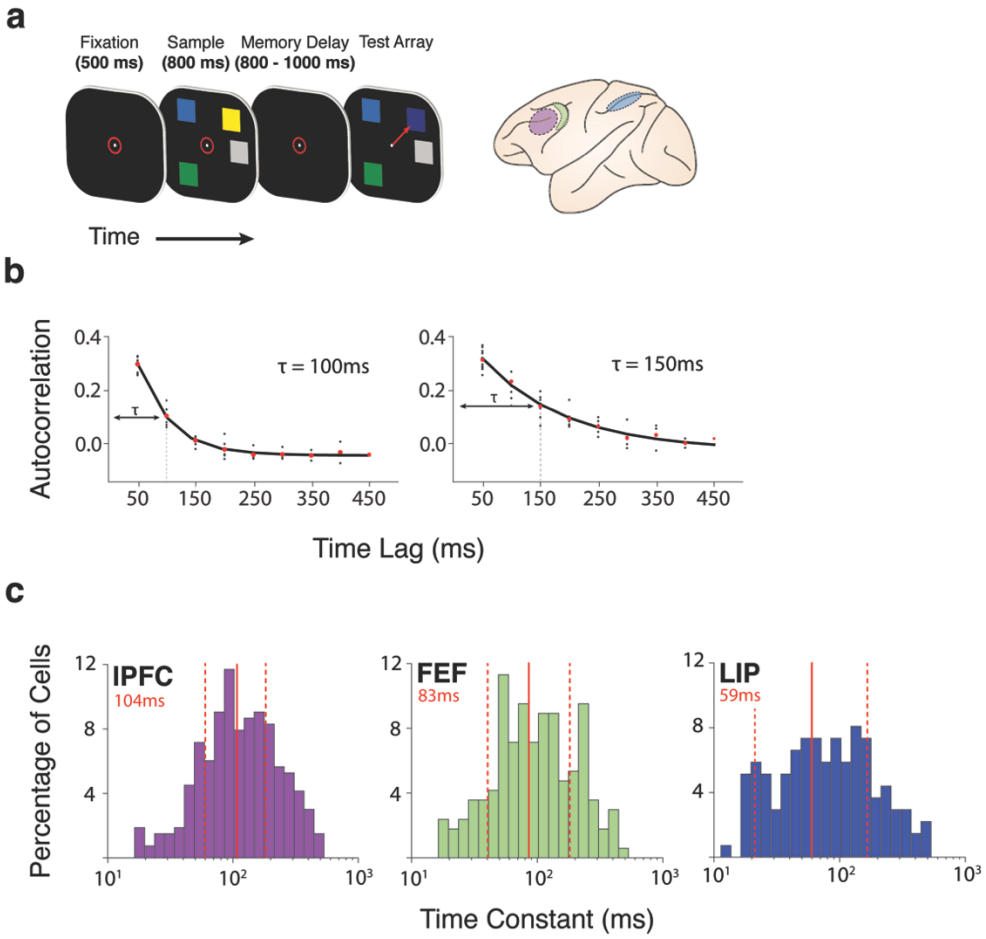


Figure 3.1: Experimental paradigm & estimation of intrinsic timescales: a) Delayed change detection task: Subjects fixated to initiate the start of a trial (red circle, 500ms). A sample array was presented for 800ms consisting of two to five items. After a memory delay (800-1000ms) a test array was displayed that was identical to the sample, except one item (the target) had changed color. Animals indicated target location by a saccade (see [51]). The right panel shows recording sites: lPFC (purple); FEF (green) and LIP (blue). **b)** Autocorrelation decay functions of two example cells estimated over the duration of the fixation period (left: short τ; right: long τ). The red points denote the mean autocorrelation for a specific time lag. Each black dot marks the autocorrelation of each time-bin x and any other time-bin x' at the lag specified on the x-axis. Solid black lines denote exponential fits to the red points. **c)** Histograms of the distribution of intrinsic timescales estimated separately for each cell and each brain region lPFC (n=265); FEF (n=168); LIP (n=136). Color code corresponds to the brain schematic. Red vertical lines denote mean and standard deviations, estimated from log transformed τ values (mean values are displayed in red).

Defining neural populations by their intrinsic timescale

Our main question was how the intrinsic temporal stability of a neuron (Murray et al., 2014) contributes to the dynamic processes underlying a WM task. To investigate this question, we computed the decay time constant of each neuron's autocorrelation

function during the fixation period (Murray et al., 2014; Chaudhuri et al., 2015; Cavanagh et al., 2016). The resulting quantity is referred to as the intrinsic timescale (tau, τ). A relatively long τ indicates stable firing during fixation, as opposed to a short τ indicating more dynamic baseline firing patterns. Figure 3.1b shows two example cells from lPFC with different τ's. We plotted each cell's τ as a function of brain region (Fig. 3.1c). Average τ values differed across brain regions ($p<0.0001$, *Kruskal-Wallis*). Intrinsic timescales in lPFC were larger than in LIP ($p < 0.001$, *Kruskal-Wallis followed by Dunn's test*), but not FEF ($p<0.08$). Average τ values in LIP and FEF were not significantly different from each other ($p = 0.14$). Importantly, we also found that our observed τ values did not depend on the average fixation firing rate in any of the brain regions (Spearman rank correlation of τ with firing rate; lPFC: $r=0.06$, $p=0.3$; FEF: $r=0.12$, $p=0.1$; LIP: $r=0.08$, $p=0.3$), nor was there any significant relationship between tau and firing rate within each task epoch (Supp. Fig. 3.1).

Cells with larger intrinsic timescales carry more WM-related information

To characterize the influence of individual intrinsic firing stability on task involvement, we first quantified the amount of information each cell carried about the task relevant features (i.e., memory item). Here, we used the percentage of variance explained (ωPEV) statistic to measure the extent to which the variability in neural firing rate was determined by color and location (Buschman et al., 2011).

To determine the relationship between intrinsic timescale and single cell item selectivity, we sorted cells by τ and plotted their respective ωPEV over time (Fig. 3.2a). Visual inspection of the sorted neural populations reveals a clear pattern towards stronger item

encoding as a function of increasing τ in all three brain regions. Splitting the neural population according to their respective median τ value and averaging the ωPEV for each median split draws out this conclusion (Fig. 3.2b). In lPFC, long τ cells carried more information, especially during the delay period (p=0.003, *cluster-based permutation test*), while in FEF long τ cells were more informative throughout both the sample and delay periods (p=0.008, p=0.02). Interestingly, there was little difference between cell types during the initial transient responses associated with memory encoding, or the response probe at the end of the trial: the functional benefit for long τ cells only emerges during the more stable task epoch. LIP cells showed a similar trend during the sample period, but not during the delay period, where mnemonic coding was poor.

To further investigate the observed relationship, we computed a rank correlation between each cell's τ and ωPEV value for each time point in the trial (Fig. 3.2c). The positive correlation time courses confirm our results. In lPFC, the correlation between τ and ωPEV shows an initial, transient bump ($p<0.01$, *cluster based permutation test*), as also seen in the median split plots (Fig. 3.2b). This is followed by a progressive increase during the delay period (p=0.0001). In FEF we observed fast rising correlation values during the sample period, which stayed high throughout the task (p=0.002). In LIP the correlation between ωPEV and τ is less straightforward, showing a trend for larger correlation values during the sample period of the task (no cluster survived thresholding), followed by a drop-off during the delay period. Although not directly correlated with τ, the average firing rate during fixation might still influence the relationship between τ and the ωPEV. Therefore, we regressed the average ωPEV estimated from the periods of significant Pearson correlation (Fig. 3.2c) against τ

factoring out average fixation firing rate (lPFC: β=0.19, p=0.004; FEF: β=0.25, p=0.008; LIP: β=0.2, p=0.017).

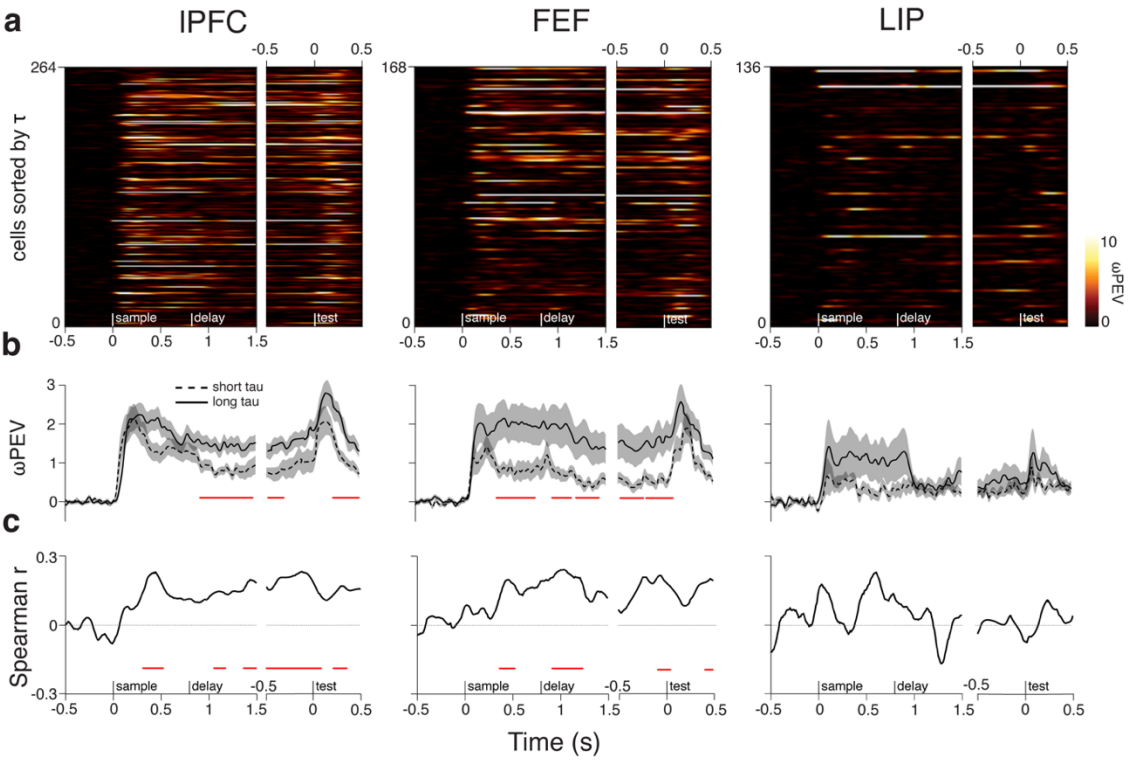


Figure 3.2: Item information and intrinsic timescales. a, b & c: a) Cells sorted by their intrinsic timescale. The color scale indicates the amount of item information i.e. ωPEV present at each time point in a given cell. Traces are smoothed with a Gaussian kernel (s.d. = 50 ms) along the time axis and nearest neighbor interpolation along the cell axis. Initially, trial time is shown with respect to sample onset. However, the duration of the memory delay is variable (800-1000), therefore, we show the last part of the trial (late memory delay until the execution of behavioral response) separately, with trial time shown in reference to presentation of the test screen. The same convention is used for all brain areas (lPFC: n=265; FEF: n=168; LIP: n=136). **b)** Median splits of cells according to their intrinsic timescales. The black dashed line denotes the time resolved average over all cells belonging to the first half of the split i.e. short τ. The black solid line denotes the time resolved average ωPEV over all cells belonging to the second half of the split i.e. long τ. The shaded area represents the standard error of the mean (s.e.m.). The red solid line marks a significant difference between the two splits. **c)** Spearman rank correlation of the intrinsic timescales and the time-resolved ωPEV. The red solid line marks a significant correlation coefficient. Correlation time-courses were smoothed with a Gaussian kernel (s.d. = 50ms); statistics were computed on unsmoothed traces.

Furthermore, we assessed the proportion of cells with significant item information within our tau splits. Within each region, a larger proportion of long tau cells showed significant ωPEV at any time point in the trial: lPFC (long tau: 113/132; short tau: 90/132), FEF (long tau: 63/84; short tau: 58/84), and LIP (long tau: 29/67; short tau:

24/67). When focusing on the significant cells only, while performing a new median split by tau within those, average ωPEV traces resembled the ones computed from the original cell population (Supp. Fig. 3.2).

Intrinsic timescale predicts encoding onset during the sample period in lPFC

Cells associated with longer τ values carried more information, which was especially prominent during the delay period in lPFC. In the visual system, smaller receptive fields are useful for detection of rapid fluctuations in stimuli (Buračas et al., 1998). Analogous, we considered whether short τ cells might therefore be faster at encoding at the time around stimulus onset, since such a division would point towards a differential weighting of cells according to τ for either fast perceptual encoding and longer-term storage of the stimulus.

To investigate this hypothesis, we limited our analysis to cells that showed significant item encoding (ωPEV) at any point during the sample period, since this allowed us to quantify significant encoding onsets. Visual inspection of the average ωPEV in short and long τ subpopulations suggests that short τ cells might encode more information earlier during the sample period in lPFC (Fig. 3.3a; no cluster survived thresholding). To examine this in more detail, we computed the fraction of currently significant cells for each time point while averaging over τ values for this fraction (Fig. 3.3b). We found that, especially during the early period of the sample epoch, as more cells became task-selective, the average τ value for those task-engaged cells increased as well. This observation was complemented by a significant positive correlation between τ and encoding onset time (time of first significant ωPEV) during the time period in which the

fraction of encoding cells reached 80% percent of its maximum (first 200ms of the sample period) (r=0.36; p<0.001, *Spearman rank correlation*). This correlation remained significant even after taking the whole sample period into account (r=0.16; p=0.03). This result clearly shows that cells with a shorter τ actively encode task information earlier during stimulus presentation in lPFC. In FEF and LIP we found no such relationship (Supp. Fig 3.3).

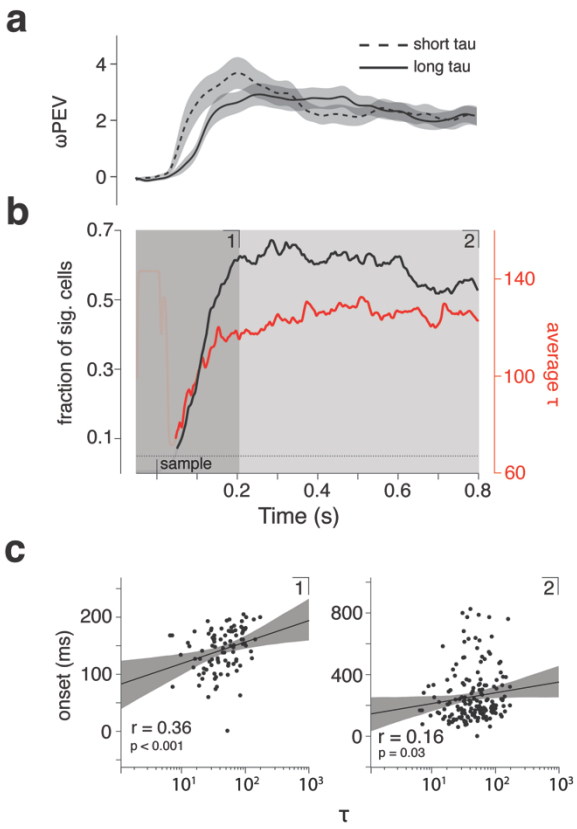


Figure 3.3: **Evaluation of sample period onset times in relation to intrinsic timescales for lPFC. a)** Average ωPEV per median split (dashed: short τ; solid: long τ) for sample period only; Shaded region indicates s.e.m.. **b)** Black line indicates the fraction of significant cells (ωPEV; p<0.05) at each time point from all cells showing a significant ωPEV at any point in during the sample period. The red line shows the average τ value of cells having a significant ωPEV at that time point. The shaded region marks the time period at which the fraction of significant cells first reached 80% of its maximum value. Occluded parts mark the first time period with fewer than ten significant cells. Numbers in the upper part of the panels refer to **c**. **c)** Left panel: Scatter plots shows onset times (ms) of significant ωPEV estimated during time period **1** i.e. until the fraction of significant cells reached 80% of its maximum value, versus log transformed τ values (r=0.36; p<0.001; n=96, *Spearman rank correlation*). Each dot marks a cell. Right panel: The same scatter plot but with onset times estimated from the entire sample period **2** (r=0.16; p=0.03; n=176, *Spearman rank correlation*). Black lines indicate linear fits to the data with the shaded area depicting the 95% confidence interval of the fit.

Population analysis reveals a more stable code for cells with longer intrinsic timescales

Next, we extended our analysis to investigate how τ influences coding dynamics on the level of the neuronal population. Recent studies successfully used cross temporal pattern analysis to characterize dynamics during a range of WM tasks (Meyers et al., 2008; Stokes et al., 2013; Spaak et al., 2017; Murray et al., 2017). Here, we applied the same method to investigate potential differences in our two subpopulations of neurons (short vs long τ). In brief, we split the observed trials into two independent halves and computed the averaged firing rate, per condition, per neuron, for each half. We then computed all pairwise differences within each split, and correlate the pattern of pairwise condition differences of the whole population at each time point **t** in split one with every time point **t'** in split two, yielding a two-dimensional matrix representing a discriminability score (see Methods; also **Chapter 2**). If the discriminating pattern is stationary over time, the within-time correlation values should resemble the between time-point ones; i.e. the pattern should cross-generalize over time. Conversely, a dynamic pattern will result in correlation values lower than the corresponding time-points along the diagonal of the matrix.

To probe whether our previous univariate observations generalize to the population, we first applied the above described decoding approach within-time i.e. for pairs of same time-points between splits (Fig. 3.4a). Across regions, within-time discriminability time-courses strongly resembled those from the ωPEV analysis (Fig. 3.4a). In lPFC, short τ cells initially showed a higher discriminability score ($p<0.01$, *cluster based permutation test*) which decayed during the delay period, whereas mnemonic information was more strongly represented by long τ cells ($p<0.01$). In FEF, long τ cells

showed higher discriminability throughout the task ($p<0.01$) while in LIP, item information was only discriminable during the sample period. Here, again, long τ cells showed higher item discriminability than short τ cells ($p<0.01$).

Figure 3.4b shows the across-time extension of our decoding approach. Note, the diagonals of the respective matrix plots correspond to time-courses in Fig. 3.4a. In general, cross-temporal population coding was more prominent and temporally stable for the long τ subpopulation. In lPFC, white contours indicating a significant discriminability score ($p<0.001$, *cluster based permutation test*) extend over the entire task duration for the long τ subpopulation (Fig. 3.4b). In contrast, in the short τ subpopulation off-diagonal discriminability drops to chance level at the transition between sample and delay periods and also within the delay period. Subtracting the discriminability matrices computed for the two subpopulations (long τ - short τ) reveals significantly stronger and more stable decoding during the delay period within long τ cells ($p<0.01$, *cluster based permutation test*). In FEF, the discriminative pattern of the long τ cells strongly cross-generalizes both within task epoch as well as across task epochs (white contours, $p<0.001$). This strong cross-generalization is absent in the short τ subpopulation which shows significantly less cross-temporal decoding ($p<0.01$). In LIP differences between τ populations are confined to the sample period where long τ cells show a stronger discrimination score ($p<0.01$), which is remarkably stable ($p<0.001$).

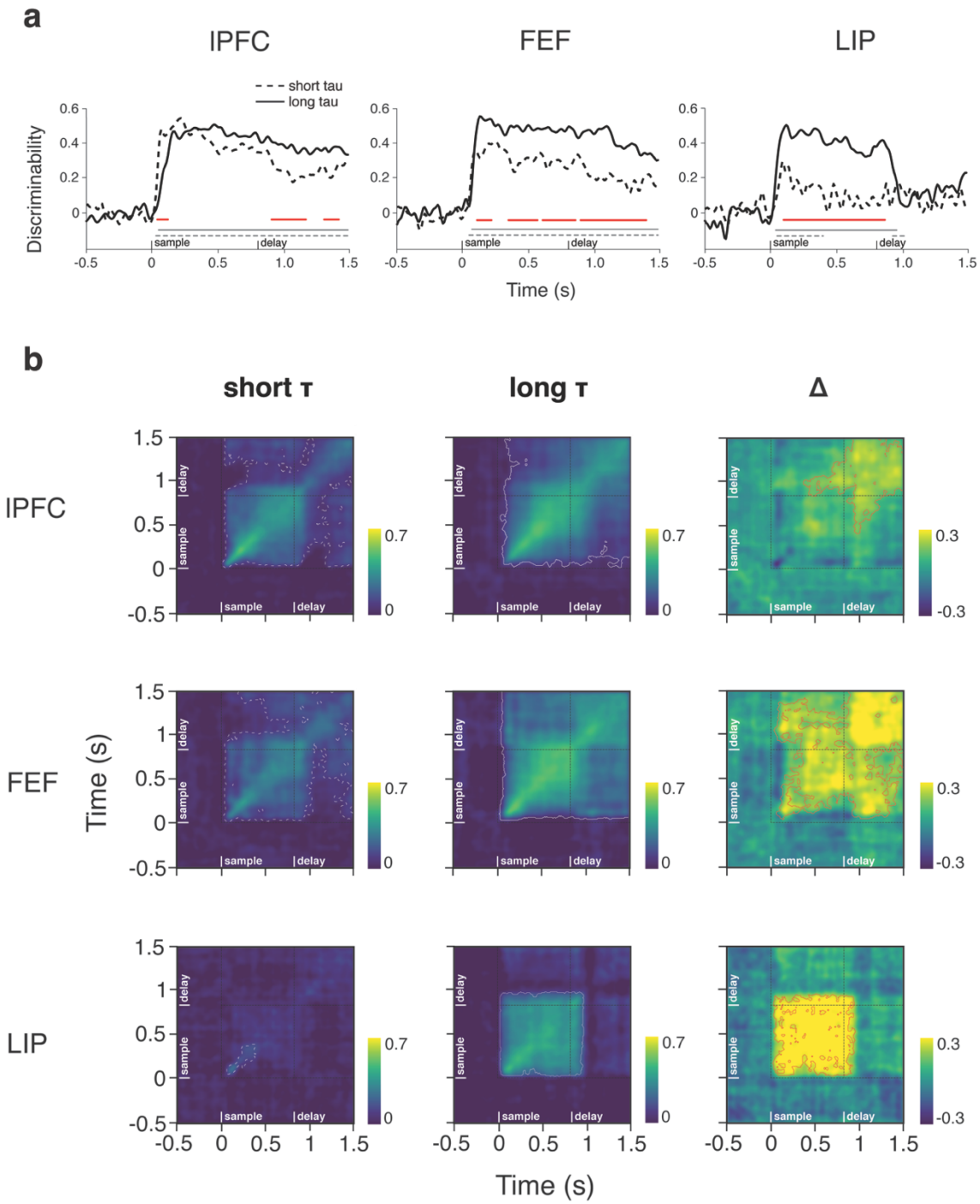


Figure 3.4: Within-time & cross-temporal decoding and intrinsic timescales. a) Within-time decoding. Solid black lines indicate decoding discriminability for long τ cells. Dashed black lines indicate decoding discriminability for short τ cells. Grey lines indicate significant discriminability ($p<0.001$, *cluster based permutation test*). Red lines indicate significant differences between the short and long τ cells ($p<0.01$, *cluster based permutation te*st). **b**) The cross temporal discriminability is color coded. Yellow colors indicate strong decoding while blue colors indicate weak or no decoding. White dashed and white solid contours indicate significant cross-temporal discriminability within respective subpopulation (cluster based permutation test; $p<0.001$). Black dashed lines distinguish task periods. First column: short τ subpopulations (dashed). Middle column: long τ (solid). The right column shows the difference between the two first columns i.e. long τ - short τ. Red contours mark a significant difference ($p<0.01$, *cluster based permutation test*) PFC (n=265); FEF (n=168); LIP (n=136).

To illustrate population decoding for both item features separately, we conducted within and across-time temporal discriminability analysis for location and color only (see Methods) (Supp. Fig. 3.4 & 3.5; for statistical test values refer to figure legends). For locations only, differences between decoding in τ subpopulations in lPFC and FEF were similar to those observed for combined item decoding (i.e. color & location; Fig. 3.4) (Supp. Fig. 3.4). In LIP, location information was weak but present during the delay period ($p<0.05$), as evident from the cross-temporal plots. Interestingly, long τ cells showed a similar re-emergent pattern during sample and late delay periods. Apart from a very weak effect in FEF, color information was only present in lPFC (Supp. Fig. 3.5). Here, the long τ subpopulation showed stronger within and across time discriminability scores, especially during the late sample and early delay periods. Color encoding was only weak in FEF and absent in LIP, preventing an informative comparison between τ subpopulations.

Lastly, we wondered whether in principle our method would be able to uncover a relationship between neuronal signals and behaviour, under the assumption that the decoded information is a behaviourally relevant read-out. To this end, we implemented a correct versus error trial analysis. More specifically, we wanted to know whether there was any difference between the decodability of the item that was going to change its color i.e. the target, on trials were the monkey indicated the correct location of the color change (correct trial) and when it indicated the wrong location (error trial). Our correct versus error trial analysis revealed a significant decrease in target decoding during the delay period in lPFC on error trials (Supp. Fig. 3.6). This effect was also clearly visible on the off-diagonals of the cross-temporal plots, were target information of the sample period was similar to that in the delay period only within correct trials (Supp. Fig. 3.6).

There was a similar trend in FEF (Supp. Fig. 3.6) but our analysis was too noisy to make any interpretations for LIP (Supp. Fig. 3.6). However, this observation confirms the behavioural relevance of neural activity in lPFC captured by our decoding approach.

Intrinsic timescales predict single cell temporal coding dimensionality

So far, our results show the existence of a link between an individual cell's intrinsic firing stability and the magnitude of its item encoding. However, the ωPEV does not make any assumption about the consistency of neural responses over time nor does it take into account a neuron changing its selectivity (Enel et al., 2016; Spaak et al., 2017; Sigala et al., 2008). While we show that on the population level, intrinsic timescale affects the stability of information across time, we have no single cell measure quantifying temporal coding stability for an individual cell. Hence, we chose to evaluate the effective temporal coding dimensionality (N_{eff}; where N stands for the number of principal components) of each cell via principal component analysis (PCA; see Methods). In neural population analyses, PCA is classically used for dimensionality reduction; i.e. each neuron counts as a dimension and each trial or time point counts as a point in neuronal state space. Here, we apply the same method to a single neuron with adjacent time bins counting as dimensions and mean condition firing rates counting as points in 'time state space'. Specifically, a high N_{eff} indicates a high temporal dimensionality which can be caused by either non-specific activity, unstable firing or switching selectivity.

We estimated the N_{eff} in a stepping 500ms time window composed of 10 independent 50ms bins (step size of 50ms). This particular relation between time window and bin

size was chosen since it allowed us to evaluate single epochs of the task while capturing most of the temporal coding dimensionality within a given task epoch, and across brain regions (Supp. Fig. 3.7). As expected, N_{eff} was highest during the fixation period, when condition specific firing was absent (Fig. 3.5a). Eliminating condition specific firing by shuffling condition labels yielded N_{eff} similar to those of the fixation period (data not shown). In lPFC and FEF, cells with longer τ values displayed a lower temporal dimensionality during sample and delay periods ($p<0.001$, *cluster based permutation test*) (Fig. 3.5a). In LIP, the difference between N_{eff} of the two split halves was significant only around the sample period. The correlation time-courses (Spearman r) in Fig. 3.5b confirm this observation. In FEF, the correlation between N_{eff} and τ steadily increases over time, peaking during the delay period ($p<0.01$, *cluster based permutation test*). In lPFC a strong correlation is evident already early during the sample period ($p<0.01$), which dips between the two task periods, suggesting a τ independent state transition, in order to re-emerge during the delay period ($p<0.001$). In LIP a dependence of N_{eff} on τ is clearly only present during the sample period ($p<0.01$). Scatter plots of all neurons/area in Fig. 3.5c show the same correlation but for the average N_{eff} estimated from independent time periods within task epochs. Since N_{eff} gives an estimate of the temporal variability of neuronal firing, we should expect to observe a higher N_{eff} for shorter τ values during the fixation period. In fact, we found such a relationship in lPFC (Fig. 3.5c; r=-0.15, p=0.013, *Spearman rank correlation*) but failed to observe a significant effect in FEF (r=-0.13, p=0.1) and LIP (r=-0.1, p=0.23).

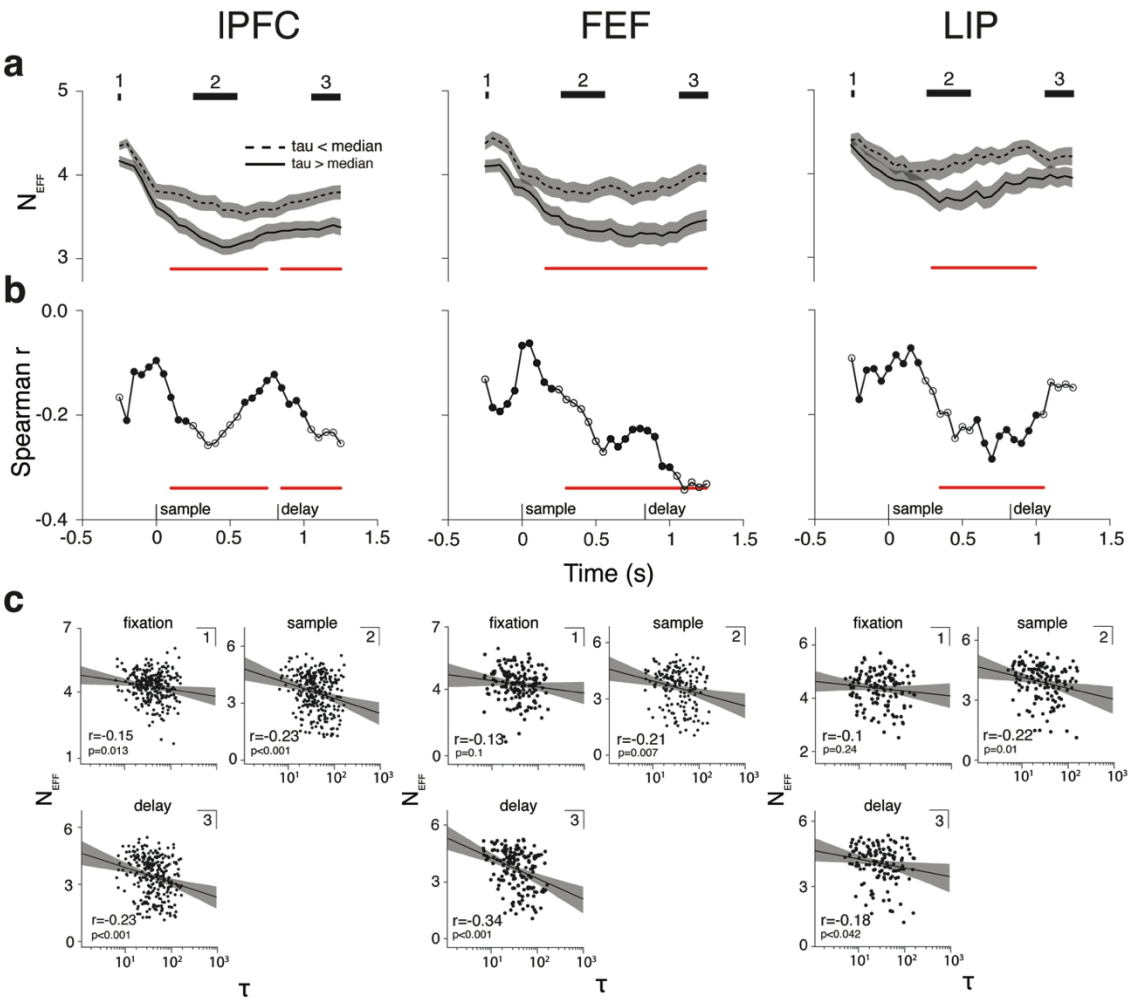


Figure 3.5: Effective temporal dimensionality and intrinsic timescales. a) Effective temporal dimensionality (N_{eff}) estimated at a single cell level (N_{eff} was estimated from a sliding 50ms*10 (500ms) time window shifted by 50ms on each step. Dashed black line: Average N_{eff} for short cells τ cells. Solid black line: Average N_{eff} for long τ cells. Red horizontal line: significant difference between the two splits halves (p<0.01, *cluster based permutation test*). Shaded regions mark s.e.m. Black bars mark N_{eff} estimated within independent trial epochs i.e. were the sliding window did not overlap with distinct epochs (**1**: fixation; **2:** sample; **3**: delay). **b**) Spearman correlation between N_{eff} and log transformed τ values. Each dot marks the correlation coefficient as a function of time i.e. the center of the time window from which N_{eff} was estimated. Red line: significant Spearman correlation coefficient (p<0.01, *cluster based permutation test*). Hollow dots mark independent trial epochs (**1**: fixation; **2:** sample; **3**: delay). **C**: Scatter plots for N_{eff} estimated from independent trial epochs (number in left corner refers to number on black bar) and τ values. Each dot represents a neuron: lPFC (n=264); FEF (n=167); LIP (n=134).

Robust temporal dynamics serve as a feature for discrimination

Next, we hypothesized that if a short τ cell genuinely encodes item information, albeit in a temporally dynamic way, the temporal activity pattern should carry information not

evident in a purely stable cell. In brief, per neuron, we split trials into two independent halves and quantified the correlation between the temporal pattern of item specific activity estimated over 200ms between the two split halves (see Methods). If the temporal activity trace is corrupted by noise, correlation between the split halves (i.e. the discriminability score) should be small. Similarly, a condition specific but temporally stable trace would also translate into a comparatively small discrimination score. However, a dynamic temporal activity pattern should result in a relatively high discrimination score.

Indeed, item information was discriminable in the temporal coding dynamics inherent in both short and long τ cells ($p<0.001$, *cluster based permutation tes*t) (Fig. 3.6, for lPFC & Supp. Fig. 3.8, for FEF & LIP). In lPFC, temporal discriminability time courses were strikingly similar during both the sample and test periods. During the delay period, temporal discriminability initially rose but then stayed significant only for short τ cells ($p<0.001$), which showed significantly higher decoding during part of the delay ($p=0.017$). In FEF and LIP we did not observe any significant difference between the two subpopulations separated by τ (Supp. Fig. 3.8). Interestingly, in all three brain regions, our method seems to capture transitions between epochs as decoding falls back to baseline as task epochs unfold.

Correlating τ and temporal discriminability over time accentuates the observation made from the median splits. In lPFC the correlation time-course shows a significant dip during the delay period ($p<0.001$) emphasizing stronger temporal discriminability for short τ cells (Fig. 3.6). Interestingly, a similar dip is evident in the FEF correlation time

course at a later time point in the delay ($p<0.01$) which was only hinted at in the median split plots (Supp. Fig. 3.8a).

Note, it is improbable for any cell to encode information in a perfectly stable way and strong item encoding is likely to drive much of the observed correlation. However, keeping in mind the weaker item encoding for short τ cells as observed by the ωPEV, an equal or even stronger temporal discriminability score for the same cells strongly speaks for a robust but dynamic coding within this subpopulation.

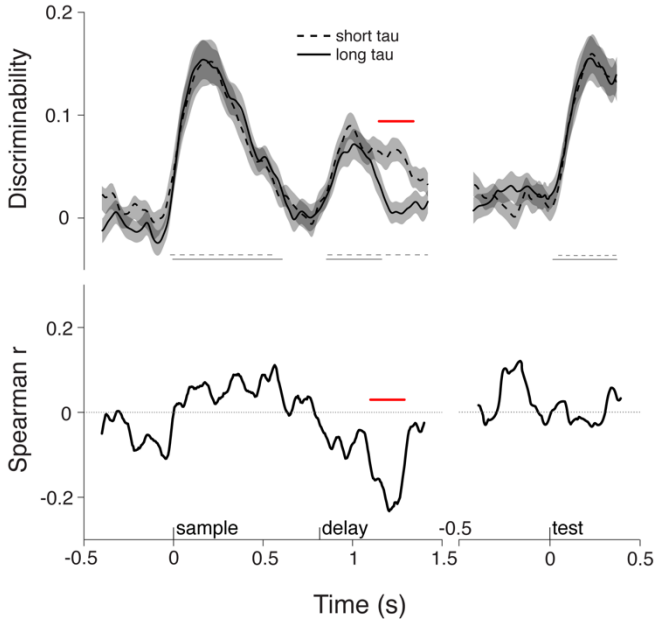


Figure 3.6: Temporal discriminability in lPFC. The upper panel shows the temporal discriminability score (see methods) averaged over neurons. Black solid line indicate the average over long τ cells and black dashed lines indicate the average over short τ cells. Shaded area indicate s.e.m. Thin grey lines indicate significant discriminability ($p<0.001$, *cluster based permutation test*), for respective τ splits. The red solid line depicts a significant difference between the two splits of the data ($p<0.05$, *cluster based permutation test*). The lower panel shows the Spearman rank correlation between τ and discriminability score for each time point. The Solid red line denotes a significant correlation ($p<0.05$, *cluster based permutation test*)

Functional connectivity at rest correlates with intrinsic timescales

Next, we tested possible sources leading to our observed τ values. A recent modelling study proposed that intrinsic timescales could, in part, arise through local connection densities (Chaudhuri et al., 2015). This observation raises the possibility that neurons associated with a long τ value might be part of a local network hub with a high concentration of incoming and outgoing connections (Chaisangmongkon, 2017; Runyan et al., 2017). To test this hypothesis, we probed functional connectivity using spike count correlations (rSC), also referred to as noise correlations (Cohen and Kohn, 2011) (see Methods). Spike count correlations were estimated from the fixation period since we wanted to determine whether τ depended on functional connectivity at rest i.e. before a stimulus synchronized network activity. We found a significant correlation between τ values and rSC when pooling over all brain regions (r=0.18, p=0.004, *Spearman rank correlation*) (Fig. 3.7 upper left panel). Spike count correlation can depend on basic firing rate (de la Rocha et al., (2007)). However, we found no evidence for a significant correlation between rSC or τ and fixation activity, respectively (Supp. Fig. 3.9; r=-0.05, p=0.37; r=-0.05 p=0.38). Furthermore, the relationship between rSC and τ was also present when controlling for baseline firing using a multiple linear regression model (β=0.03; p=0.007). Together, this implies that differences in fixation activity were not responsible for the positive relationship between rSC and τ. Separating out individual brain regions revealed a more diverse picture: In lPFC, rSC and τ were not correlated (r=0.04, p=0.63, Fig. 3.7 lower left panel), while both FEF and LIP showed a positive correlation between rSC and τ (FEF: r=0.28, p=0.015; LIP: r=0.33, p=0.013), which held true when controlling for baseline firing using the multiple regression model (FEF: β=0.05, p=0.02; LIP: β=0.04, p=0.02).

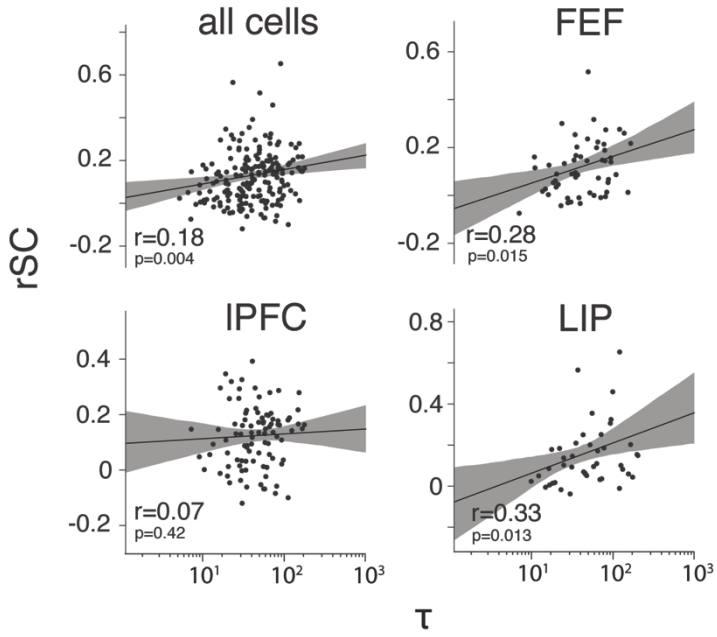


Figure 3.7: Spike count correlations and intrinsic timescales. Scatter plots for rSC on the y-axis versus the τ values on the x-axis. rSC is estimated during the fixation period and between cells on the same contact. Each dot depicts one cell. The r value within panels depicts the Spearman correlation coefficient and associated p-value, for the two variables. All cells: n=255; lPFC: n=126; FEF: n=74; LIP: n=55. Black solid lines show linear regression lines. Shaded region depicts 95% confidence interval.

3.5 Discussion

Heterogeneity in temporal tuning observed during WM is related to baseline firing dynamics of individual neurons. Specifically, we found that cells with a longer intrinsic timescale carried more information throughout the task epochs, and particularly during the delay period in prefrontal areas. Important to WM function, mnemonic coding of long timescale cells was more stable across time in lPFC and FEF. In contrast, short timescale cells responded more rapidly during encoding in lPFC, and carried rich temporal information. Overall, we demonstrate an intrinsic distinction between cells of a functionally heterogeneous pool, which perform complementary computations: rapid encoding or longer-term storage of information.

The core finding of this study is that the baseline temporal stability of individual neurons in lPFC predicts how robustly they code WM-related information, especially during the mnemonic delay period. Early studies of WM focused on persistent activity in lPFC representing specific items in memory (Miller et al., 1996), however more recent studies also reveal more complex dynamics during encoding and maintenance (Lundqvist et al., 2016; Meyers et al., 2008; Stokes et al., 2013; Spaak et al., 2017). In particular, it has been shown that single cells become active at different time points during the trial, encode information only transiently, possibly multiple times and/or switch their selectivity over time (Shafi et al., 2007; Spaak et al., 2017; Cromer et al., 2010). Importantly, we now find that the metric of intrinsic timescales as estimated at rest (i.e., pre-trial fixation period) predicts the coding dynamics during WM.

Recent evidence shows that despite transient dynamics during sample and early delay periods, information in the later delay period was maintained in a relatively stable population code (Spaak et al., 2017). Additionally, a study by Murray et al. (2017) showed that during WM, stable and dynamic population codes coexist. Conceptually, in those studies, dynamic and stable mnemonic representations refer to orthogonal subspaces in high dimensional neural state space (Murray et al., 2017). Here, we extend those observations by proposing that they are linked to intrinsic properties of neuronal subpopulations: long timescale cells more strongly contribute to the stable subspace whereas short timescale cells contribute to the dynamic subspace. Our data also shows that this is mostly the case during the delay period in prefrontal areas (Fig. 3.2 & Fig. 3.4). This contrasts with LIP, where encoding did not significantly differentiate between timescale subpopulations during the delay, where mnemonic coding was absent. Rather, larger timescale cells in LIP stably encoded more information during the sample period

(Fig. 3.4). A previous study recording from monkey LIP during a simple delayed response task, showed that baseline firing stability modulated stimulus specific activity during the delay period (Nishida et al., 2014). In that study baseline firing stability was estimated by a relatively coarse metric: The raw correlation value between spike counts within time-bins separated by a distinct lag. The tau metric proposed by Murray et al. (2014) more elegantly captures a neuron's intrinsic dynamics. First, the metric is not dependent on specific temporal lag used for estimating the temporal autocorrelation, but rather captures its temporal profile over a number of lags. Second, as a second-order property, the tau metric is inherently more robust to differences in absolute neural activity. Tau, therefore, more accurately captures a cell's intrinsic time window of integration than raw autocorrelation at any specific temporal lag. Furthermore, the complexity inherent in our task (color, locations & load effects) as well as relatively poor encoding within the LIP population might have diluted an otherwise clearer relationship. Using location information as discriminative feature revealed some cross-temporal generalization extending to the delay period in LIP (Supp. Fig. 3.4). In FEF, the difference between short versus long timescale cells was striking in terms of magnitude of information encoding during sample and delay period (Fig. 3.2 & Fig. 3.4). Different from lPFC, the discriminative pattern within the long timescale subpopulation cross-generalized more strongly from the sample, all the way through the delay period. Note, in our task, strong, temporally stable signals as observed in FEF are unlikely to reflect oculomotor preparation since monkeys can not prepare a response in advance. Importantly, by analysing local field potentials (LFPs) on a single trial basis, Lundqvist et al. (2016) showed that WM was manifested by the rate of discrete bursts in gamma activity. They argue that observed persistent activity and stable population coding arises through trial averaging of discrete idiosyncratic activity bursts. Although this result

might seem contradictory to the temporal stability observed in the long tau cells, it is important to point out that the two results focus on very different neural signatures (spiking and LFP), at very different temporal scales (100 s vs 1000 s of ms). It would be interesting to see future work that explores the relationship between intrinsic neuronal timescales and LFP oscillations, and their relation to WM. For now, dynamic bursts in both the LFP gamma signal and stable spiking are probable mechanisms underlying WM.

Results from an information theoretic approach suggest WM information can be stored more faithfully if the brain first encodes information appropriately before passing it to persistent activity networks (Koyluoglu et al., 2017). Within the same study, the authors showed that human performance is better fit by such a two-step model than by a model mimicking direct storage in persistent channels (Koyluoglu et al., 2017). Here we show that the measure of intrinsic timescale can help explain differences in temporal dynamics and therefore distinct computational roles within a complex WM task. In lPFC we saw that stable coding within the long timescale subpopulation generalizes only weakly across the different epochs of the task (i.e., sample to delay period) (Fig. 3.4), suggesting that in lPFC, the neural code transitions between different epochs, ultimately to be stored in a more stable format during the delay epoch through cells with longer intrinsic timescales. This seems to be in accordance with previous observations in PFC (Stokes et al., 2013; Spaak et al., 2017) and with the two step model described by Koyluoglu et al. (2017), while emphasizing distinct networks for perception and storage which are potentially both present during the sample period (Sigala et al., 2008; Markowitz et al., 2015; Pinto and Dan, 2015; Mendoza-Halliday and Martinez-Trujillo, 2017). A similar conclusion had been reached by Murray et al., (2017), who suggested orthogonal

mnemonic and perceptual representations during the sample period with the latter decaying during the delay period. Contributing to the evidence of distinct roles for cells with different timescales, we found that more dynamic cells, as defined by intrinsic timescale, coded for item information relatively early during the sample period. This onset cascade early in the sample period implies a more specific role for short timescale cells since they more rapidly detected the presented stimuli. This result might further play into a two-step model (i.e. rapid perceptual encoding and transfer to a robust stable code) of WM encoding leveraging the presence of a heterogeneous pool of available intrinsic timescales. It is important to note that while discussing different neurons being recruited for certain tasks, those neurons i.e. their respective timescales lie in a continuum rather than belonging to categorically distinct subpopulations. It therefore seems likely that contributions of individual neurons are weighted according to evolving task demands (Raposo et al., 2014). Along those lines, functional subpopulations are best thought of as distinct dimensions of the neuronal population rather than specific classes of cell types with distinct roles. In our data this notion is specifically apparent during the gradual evolvement in item encoding in lPFC (Fig. 3.2), and specifically during the onset cascade where cells became gradually engaged in the task as a function of tau (Fig. 3.3).

High temporal autocorrelation implies temporal stability, which seems well suited for WM. On the other hand, temporal autocorrelation also translates to a lower temporal dimensionality, which necessarily limits the informational capacity of coding over time. A recent, influential study found that high neural dimensionality is crucial for complex behaviour (Rigotti et al., 2013). Specifically, the dimensionality of the neural representation is expanded through neurons exhibiting mixed selectivity in response to

task factors, maximizing the possibility space for linear classification i.e. readout of the population activity (Rigotti et al., 2013). This view can be extended to the temporal domain, where dynamic changes in neural activity over time i.e. switching selectivity are reflected in a high temporal dimensionality. Indeed, the information potential of a highly dynamic network presentation is directly proportional to the statistical independence between time points (Stokes, 2011). Here, we could show that a faster decay of temporal autocorrelation at rest directly translates to a higher coding dimensionality during the task (Fig. 3.5). While our measure of temporal coding dimensionality (N_{eff}) captures selectivity to task parameters, it also reflects the temporal stability or instability of the tuning profile, extending the ωPEV measure substantially. This is evident in the clear epoch transitions in lPFC: long timescales cells maintain information in a lower dimensional temporal state only within each task epoch, not in transitions between epoch. Conversely, short timescale cells exhibit a higher temporal coding dimensionality even before the start of the delay period. In FEF the relationship between intrinsic timescale and temporal coding dimensionality steadily increases throughout the task, peaking during the delay period, an observation which again adds on the observed ωPEV time courses. Complementing those observations, we found that robust but temporally dynamic coding can discriminate information which would otherwise be lost in a purely stable code (Fig. 3.6). Additional information inherent in temporal dynamics was particularly evident in short timescale cells in lPFC during the delay. Here, short timescale cells showed weaker item encoding as quantified by the ωPEV. Nevertheless, their inherently high dimensional coding properties could be used to decode memory-related information.

Temporal dynamics are especially important within the context of cognitive flexibility. According to the principle of adaptive coding, PFC represents information in a dynamic and therefore flexible manner, where neural resources can be recruited 'on the fly', according to behavioural demands (Duncan and Miller, 2002). A recent study investigating neural correlates of cognitive flexibility in monkeys found that neurons effectively switching their selectivity over time were most active when cognitive control was demanded (Blackman et al., 2016). Crucially, most of the variance at the population level could be attributed to time or the interaction between time and trial type, showing that different cue-probe combinations are encoded differently in the distinct trial phases (Blackman et al., 2016). Furthermore, a recent study in monkeys suggested that the interplay between transient and stable coding guides the evolution of category based decisions during WM (Chaisangmongkon, et al., 2017). The advantage of a temporally stable representation during WM seems straightforward, also because such a low dimensional subspace could be read out over time by a set of fixed weights (Murray et al., 2017). On the other hand, such a low dimensional representation would limit the information capacity for coding transient events. An intrinsic reservoir of continuous neuronal timescales therefore seems particularly well suited for parallelizing either computation, as implied within the present study. In fact, depending on the task at hand different regions might adjust weightings of cells with different tau values. As an example: This suggests that intrinsic timescale might be necessary but not sufficient for involvement in WM. It is possible that LIP might be more involved in other types of maintenance (e.g., a delayed saccade (Nishida et al., 2014)), or integration of perceptual information. Therefore, although tau generally scales with position along the cortical hierarchy (Murray et al., 2014), there might be some further specificity within equivalent levels of the hierarchy according to the type of memoranda or task. More generally, both

the engagement of brain regions along the hierarchy according to diverse temporal integration demands (Runyan et al., 2017; Honey et al., 2012), and/or their weighting of cells according to tau seem to be at play (Scott et al., 2017; Bernacchia et al., 2013; Cavanagh et al., 2016) when processing complex cognitive tasks. It will be important to further investigate the recruitment of neurons according to their intrinsic timescales when mnemonic information has to be integrated with flexible task demands.

Our initial quantification of baseline autocorrelation (Fig. 3.1) adds to the accumulating evidence for a hierarchy of intrinsic timescales in the monkey (Chen et al., 2014; Murray et al., 2014; Ogawa and Komatsu, 2010; Chaudhuri et al., 2015; Cavanagh et al., 2016), mouse (Runyan et al., 2017) and human cortex (Hasson et al., 2008; Honey et al., 2012; Gauthier et al., 2012). Prefrontal areas lPFC and FEF showed longer intrinsic timescales than LIP, which coincides with their position in the cortical hierarchy (Felleman and van Essen, 1991). In the temporal domain, longer processing timescales directly relate to the integration or maintenance of information over time (Runyan et al., 2017), which is useful for higher cognitive functions such as decision making or WM (Runyan et al., 2017, Wang, 2013). Indeed, prefrontal areas and specifically lPFC, which showed the longest average intrinsic timescale, has served as the major focus for WM research over the past decades (Goldman-Rakic, 1995) and is generally considered crucial for WM function (Butters and Pandya, 1969; Stamm, 1969).

It is tempting to argue that neurons exhibiting longer intrinsic timescales form part of a continuous attractor network. Classical attractor models of WM persistent activity (Compte et al., 2000; Wang, 2001; Machens et al., 2005; Wong and Wang, 2006) imply functional clustering i.e. increased connectivity between clusters of similarly tuned

neurons, which has been observed and been linked to persistent activity in some tasks (Masse et al., 2017). Local connection density per se might be a good predictor of intrinsic timescales and in turn persistent activity (Goldman-Rakic, 1995; Runyan et al., 2017; Chaudhuri et al., 2015). In a recent study, Chaudhuri and colleagues constructed a large scale brain model aimed at investigating possible mechanisms leading to a hierarchy of intrinsic timescales similar to the one observed in Murray et al., 2014. They found that both long range connections ('inter-areal loops') as well as local connection densities determined the relative size of intrinsic timescales across brain areas. In line with this, a recent study in mice found that areas with higher functional coupling showed larger processing timescales on the level of the population (Runyan et al., 2017). Together, those observations raise the possibility that an individual intrinsic timescale might be determined by the concentration of incoming and outgoing connections a given neuron receives i.e. whether it forms part of a local 'network hub' (Chaisangmongkon, 2017). Our results for FEF and LIP indeed suggest that neurons with longer intrinsic timescales might pertain to such a network hub (Fig. 3.7). In lPFC we failed to observe a relationship between a neuron's functional connectivity and intrinsic timescale, possibly indicating more flexible connectivity regimes (Mongillo et al., 2008; Szatmáry and Izhikevich, 2010). It is important to note that spike count correlations do not necessarily indicate anatomical connectivity and their source does not have to be local (Cohen and Kohn, 2011). Finally, there exists a multitude of cell-intrinsic or network mechanisms underlying persistent activity accounts (Zylberberg and Strowbridge (2017)). This plurality naturally applies to possible sources for the heterogeneity of observed intrinsic timescales.

Through analyzing simultaneous electrophysiological recordings from three brain regions: lPFC, FEF and LIP, we found that the temporal tuning heterogeneity previously observed during WM is predicted by baseline firing stability of individual neurons. We showed that intrinsic timescales effectively shape the temporal dimensionality of the encoded information. In the prefrontal regions lPFC and FEF, neurons with longer intrinsic timescales encoded task information more strongly while storing it in a comparably stable, low dimensional format throughout the delay. In contrast, cells exhibiting short intrinsic timescales encoded relevant features in a more dynamic fashion, maintaining a high temporal coding dimensionality throughout the task, which in lPFC, could be robustly decoded even during the delay period. Additionally, in lPFC, the presented sample stimulus was more rapidly encoded by short timescale cells. Lastly, we presented some evidence for distinct mechanisms contributing to the emergence of intrinsic timescale heterogeneity. In summary, we suggest that WM constitutes the dynamic recruitment of neurons with different intrinsic timescales, possibly shaped by local connectivity regimes, to optimize functionally coexistent computations such as the encoding and storage of information.

Chapter 4. Interneuronal correlations at longer timescales predict decision signals for bistable structure-from-motion perception.

4.1 Abstract

Perceptual decisions are thought to depend on the activation of task-relevant neurons, whose activity is often correlated in time. Here, we examined how the temporal structure of shared variability in neuronal firing relates to perceptual choices. We recorded stimulus-selective neurons from visual area V5/MT while two monkeys (*Macaca mulatta*) made perceptual decisions about the rotation direction of structure-from-motion cylinders. Interneuronal correlations for a perceptually ambiguous cylinder stimulus were significantly higher than those for unambiguous cylinders or for random 2D motion during passive viewing. Much of the difference arose from correlations at relatively long timescales (hundreds of milliseconds). Choice-related neural activity (quantified as choice probability; CP) for ambiguous cylinders was positively correlated with interneuronal correlations and was specifically associated with their long timescale component. Furthermore, the slope of the long timescale - but not the instantaneous - component of the correlation predicted higher CPs towards the end of the trial i.e. close to the decision. Our results suggest that the perceptual stability of structure-from-motion cylinders may be controlled by enhanced interneuronal correlations on longer timescales. We propose this as a potential signature of top-down influences onto V5/MT processing that shape and stabilize the appearance of 3D-motion percepts.

---

*The experiment in this chapter has been published as: Wasmuht, D. F., Parker, A. J., Krug, K. Interneuronal timescales at longer timescales predict decision signals from bistable structure-from-motion perception. *Sci. Rep.* **9**, 11449 (2019)

4.2 Introduction

Understanding the relationship between the firing of cortical neurons and the formation of perceptual decisions has been one of the central goals in systems neuroscience (Barlow, 1972; Parker and Newsome, 1998). Investigations in the extrastriate, visual area V5/MT of macaque monkeys have played an important in this endeavour. V5/MT neurons show selective tuning for motion direction (Dubner and Zeki, 1971) and binocular disparity (Krug and Parker, 2011; Maunsell and Essen, 1983) and causally contribute to the percept of both motion and binocular depth (Krug et al., 2013; Salzman et al., 1990). It was first observed in V5/MT that neuronal activity correlated with the percept reported by the monkey on a trial to trial basis, even when the stimulus exciting the given neuron was fixed (Britten et al., 1996). A measure of this correlation is choice probability (CP), which refers to the probability that an independent observer correctly predicts an upcoming choice, solely based on the neuronal spiking activity. While significant CPs have been described in a number of sensory cortical areas and using different perceptual tasks (Britten et al., 1996.; Celebrini and Newsome, 1994; Dodd et al., 2001; Cook and Maunsell, 2002; Uka and DeAngelis, 2004; Uka et al., 2005; Gu et al., 2008; Liu et al., 2013; Nienborg and Cumming, 2009; Verhoef et al., 2010; Shiozaki et al., 2012), their origin is debated (Cumming and Nienborg, 2016; Krug, 2004; Wimmer et al., 2015).

Determination of the origins of choice-related firing in cortical sensory areas will shed light on the emergence and sequence of decision-making processes. In a simple feed-forward model, variability in choice is directly caused by variability in the evoked response of sensory neurons, captured in their CP. Early theoretical models considering

this interpretation, suggest that a decision is determined by comparing the CPs of pools of oppositely tuned neurons (Britten et al., 1996; Shadlen et al., 1996; Shadlen and Newsome, 1998). This bottom-up model can be contrasted with a top-down model, in which choice-related variability in the sensory response represents decision-linked modulatory feedback. In this case, the observed CP might be a reflection of latent internal states, such as attention or intrinsic biases, which may affect decision-making at multiple stages (Bondy et al., 2018; Herrington et al., 2009; Krug, 2004; Kwon et al., 2016; Nienborg and Cumming, 2009; Roelfsema and Spekreijse, 2001; Wimmer et al., 2015).

Theoretical and experimental evidence directly links CP to shared response variability among sensory neurons (Cohen and Newsome, 2008; Haefner et al., 2013; Parker, 2013; Pitkow et al., 2015; Shadlen et al., 1996), also termed noise- or spike count correlation (rSC). Specifically, a neuron that is not directly linked to a decision process might show a significant CP simply because it is correlated with neurons that are. Hence, CP observed in a neuron depends both on the neurons relative read-out weight, i.e. contribution to the decision, and the structure of its correlations with other neurons (Haefner et al., 2013; Shadlen et al., 1996). A variety of mechanisms may generate such shared interneuronal correlations in sensory cortical areas (Cohen and Kohn, 2011; Parker, 2013; Doiron et al., 2016; Kohn et al., 2016). First, correlations might arise between neurons that receive common input from sensory afferents. On this view, correlations should be largely driven in a bottom-up fashion, perhaps through factors such as eye movements (Bair et al., 1998) or stochastic fluctuations in the stimulus (Bair et al., 2001). Secondly, the common signal could arise through recurrent local connectivity (Helias et al., 2014; Pernice et al., 2011) including long range horizontal

connections between neurons of similar tuning properties (Rosenbaum et al., 2016). Such ideas find support in the spatial structure of interneuronal correlations since they become larger with physical proximity and similarity in stimulus preference of neuronal pairs (Bair et al., 2001; Kohn and Smith, 2005; Rosenbaum et al., 2016; Smith and Kohn, 2008). Thirdly, there also exists substantial evidence for a more global and flexible source for correlations, possibly provided by top-down efferent projections (i.e. feedback connections) (Cumming and Nienborg, 2016; Maunsell, 2015). This type of correlation is associated with processes downstream of the sensory area such as feature-based attention (Cohen and Maunsell, 2009; Mitchell et al., 2009; Ruff and Cohen, 2014a) or task instructions related to a decision (Bondy et al., 2018; Cohen and Newsome, 2008).

Turning to CPs, the question of their emergence therefore becomes linked in part to determining the source and structure of the underlying interneuronal correlations. For identification of potential sources for noise correlations, it is crucial to examine the timescales on which those correlations unfold. Noise correlations typically act on timescales in the range of a few tens to hundreds of milliseconds (Bair et al., 2001). Although measured timescales vary between studies (Cohen and Kohn, 2011; Ecker et al., 2010), it is thought that feed-forward i.e. bottom-up correlations, caused by the external stimulus, act on a faster timescale than those induced by top-down i.e. feedback signals (Nienborg et al., 2012; Zhang and Abbott, 2000) but see (Hupé et al., 2001). A recent study by Wimmer et al. (2015) used a hierarchical network model consisting of a sensory and a higher order integration circuit, which was trained on the classic random dot motion task. They found that CP measured in the sensory layer could be explained by a combination of a fast-acting bottom-up signal and a slower top-down component

from the integration layer. The latter was proposed to increase over time injecting additional correlations into the sensory layer, which led to higher measurable CPs in those neurons.

Here, we recorded from V5/MT neurons while monkeys judged the rotation direction of a structure-from-motion (SFM) cylinder. Perceptual interpretation of this stimulus requires integration of motion direction and binocular depth information (Krug et al., 2013). Previous studies have shown that the perceptually bistable version of the cylinder stimulus (i.e. the ambiguous cylinder) is associated with relatively large CPs when compared to the motion discrimination task (0.67 vs. 0.56 respectively) (Britten et al., 1996; Dodd et al., 2001). This might reflect a strong cylinder percept even in the ambiguous case and differences in population coding for visual motion and stereo depth. However, this comparison is taken across different studies in different laboratories with animals trained on different tasks. Applying an interleaved design including cylinders with varying disparity and random motion stimuli, we investigated spike count correlations, the timescale of the underlying neuronal interactions, and their relationship to the time-course of choice related activity, measured as CP. We find that the ambiguous version of the cylinder is associated with high interneuronal correlations at longer timescales. Furthermore, the long timescale component of these correlations predicts the large CP values observed for the ambiguous cylinder. We propose that this distinct correlation signature may reflect a top-down influence onto V5/MT processing, which could convey and stabilize the perceptual appearance of bistable visual stimuli.

4.3 Methods

Physiological recordings have been reported in detail before (Dodd et al., 2001; Krug et al., 2004). In brief, two adult rhesus monkeys (*Macaca mulatta*) were trained to perform a two alternative forced choice task (2AFC) for fluid rewards. A headpost, scleral search coils and a recording chamber were implanted. We recorded a total of 93 sites (29 in monkey bill and 64 in monkey rin) in extrastriate visual area V5/MT using single microelectrodes. Stimuli and perceptual tasks were optimised for the disparity and motion tuning of each recorded single neuron. All procedures were performed in accordance with United Kingdom Home Office regulations and European Union guidelines on animal experimentation (EU directive 86/609/EEC; EU Directive 2010/63/EU). Animal protocols passed local and national ethical review and were licensed by the UK Home Office.

Stimuli and experimental paradigm

Cylinder stimuli are defined by two transparent planes of randomly placed (seeded anew every trial – i.e. not frozen noise) black and white dots moving in opposite directions on a mid-gray background; for further details of the stimuli, set-up and task see (Dodd et al., 2001). The velocity of the dots is a sinusoidal function of spatial position, which gives rise to a sensation of depth-from-motion. When no depth order is specified (zero binocular disparity), the direction of rotation of the cylinder is ambiguous. For long durations of viewing (several seconds or more), the stimulus is perceptually bistable and spontaneously flips its perceived configuration (Krug et al., 2008). For shorter durations (including the 2s presentations used here), the stimulus is stable but changes from one

presentation to the next (Dodd et al., 2001; Nawrot and Blake, 1993). Adding binocular disparity to the moving dots provides sensory evidence that disambiguates the motion percept into rotating clockwise (CW) or counterclockwise (CCW). The amplitude of the disparity signal controls the degree of perceptual ambiguity of the cylinder. As the amplitude of the disparity signal increases, the variability of the judgments decreases and the reported motion rotations resolve into unambiguous decisions in favour of CW or CCW rotation.

The ambiguous and unambiguous cylinder stimuli were presented in a pseudo-random sequence of trials, into which presentations of random motion stimuli were also pseudo-randomly interleaved. These stimuli had a binocular disparity of zero and their dots had a lifetime of one frame. At the beginning of each trial the monkeys were required to fixate on a point in the middle of the screen. Upon successful fixation, a stimulus was presented in the receptive field of the recorded neuron for 2s. On cylinder trials, two targets appeared on either side of the fixation marker after stimulus presentation and the monkey had to indicate its perceptual choice about rotation direction with a saccade. Animals received a fluid reward for maintaining fixation and making a correct choice about the rotation direction of the unambiguous cylinder. Reward was given at random for the ambiguous cylinder. In response to presentations of random motion stimuli, animals were rewarded for maintenance of fixation only, no choice targets appeared and no choice was required.

Spike sorting

We used Parylene-coated tungsten electrodes with an impedance of around 0.5 – 1.2 MOhm (Microprobe Inc.) that allowed clear single unit (SU) recordings. After amplification (BAK electronics), filtering (200 Hz to 5kHz) and digitization (32 kHz), electrode signals were stored to disc as well as classified online (DataWave Technologies). Online, experiments were only carried out when a SU could be clearly separated from the lower amplitude multi-unit (MU) activity and the SU had clear direction and disparity tuning, to which the test stimuli were matched (for details of this experimental procedure see Dodd et al. (2001)).

Spike sorting was repeated *de novo* on the stored electrode signals offline in a custom-built C-programme. For each recording site to be included, the SU activity had to have a clearly distinguishable spike waveform that was visually checked and maintained throughout all the experiment runs. Quantitative analysis of the spike waveforms was performed using a range of parameters, most commonly 'spike width', 'spike energy', 'spike excursion'. In order to be included the SU spike values had to form a separate cluster away from the remaining MU activity – for all measures. The selection for the SU was adjusted until only the spikes forming this cluster were included. The MU activity was sampled by taking signals above a Schmitt Trigger level adjusted for each experiment minus the identified SU activity, so that the MU activity for 0° disparity cylinders was between 10s to 200 spikes/s.

There inevitably is a small fraction of spikes that are misclassified in any computational sorting method. To test whether our results were robust to small errors in spike classifications we simulated deliberate misclassifications. In short, we assigned varying fractions of random spikes from the SU to the MU and vice versa to create shuffled spike

trains. We repeated the analysis of our main results (Fig. 4.3) using the shuffled spike trains to confirm that our results were not explained by moderate errors in spike sorting (see Supp. Fig. 4.10).

Data preprocessing

All analysis was performed on spike sorted data. For all correlation analysis we compared the isolated SU with the exclusive MU response recorded on the same electrode. We excluded sites with a firing rate < 10 spikes/s estimated across trials and excluded without a spike. Furthermore, we required activity to be recorded for at least five trials for a given condition. Those criteria eliminated 5 sites from our original sample of 93 sites (n=88). Included sites showed significant SU tuning for binocular disparity over a wider range (±0.9°) and direction of motion (360°) ($p<0.05$, *ANOVA*). Analyses were implemented using custom code written in Matlab (The MathWorks Inc., Natick, MA).

Estimation of tuning parameters

Fitting of tuning curves and parameter estimation was performed on the square root of the firing rate (spikes/s) for each 2 s stimulus presentation to remove the linear relationship between variance and firing rate (Prince et al., 2002; Tolhurst et al., 1981). Curve fitting was performed using the Matlab implementation of Nelder-Mead simplex algorithm (Nelder and Mead, 1965), which sought a local minimum to the squared error. All procedures were carried out for SU and MU activity, if not indicated otherwise.

To estimate motion direction tuning, recording sites were probed with a patch of dots

coherently moving across the receptive field at a minimum of 6 different directions, over a range of 360°. Mean responses to motion directions were fit with a von Mises distribution. Von Mises distributions generally provided excellent fits to the data, accounting for >90% (median across neurons) of the variance in the mean response across directions. Tuning for the direction of cylinder rotation (CW vs CCW) was estimated in responses to cylinder stimuli presented over a narrow range of cylinder disparities around 0°: - 0.06° to 0.06°. Here, negative disparities indicate CW rotation whereas positive disparities indicate CCW rotation. A cumulative Gaussian was fitted to the mean response for each cylinder disparity, applying the same procedure as for the direction of motion tuning. Cumulative Gaussians provided a good fit accounting for >90% (median across neurons) of the variance in the mean response across cylinder disparities.

To extract parameters describing the strength of selectivity to a given parameter (motion direction or cylinder rotation), we calculated the F-index (Prince et al., 2002), given by

$$F_{index} = \frac{MS_{treatment}}{MS_{error} + MS_{treatment}} \tag{4.1}$$

where, MS_{error} and $MS_{treatment}$ represent the familiar mean square terms from a one-way ANOVA performed on the trialwise responses to either cylinder stimuli or moving random dot stimuli. MSE_{error} is the mean within disparity or motion direction variance, and $MS_{treatment}$ is the mean between disparity or motion direction variance. The F_{index} quantifies the modulation of the tuning function by the tested parameters (neuronal property) while taking into account the reliability of the measured responses (experimental property) (Prince et al., 2002). We calculated $F_{indices}$ for both cylinder

rotation ($F_{cylinder}$) and motion direction tuning (F_{motion}), respectively. Tuning of MU and SU was deemed as strong if F_{index} was greater than 0.1.

Tuning similarity between SU and MU responses were evaluated using the r_{Signal} metric, which refers to the Pearson correlation between two raw tuning functions. Additionally, for cylinder rotation tuning, we assessed similarity by comparing the maxima of the two cumulative Gaussians fits for SU and MU, which had to lie within the same rotation direction (CW vs CCW) as indicated by the sign of the disparity. For motion direction, the maxima of the von Mises fits had to lie within 90° of each other to be considered as similarly tuned.

Spike count correlations

Spike count correlations (rSC) (also known as noise correlations) (Cohen and Kohn, 2011; Zohary et al., 1994; cacultions of rSC equal those on **chapter 3**) were calculated using repeated trials within a condition. Spikes were counted for each trial over the full 2 s duration and converted to spikes/s. To avoid contamination of our estimates of r_{sc} by outlier responses, we removed trials for which the response of either SU or MU was >3 SDs different from its mean response computed over all trials (Zohary et al., 1994). We then normalized firing rates on each trial for both mean firing and slow drifts in neural excitation by computing a z-score for SU and MU firing rate on each trial using a sliding window of 5 trials before and after the current trial. We used Eq. 3.4 from **chapter 3** to compute z-scores. We then computed rSC as the Pearson correlation between the SU and MU z-scores. All statistics involving rSC values were carried out after transforming them to Fisher's z.

Spike time correlations

We computed spike train cross-correlograms (CCGs) and auto-correlograms (ACGs) (Perkel et al., 1967) using binary spike trains at a 1 ms resolution. The trial-averaged cross correlation of two spike trains is defined as:

$$C_{jk}(\tau) = \frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T} x_j^i(t)x_k^i\,(t+\tau) \tag{4.2}$$

Where x_j^i or x_k^i can be 1 or 0 depending whether neuron *j* or *k* (= MU and SU) fire an action potential at time-point *t* in trial *i*. *T* is the duration of the trial in milliseconds and *M* the number of trials. The parameter τ defines the time-lag (in ms) i.e. the time-window the cross-correlation is computed over. For the auto-correlation, *j=k*. To compute the CCGs and ACGs, the trial averaged cross-correlation is simply divided by the product of a triangular function and the geometrical mean of the mean firing rates of neuron *j* (λj) (e.g. SU) and *k* (λk) (e.g. MU):

$$CCG(\tau) = \frac{C_{jk}(\tau)}{\Theta(\tau)\sqrt{\lambda_j * \lambda_k}} \tag{4.3}$$

The triangular function $\Theta(\tau)$ represents the extent of overlap of the spike trains as a function of the discrete time-lag τ:

$$\Theta(\tau) = T - |\tau|, \ (-T < \tau < T) \tag{4.4}$$

i.e., that there are *T* opportunities for simultaneous events in a trial of length *T* but only *T*-1 opportunities for coincidences at time-lags of 1 ms, etc. Substituting equation (4) into equation (3) corrects for the triangular shape of the C_{jk} caused by the finite duration of the data. We chose to normalized the CCGs and ACGs by the geometric mean spike rate, as opposed to the product of spike rates or another factor, for two reasons. First, it is the most commonly used normalization, so it facilitates comparison with previous studies (Bair et al., 2001; Eggermont and Smith, 1996, 1995; Kohn and Smith, 2005; Mastronarde, 1983; Solomon et al., 2015) [similar to the normalization of the joint peristimulus time histogram (Aertsen et al., 1989)]. Second, this normalization provides results that are most comparable with measures of spike count correlation (Bair et al., 2001; Kohn and Smith, 2005). In order to correct for stimulus induced co-variations in SU & MU firing rates, we computed a shift predictor:

$$C_{jk}^{*}(\tau) = \frac{MS_{jk}(\tau) - C_{jk}(\tau)}{M - 1} \tag{4.5}$$

In which S_{jk} denotes the cross correlation of the post-stimulus time histograms (PSTH) (Bair et al., 2001). Substituting C_{jk}^{*} for C_{jk} in equation (3) gives the final shift predictor, which was then subtracted from the CCGs and ACGs, respectively. We also computed CCGs, ACGs and shift predictors after excluding the first 300ms of each trial to avoid processing the initial transient onset response, which could influence correlation strength. However, we found only negligible differences.

To calculate the timescale of correlations we used the rCCG metric derived by Bair et al. (2001).

$$rCCG(\tau) = \frac{\int_{\tau=-t}^{t} CCG(\tau)}{\sqrt{\left(\int_{\tau=-t}^{t} ACG_j(\tau)\right)\left(\int_{\tau=-t}^{t} ACG_k(\tau)\right)}} \tag{4.6}$$

The rSC is equal to the integral over the full shift predictor corrected CCG i.e. time-lags covering the full duration of the trial, divided by the geometrical mean of the respective integrals of the ACGs (also corrected by the shift predictor) of the MU and SU (Bair et al., 2001). Integrating different areas around the central peak of the rCCG allows for an assessment of different timescales of correlations.

Choice probabilities

Choice probabilities (CPs) were calculated on the mean SU firing rate in response to the ambiguous cylinder over the stimulus duration (2s). Trials were pooled according to the choice made by the monkey (CW vs CCW). CP is inferred by quantifying the separation between the two distributions. Specifically, this is done by calculating the area under the ROC curve (Britten et al., 1996; Green and Swets, 1966). We included neuronal data only from blocks of trials in which the animal's behavioural performance demonstrated that it was accurately responding to the cylinder disparity. Therefore, we required that the animal scored over 80% correct at the largest values of cylinder disparity tested, typically 0.03°, and that the animal's responses to the ambiguous cylinder were unbiased (experiments in which the same target was selected on more than 80% of the ambiguous

cylinder trials were discarded). Our selection criteria led to the exclusion of 22 recording sites from our initial sample (n=66).

Statistical testing using cluster-based permutation tests

Throughout this study we used a non-parametric cluster based permutation test (Maris and Oostenveld, 2007). The method compares some observed test statistic with a constructed null distribution while controlling for multiple comparisons across time. Depending on the question at hand we constructed null distributions in three ways: 1) By randomly permuting condition labels when comparing cylinder conditions with each other or with the random motion condition. 2) By randomly permuting CP values across sites when comparing median splits of the population or computing correlation coefficients across the population. 3) By randomly permuting correct and incorrect trials when evaluating CPs. The relevant test statistic or raw difference was computed for the observed data as well as for each of 10000 permutations. For 1) and 2), a two-sample t-test was applied for each time-point and observed clusters of contiguous p-values < 0.05 were compared against the distribution of maximal clusters obtained from the permutations. If the size of the observed cluster exceeded the 95^{th} percentile of the null distribution it was deemed significant. In one case – Fig 4.6c – a Pearson correlation coefficient was computed. For 3) the observed CP value was compared to the constructed distribution of CP values and was classified as significant if it exceeded the 95^{th} percentile of that distribution.

Calculating time-courses of spike count correlations and choice probabilities

To calculate the time-course of rSC or CP over the course of the trial, we counted all spikes falling into a sliding 100 ms time-window (T) (1ms bin-size). CPs and rSC values were then calculated and de-trended as defined above - across trials but within a given time-window.

To obtain rSC matrices, we restricted our spike count data to come from independent time-windows (i.e. time-bins) by averaging across time-points centred at t = T/2, 3/2T, 5/2T and so on resulting in 20 time-bins for the trial duration of 2s. Then, for each site we calculated the rSC between time-bin t_i across trials in neuron one (i.e. SU) with the same and any other time-bin t_j across trials in neuron two (i.e. MU). All pairwise combinations give the rSC matrix for a site. The average of those matrices gives the rSC matrix for the tested population.

To estimate the time-course of rSC at varying time-lags across the trial, for each time-bin t_i we averaged the corresponding leading and lagging correlation values obtained from our rSC matrix (per site). For simplicity, we refer to any rSC values derived in that way as rSCτ, where the τ represents the dynamically chosen number of leading and lagging rSC values making up rSCτ. To investigate correlations at longer timescales, τ was set to four, which corresponds to an effective time lag/lead of about -450ms to +450ms, excluding the main diagonal pairs where $t_i = t_j$ i.e. τ=0 (i.e. +/- 50 ms). We also examined separately the main diagonal, i.e. $\tau = 0$. In this case rSCτ was referred to as instantaneous correlation.

To identify trends in the rSCτ time-courses, we fitted a 2nd order polynomial to the rSCτ values obtained for each time-bin and took the derivative of our fitted function. Although

estimated from independent time-bins, this derivative is now available for each time-point. To test whether the rising or decaying rSCτ predicted CP at any time-point in the trial, we calculated the Pearson correlation between derivatives obtained for each site at any time-point with the CP calculated for the SU at the same or any other time-point. Finally, to analyse how trends in rSCτ estimated over the whole trial affect CP time-course, we took the average derivative of our polynomial fits to rSCτ and split our population into sites that showed a generally rising (derivative > 0) or decaying (derivative < 0) rSCτ time-course. After these steps, we computed the average CP time-courses for each subpopulation.

4.4 Results

Two monkeys (*Macaca mulatta*) performed a perceptual decision-making task, judging the rotation direction of structure-from-motion (SFM) cylinder stimuli that had different degrees of ambiguity (defined by the specific conjunction of motion direction and binocular disparity). Cylinder stimuli were pseudo-randomly interleaved with random dot motion stimuli with zero net motion, for which no decision was required. While monkeys were performing the cylinder task, we recorded electrical activity from 88 sites (after exclusion, see Methods) in cortical area V5/MT with single electrodes. Visual stimuli were optimized online for the disparity- and direction-selectivity of the recorded single neuron. Single unit activity (SU) was isolated from its surrounding multi-unit activity (MU) to quantify interneuronal correlations in response to the different conditions. SUs included in this study were tuned for binocular disparity and direction of motion ($p<0.05$, *ANOVA*). Some of the SU data have been previously reported (Dodd et al., 2001). The central aim of this study was to uncover interneuronal correlation

structures underlying the processing of bistable SFM stimuli during a perceptual decision-making task.

Higher spike count correlations for the bistable cylinder stimulus than for random motion.

First, we directly compared interneuronal correlations in response to ambiguous cylinder stimuli (zero disparity; evoking a bistable percept) to those evoked by a random dot stimulus (Newsome et al., 1989; Zohary et al., 1994) in order to assist comparison with previous studies. Interneuronal correlations were quantified as trial-by-trial spike count correlations (rSC; also termed noise correlations; see also **Chapter 3**), which were calculated for the two stimulus types separately. In brief, for each site we correlated the z-scored SU and MU activity for repeated trials of the same stimulus type (Fig. 4.1a). The resulting correlation coefficients (rSC) represent the common trial-by-trial fluctuations of shared 'noise' between the SU and surrounding MU in response to a given stimulus. The average rSC for all sites probed with the ambiguous cylinder stimulus was rSC=0.3 (n=88) (Supp. Fig. 4.1), rising to rSC=0.37 (n=52) for sites with matching and strong SU and MU tuning (see Methods). In contrast, rSC in response to the random motion stimulus was rSC=0.16 (Supp. Fig. 4.1) on all sites probed with this stimulus (n=53) and rSC=0.17 on sites with matching and strong direction tuning (see Methods) (n=41). The rSC values obtained for the random motion stimulus are in accordance with those found previously in V5/MT, analysing correlation for pairs of SUs recorded from a single electrode (Bair et al., 2001; Zohary et al., 1994).

Next, we directly compared rSC values for bistable cylinders and random motion for

sites that were recorded with both stimuli pseudo-randomly interleaved (n=52). rSC was consistently higher in response to the ambiguous cylinder than to random motion at the same site (Fig. 4.1b; $p<0.001$, n=52, *sign test*). This difference is also evident when comparing mean rSC for both stimuli ($p=0.002$, n=52; *t-test on Fisher's z*) (Fig. 4.1c) and holds true for cells with matching and strong tuning ($p=0.023$, n=26) despite an increase in rSC values for both stimulus types. Data analysed separately for individual monkeys showed the same trend (Supp. Fig. 4.2).

Spike count correlations can depend on firing rate (de la Rocha et al., 2007). However, we find no significant correlation between rSC and mean firing rate (geometric mean of SU and MU firing rates), within either stimulus condition (ambiguous cylinder: $r=0.017$, $p=0.9$; random motion: $r=-0.09$, $p=0.53$; *Pearson correlation*; n=52 sites), nor is there any significant correlation between the change in rSC between the two stimuli and the difference in mean firing rate ($r=-0.1$; $p=0.46$; *Pearson correlation*; n=52 sites). Thus, the stimulus-related differences in rSC were not merely linked to changes in mean neuronal firing.

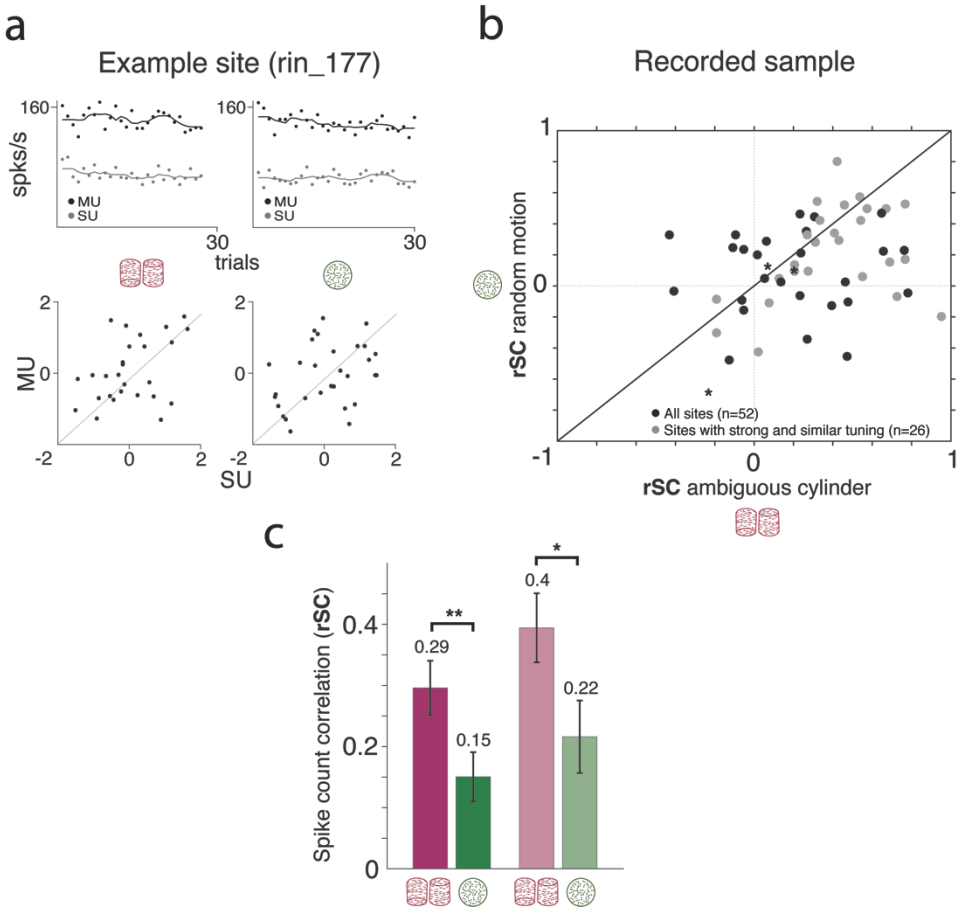


Figure 4.1: Comparing spike count correlations for the ambiguous cylinder and random dot stimulus. **a)** Example site (rin_161); recorded activity in response to an ambiguous cylinder stimulus (left; indicated by the purple schemata) and random motion (right, green schemata). Upper two panels: Dots represent trialwise average firing rate (spks/s) of the SU (grey) and MU (black) recorded on the same electrode. Solid lines indicate the running mean computed over 11 trials centred on the current trial. Lower two panels: Scatter plots of SU and MU z-scores for each stimulus. The Pearson correlation between the z-scores gives the rSC (left; rSC=0.43; right: rSC=0.29). **b**) The rSC values for the random motion stimulus (y-axis) are plotted against the rSC values for the ambiguous cylinder (x-axis) from the same sites. Data points show all sites that were recorded with both stimuli pseudo-randomly interleaved (n=52). Grey dots show sites that showed strong and matching SU and MU tuning (n=26 of 52) (see Methods). **c**) Bar plot comparing the average rSC between the two stimuli; ambiguous cylinder (purple) and random motion (green). Bright bars show average values for all sites (n=52); pale bars show average values for sites with strong and matching SU and MU tuning (n=26 of 52). *p<0.05, **p<0.01; *t-test on Fisher's z*. Error bars indicate the s.e.m.

Previous studies observed a positive relationship between rSC and measures of tuning similarity in visual cortex (Cohen and Kohn, 2011), specifically in area V5/MT (Bair et al., 2001; Huang and Lisberger, 2009; Solomon et al., 2015). Those observations suggest that shared common input endows nearby neurons with similar tuning properties and

makes them subject to similar noise sources. We quantified tuning similarity using the Pearson correlation between SU and MU tuning curves (r_{Signal}). For the coherently moving random dot stimulus, average r_{Signal} was large and positive: r_{Signal} = 0.68; and similar to that for ambiguous cylinders: r_{Signal} = 0.62. There was a weak relationship between rSC and r_{Signal} for the ambiguous cylinder (r=0.25, p=0.017, n=88; *Spearman rank*). This observation was matched by a similar, non-significant trend for the random motion stimulus (r=0.21, p=0.16: *Spearman rank*) (Supp. Fig. 4.3a). Additionally, for pairs with matching tuning properties (see Methods) we find a positive correlation between rSC and the mean tuning strength i.e. mean $F_{cylinder}$ of SU/MU pairs for the ambiguous cylinder (r=0.32, p=0.006, n=71; *Spearman rank*). This effect remained significant when applying multiple linear regression (using z-transformed correlation values), factoring out firing rate (β=0.48, p=0.008). Again, for the random motion stimulus, a similar trend was observed (r=0.24, p=0.13, n=41; *Spearman rank*) (Supp. Fig. 4.3b). Tuning strength for motion direction (F_{motion}) was assessed on a coarse scale and therefore appeared larger than its equivalent for the cylinder rotation, for which $F_{cylinder}$ was assessed with near-threshold disparities close to zero.

Higher spike count correlations for the cylinders are linked to perceptual ambiguity.

Ambiguous cylinder and random motion trials differed in both the visual stimulus and the task performed (perceptual choice for cylinders, fixation for random motion). But all recording sites were also probed with a number of non-zero disparity cylinder stimuli, which made it possible to compare rSC for different levels of stimulus ambiguity during performance of the same task. We observed that rSC decreased monotonically as the absolute value of disparity rendering the cylinder unambiguous increased (p=0.01; *one-*

way Anova on Fisher's z) (Fig. 4.2b). We only included cylinder disparities that were probed for at least ten sites. Thus, the probed disparities ranged from 0° to 0.03° in steps of 0.01°. On average, the ambiguous cylinder evoked a larger rSC than the unambiguous cylinder, when different non-zero disparity cylinder were pooled for each site ($p=0.02$, $n=50$, *paired t-test on Fisher's z*) (Fig. 4.2b, left panel). The difference became even more pronounced when we limited our analysis to sites that were recorded with all three stimulus conditions ($p=0.006$, $n=24$, *t-test on Fisher's z*) (Fig. 4.2b right panel). There was no significant difference between the average rSC for unambiguous cylinder and the random motion condition ($p=0.65$, $n=24$, *t-test on Fisher's z*). Taken together, those results suggest that the level of perceptual ambiguity of the cylinder stimulus and not the decision task itself, is linked to high rSC. The results plotted in Figure 4.2 apply for the case of the sub-set of neurons with strong and matching tuning but differences between stimuli were qualitatively similar when taking all recorded cells into the analysis. Data analysed separately for individual monkeys demonstrated the same trends (Supp. Fig. 4.4)

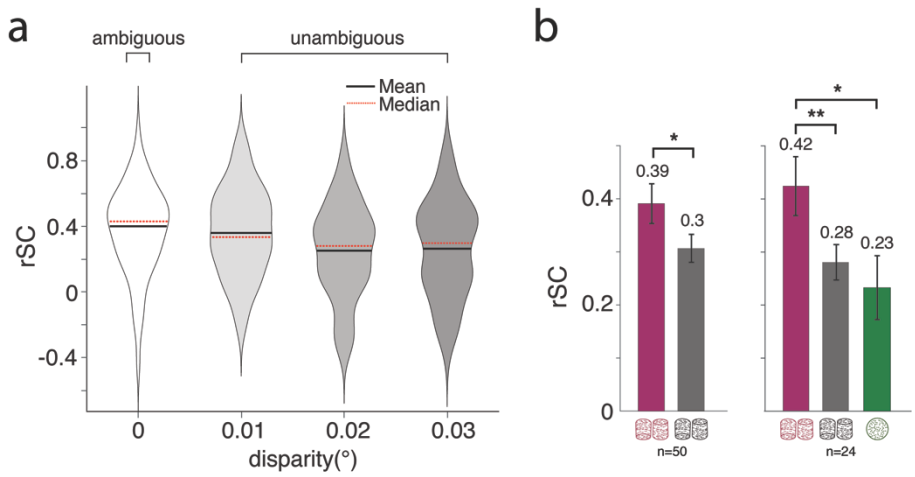


Figure 4.2: Spike count correlations for intermediate cylinder disparities. a) Violin plots show distributions of rSCs for cylinder stimuli with different disparity strengths (p=0.01; *one-way Anova on Fisher's z*). Only disparity values tested on more than 10 recording sites were included; negative disparities were inverted so that each distribution is made up of rSC values obtained from both negative and positive cylinder disparity (i.e. both rotation directions). **b)** Bar plots indicate mean rSC values for three stimulus conditions (purple: ambiguous cylinder; grey: unambiguous cylinder; green: random motion). Only sites with strong and matching MU and SU tuning (see Methods) were included. Left panel: Mean rSC values for all sites that were tested with the ambiguous cylinder (zero disparity) and at least one unambiguous cylinder stimulus (i.e. non-zero disparity). Right panel: Mean rSC values for sites that were tested with all three conditions i.e. stimuli. *p<0.05; **p<0.01; *t-test on Fisher's z*. Error bars indicate s.e.m.

Longer timescales of interneuronal correlations for ambiguous SFM cylinder

Our data show a higher rSC in response to ambiguous cylinders than to the unambiguous version of the same stimulus or to zero-coherence random dot motion. However, spike count correlations were measured using the average firing rate across the whole trial. In the following, we used the spike times from the same data to compare the cylinder and random motion stimuli in terms of spike time cross-correlations. This finer grained analysis allows inference of possible sources and structure of the correlation signal underlying the difference evoked by distinct stimuli.

We cross-correlated SU and MU neuronal spike trains on a trial-by-trial basis to obtain spike time cross-correlograms (CCGs) for each site and each stimulus condition, based

on the method of Bair et al., (2001) (Bair et al., 2001) (Fig. 4.3a). We compared the average CCG for sites with strong and matching tuning (n=24) that were recorded under all three stimulus conditions (ambiguous cylinder, unambiguous cylinder and random motion) (Fig. 4.3a). The normalized area under the full CCG provides values mathematically identical to rSC (Bair et al., 2001). Therefore, integrating the CCG over variable lags can reveal the timescale over which rSC arises. We plotted the integral under the CCG, termed rCCG, as a function of the time-lag (τ) (i.e. half width of the integration window centered at ~0) (Fig. 4.3b). At small integration windows τ<20ms, rCCG for all stimuli were similar. At longer time-lags (τ>20-30ms), rCCG values for the ambiguous cylinder kept increasing until around τ=400ms, becoming significantly larger than for both unambiguous cylinders and random motion ($p<0.01$; *cluster-based permutation test*). rCCG time courses for the two latter stimuli appeared to plateau relatively early (τ ~ 10-30ms) and were not significantly different from each other. Our observations for random motion stimuli correspond to those previously reported for SU-SU pairs, which found that rCCG for random motion reliably captured most of the rSC at a τ of 32ms (Bair et al., 2001). As expected, reported rCCG values approach corresponding rSC values for each condition (Fig. 4.3b).

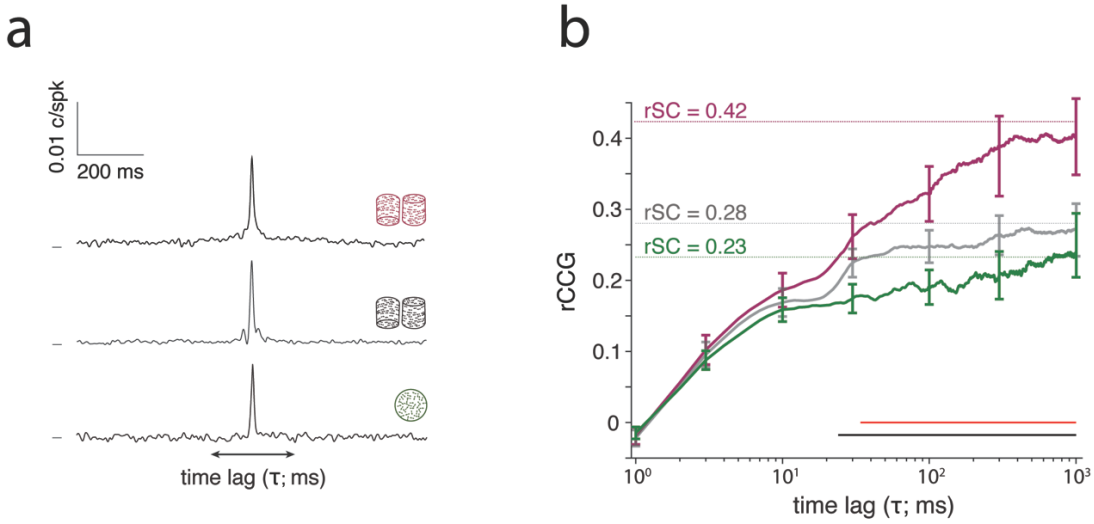


Figure 4.3: Spike time correlations and timescales of correlations. a) Averaged CCGs for all SU/MU pairs recorded for the ambiguous cylinder (purple), the unambiguous cylinder (grey) and the random motion (green) stimuli. Only sites were included that showed matching and strong SU and MU tuning (see Methods) (n=24). CCGs were smoothed with a Gaussian kernel (10ms s.d.). **b)** Average rCCG plotted as a function of the time-lag (τ) i.e. the integration window; for the ambiguous cylinder (purple), the unambiguous cylinder (grey) and the random motion (green) stimuli. Horizontal black and red lines indicate a significant difference between the purple and green and the purple and grey line, respectively ($p<0.01$, *cluster based permutation test,* n=24). Horizontal dashed lines indicate the corresponding average rSC values. rCCG courses start off at values below zero, because at time-lag zero there are no coincident spikes due isolating the SU from the MU activity on the same electrode. Error bars indicate the s.e.m.

These data suggest that while a substantial part of rSCs evoked by all three stimuli arises similarly on short timescales (around $\tau = 10$ms), a large part of the correlation associated with the ambiguous cylinder can only be explained by correlations on longer timescales ($\tau \sim$ tens to hundreds of milliseconds). These results therefore imply distinct sources of correlations that distinguish the processing of the different stimuli. Results were also replicated for each monkey separately (Supp. Fig. 4.5).

Choice probabilities and timescales of correlations

Noise-correlations have been linked to perceptual choice signals measured as choice

probability (CP) in a number of models (Chicharro et al., 2017; Haefner et al., 2016, 2013; Pitkow et al., 2015; Shadlen et al., 1996; Wimmer et al., 2015). Perceptually bistable SFM cylinder stimuli show relatively large CPs when compared to the random motion stimulus (0.67 vs 0.56) (Britten et al., 1996; Dodd et al., 2001). Therefore, we tested whether the size of the CP is associated with the magnitude or timescale of interneuronal correlations.

We quantified the CP for the ambiguous cylinder stimulus and for the isolated SUs. We could not quantify CP for random motion presentations, since no choice was required of the monkeys. The average CP for all sites assigned a CP was 0.67 (n=66; which was significantly different from chance (0.5), $p<0.01$, *permutation test*), the same value found by Dodd et al. (2001) (Dodd et al., 2001), with which this data set overlaps. For sites with strong and matching SU-MU tuning for cylinder rotation the average CP was even higher: 0.69 (n=39). Within this sample, neurons with high CPs for the cylinder were associated with higher spike count correlation values ($r=0.43$; $p=0.006$; n=39; *Pearson's r*) (Fig. 4.4a). As expected from previous assessments of neurometric thresholds (Krug et al., 2016), we also observed a strong correlation between CP and tuning strength measured as $F_{cylinder}$ ($r=0.65$; $p<0.001$; n=39; *Pearson's r*). Since rSC also depended on $F_{cylidner}$ ($r=0.29$, $p=0.05$), any relationship between CP and rSC might therefore be confounded by the correlation of either with $F_{cylinder}$. To control for this, we computed the partial correlation between CP and rSC, taking into account the effect of $F_{cylinder}$ ($r=0.34$; $p=0.03$, *linear partial correlation*), which, although reduced, remains significant.

Next, to link CP to timescales of correlations, we split the population of neurons with

strong and matching tuning (n=39) by the median CP and computed the average rCCG courses for each half (Fig. 4.4b). Additionally, we computed the average rSC values for each half. As before, neurons with higher CPs (CP > 0.68) showed higher spike count correlations (0.48 vs. 0.3; p=0.03; *un-paired t-test*). Differences between the rCCG time courses for each half appeared to emerge on relatively long timescales (τ >~30ms), but differences indicated a trend that was not significant. Taken together, high CP values associated with the ambiguous cylinder might be explained by interneuronal correlations acting on longer timescales. Trends were similar when data were analysed for each animal separately (Supp. Fig. 4.6).

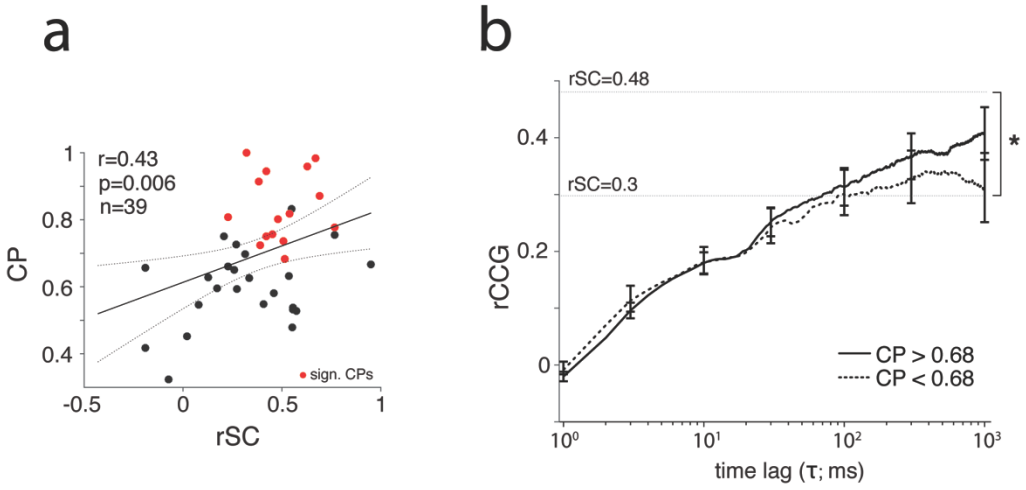


Figure 4.4: Choice probabilities and spike count correlations. a) Scatter plot of SU CP and SU-MU rSC. Each dot represents a SU that showed strong and matching cylinder rotation tuning and for which a CP value could be measured (see Methods; n=39). Red dots indicate SUs with a significant CP (p<0.05; *cluster based permutation test*). Black solid line indicates a linear fit to all data points with 95% confidence interval (dashed lines) and displayed correlation coefficient and associated p-value stem from the Pearson correlation. **b)** Average rCCG for subpopulations of sites split by the median CP associated with the SU (median CP = 0.68). Solid line: CP > 0.68; Dashed line: CP < 0.68. Error bars indicate the s.e.m. Horizontal lines indicate corresponding average rSC values. *p<0.05; *un-paired t-test.*

Interneuronal correlation time-course predicts choice probability.

So far, the measures of interneuronal correlation (rSCs, rCCGs) used here do not distinguish processes happening at specific time-points during the trial. A recent modelling study (Wimmer et al., 2015) using the random motion stimulus found that correlations induced by sensory input arose relatively early in the trial and stayed constant during the trial, showing only weak time-lagged components. In contrast, correlations due to a top-down feedback component rose throughout the trial and showed a strong component at longer time-lags. Furthermore, time-courses of lagged top-down induced correlations were predictive of CP signals (Wimmer et al., 2015). Here, we characterized interneuronal correlations arising at different time-points and timescales during the trial for all stimulus conditions and show how those correlations can predict CP trajectories for the ambiguous cylinder.

First, we computed rSC time-courses using a 100ms sliding window averaged over sites that were recorded under all three stimulus conditions and that showed strong and matching tuning (n=24) (Fig. 4.5). This instantaneous rSC for the ambiguous cylinder ramps up relatively early in the trial, when it is significantly higher than for the random motion stimulus (p=0.01, *cluster-based permutation test*). It is highest towards the middle of the trial (~800ms) [significantly higher than for both random motion (p=0.001) and unambiguous cylinder (p=0.01) stimuli]. During the second half of the trial instantaneous rSC values are similar in magnitude for all three stimulus conditions. Those results indicate that the higher correlations setting apart the ambiguous cylinder arises during the first half of the trial. Similar results were obtained for each monkey separately (Supp. Fig. 4.7).

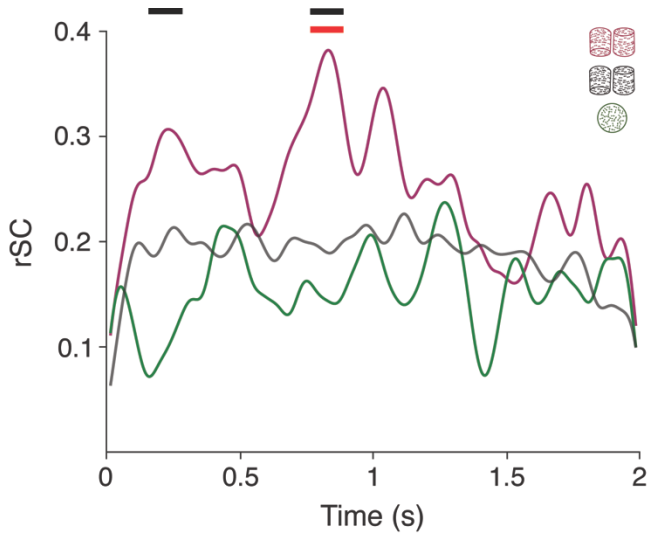


Figure 4.5: Time-course of spike count correlations. Average rSC time-courses estimated from a sliding 100ms window. Purple, grey and green lines show the average rSC time-courses for the ambiguous cylinder, the unambiguous cylinder and random motion, respectively. Averages were computed over all sites recorded under all three stimulus conditions (n=24). The black horizontal bar indicates a time window with a significant difference between the ambiguous cylinder and the random motion condition ($p<0.01$; *cluster based permutation test*). The red horizontal bar indicates a significant difference between ambiguous and unambiguous cylinder conditions ($p=0.01$). Time-courses were smoothed with a 30ms (s.d.) Gaussian kernel for display. All statistics were performed on the un-smoothed traces.

Next, we extended our analysis to incorporate correlations occurring at distinct timescales (i.e. different τ's; as in Fig. 4.4) while preserving some of the time-resolved trial structure. Hence, for each site, we extended the analysis of Fig. 4.5 across different time lags. Specifically, we computed rSC for all pairwise combinations of independent 100ms time-bins within the trial (t_i in SU and t_{j+x} in MU). We included sites associated with a CP and with strong and matching tuning for the ambiguous cylinder (n=39). In the resulting rSC matrix (Fig. 4.6a), the central diagonal of the matrix is a down-sampled version of the rSC time-course in Fig. 4.5 and corresponds to correlations occurring at short timescales ($\tau < 50$ms), which we term instantaneous correlations. Data on the off-diagonals correspond to correlations at longer timescales. Averaging the rSC matrix across the main diagonal corresponds to a down-sampled version of the CCGs in Fig. 4.4a (insets in Fig. 4.6a & Supp. Fig. 4.3a).

To probe whether interneuronal correlations at longer timescales could be predictive of CP measured at subsequent points during the trial (Wimmer et al., 2015), we isolated the long timescale component of the rSC from the rSC matrix. For each time-bin t_i we averaged over the four leading and four lagging rSC values (i.e. matrix entries) in respect to t_i (see black arrows in Fig. 4.6a) while excluding the on-diagonal rSC value. This average value is termed $rSC\tau_{=4}$ and represents the trial time resolved long timescale component of the rSC, corresponding to a lag of ±50-450ms (Fig. 4.6b). The average $rSC\tau_{=4}$ time-course, although noisy, is qualitatively similar to the instantaneous rSC time-course in Fig. 4.5 and shows a strong rise up until ~800ms, as well as a decay during the second half of the trial. In order to quantify these trends for each site, we fitted a 2nd order polynomial to the $rSC\tau_{=4}$; the derivatives of the polynomial provide a trial-time resolved index of whether $rSC\tau_{=4}$ is rising, decaying or constant. Next, for each recording site's SU, we computed CP time-courses in sliding 100ms window. Thus, for each time-point (ms) in the trial, we have two vectors: one vector of the derivatives indicating an $rSC\tau_{=4}$ trend for each site and a second vector containing the CP values for each site. When we computed the Pearson correlation between the derivative and CP vectors for each pairwise combinations of time-points, the resulting correlation matrix revealed that a rising $rSC\tau_{=4}$ during the first half of the trial is associated with a larger CP values during the second half of the trial (p=0.01; *cluster-based permutation test*) (Fig. 4.6c). In other words, the $rSC\tau_{=4}$ slope during the first half of the trial is predictive of CP magnitude later in the trial. This relationship is further illustrated when we split recording sites by the sign of their time averaged derivative i.e. whether their $rSC\tau_{=4}$, on average, rose or decayed throughout the trial. For each group (n=21 for sites with derivative < 0; n=18 for sites with derivative >0) we averaged the CP time-courses (Fig. 4.6d). Similar to Fig. 4.6c, Fig 4.6d shows that a rising $rSC\tau_{=4}$ is associated with higher

CP during the second half of the trial ($p<0.05$; *cluster-based permutation test*). Our data showed similar trends when analysed for each monkey separately (Supp. Fig. 4.8)

We then tested whether we could observe a similar relationship when taking into account only the instantaneous correlation i.e. the short timescale component of the rSC ($\tau <$ 50ms; denoted $rSC\tau_{=0}$), there was no clear relationship between $rSC\tau_{=0}$ and CP (Supp. Fig. 4.9). More specifically, the derivative of the polynomial fitted to $rSC\tau_{=0}$ did not predict CP at any time-point in the trial. Together, those results show that CP for the ambiguous cylinder is best predicted by the dynamics of correlations at long but not short timescales, which in turn might reflect a feedback signal onto V5/MT.

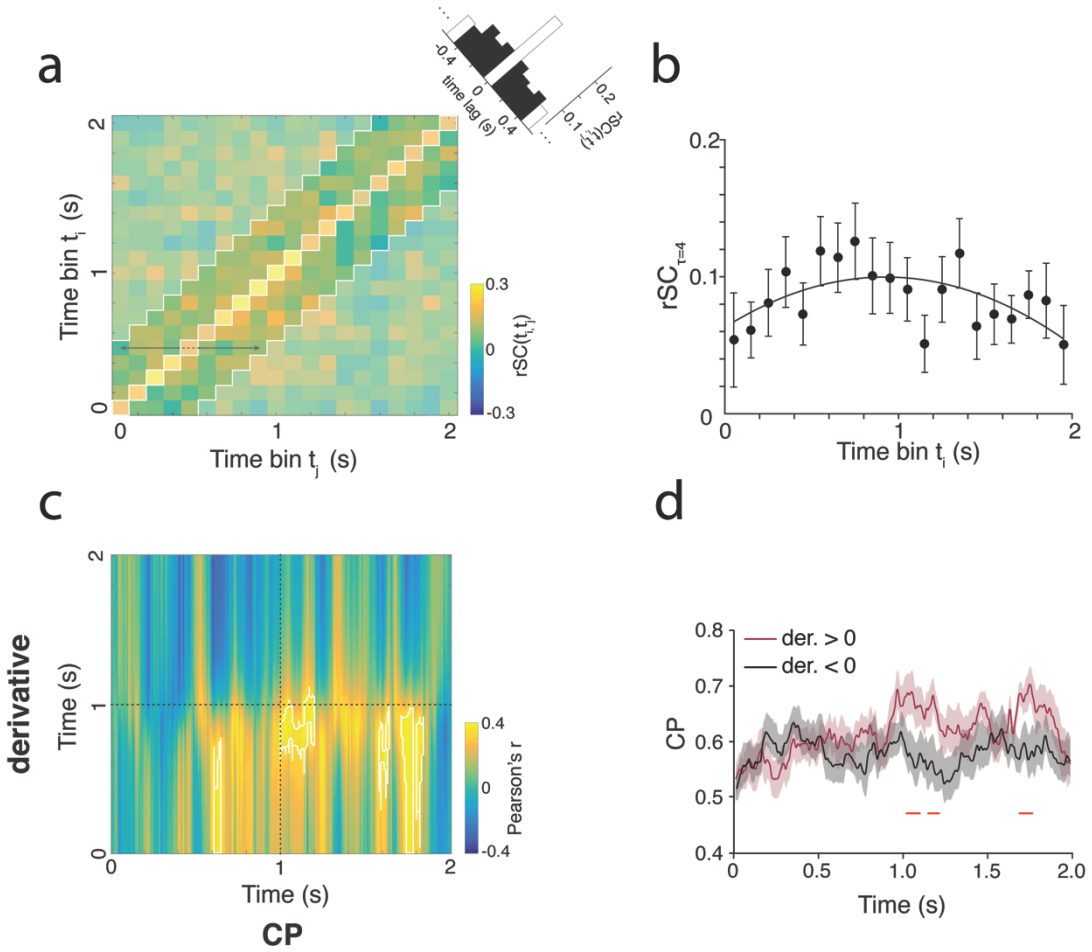


Figure 4.6: Correlation time-course predicts choice probability. a) The spike count correlation matrix for the ambiguous cylinder: The matrix entries rSC(t_i, t_j) are the rSC values computed from time-bins t_i in the SU and t_j in the MU (100ms independent time-windows) and averaged over all sites. The color bar indicates the magnitude of the rSC values. The more saturated colours indicate off-diagonal i.e. the four lagging and leading time-bin pairs with respect to t_i (excluding the main diagonal; see black arrows), that were used to compute rSC$\tau_{=4}$. Inset shows histogram of the rSC values averaged along the main diagonal. Black bars show the diagonals that were used to compute rSC$\tau_{=4}$ **b)** rSC$\tau_{=4}$ averaged over all sites are plotted as a function of time-bins t_i. Error bars show s.e.m. The black line indicates a 2nd order polynomial fit to the rSC$\tau_{=4}$. **c)** The matrix for Pearson correlation coefficients was computed between the derivative of the fitted polynomial in b) and CP across recording sites. Matrix entries simply represent the correlation coefficients obtained for each possible time-point pairing. White contours mark clusters with a significant Pearson correlation (p=0.01; *cluster based permutation test*). The color bar indicates the magnitude of the correlation. Dashed black lines mark 1s (i.e. half of the trial). **d)** The CP time-courses averaged across sites with a negative derivative (black; n=21) and positive derivative (purple; n=18). The red horizontal lines mark significant differences between the two CP time-courses (p<0.05; *cluster based permutation test*). Shaded area indicates s.e.m.

4.5 Discussion

Our data show a significant difference between the interneuronal correlations evoked by distinct visual stimuli and tasks: Interneuronal correlations measured while animals made perceptual choices about the rotation direction of an ambiguous cylinder were higher (rSC=0.42) than when the cylinder was rendered unambiguous (rSC=0.28) or when monkeys maintained fixation while viewing a classic random dot stimulus (rSC=0.23). Interestingly, the higher correlations for the ambiguous cylinder arose on longer timescales (100-400ms), indicative of a top-down influence on sensory processing. Furthermore, the long timescale – but not the instantaneous - component of the interneuronal correlations was predictive of CP time-course associated with the ambiguous cylinder. These data present a link between a top-down source injecting interneuronal correlations into sensory areas of visual cortex and the choice signals measured therein.

The reported magnitude of measured spike count correlations in visual cortex is usually small and positive but can vary across brain region, stimulus or task instruction (Cohen and Kohn, 2011). In V5/MT, reported values for rSC range from 0.1 to 0.2 for neural responses to random dot motion stimuli (Bair et al., 2001; Cohen and Newsome, 2008; Huang and Lisberger, 2009; Solomon et al., 2015; Zohary et al., 1994). We compared the isolated single unit to the surrounding multi-unit activity, rather than comparing pairs of isolated SUs with the same electrode. Nonetheless, our measured rSC for the random motion condition (0.17) was very similar to the one measured by Bair et al. (2001) (Bair et al., 2001) (0.19) and Zohary et al. (1994) (Zohary et al., 1994). In the present study, as well as the above mentioned studies, single cells were carefully selected for their

responsiveness and tuning profiles to the main visual stimulus as well as tuning similarity, properties which positively correlate with rSC, generally (Cohen and Kohn, 2011), and specifically in V5/MT (Bair et al., 2001; Cohen and Newsome, 2008; Huang and Lisberger, 2009; Solomon et al., 2015). Poor spike sorting can lead to inflated rSC values (Ecker et al., 2010; Kohn and Smith, 2005), as can pooling of cells (Rosenbaum et al., 2011). However, our measure of rSC for random motion also matches the rSC measured by Cohen and Newsome (2008), using the same stimulus, but recording isolated cell pairs with separate electrodes (rSC ≈ 0.2 for pairs with similar tuning).

Our data revealed a comparatively large value of rSC (0.37) in response to the ambiguous cylinder. In contrast to the random motion task, which purely required fixation, the cylinder task required a perceptual decision. While we could not directly compare decision-making conditions for both stimuli/tasks, our observations present evidence that it is stimulus-related properties rather than task instructions that determine rSC. For one, other data show that rSC measured for random motion does not differ markedly across task difficulties, i.e. the strength of motion coherence (Bair et al., 2001) and passive fixation (Zohary et al., 1994). Furthermore, we found that the large rSC associated with the ambiguous version of the cylinder stimulus diminished when the cylinder stimulus was rendered unambiguous, approaching the value for random motion. Thus, the perceptual bistability of the cylinder stimulus and not the decision-making component seems fundamental to the generation of high rSCs.

The cylinder and random motion stimuli differ markedly in their perceptual interpretation. The random motion stimulus contains dots, which probabilistically move in all possible directions, resulting in zero net motion. This type of stimulus does not

evoke a stable percept. In contrast, the ambiguous cylinder stimulus evokes two different percepts, each of which is stable: perceptual reports of the rotation direction indicate that there can be spontaneous switches between each of these two stable representations at a rate of tens of seconds (Krug et al., 2008; Nawrot and Blake, 1993). Crucially, the dots forming the cylinder are consumed into a single perceptual representation in a 'winner-take-all' fashion, although the sensory input remains indeterminate in nature.

Human fMRI studies in bistable perception have suggested a fronto-parietal top-down component stabilizing/destabilizing the percept generated in sensory areas through a feedback mechanism (Brascamp et al., 2018; Lumer et al., 1998; Weilnhammer et al., 2013), which might affect the circuits in V5/MT. Some evidence supporting this interpretation in our task comes from the larger reaction times (RTs) that are displayed by monkeys when deciding on the rotation direction of the ambiguous cylinder as opposed to the unambiguous version. This could be interpreted as showing that the signals stabilizing the percept become more dominant when the sensory input is weak or absent (Cicmil et al., personal communication). Alternatively, transitions in bistable vision could be a result of dynamic adaptation and inhibition within visual cortex. These processes do not necessarily rely on a feedback modulation but could arise in a feed-forward fashion (Brascamp et al., 2015) or through attractor mechanisms (Cao et al., 2016; Cicmil et al., 2015), which might also influence interneuronal correlations (Gutnisky and Dragoi, 2008; Middleton et al., 2012).

Correlations in the activity of sensory neurons arise over a variety of timescales (Bair et al., 2001; Cohen and Kohn, 2011; Kohn et al., 2016). We show that much of the correlation structure that is specific to the ambiguous cylinder arises on relatively long

timescales (tens to hundreds of milliseconds). In contrast, correlations associated with the random motion stimulus saturated at a time-lag of about 30-40ms, which matches a previous study by Bair et al. (2001). This was also the case for unambiguous versions of the cylinder stimulus. Experimental and theoretical studies have shown that feedback signals onto sensory cortex that represent cognitive factors such as attention, learning or internal bias, affect the structure of noise correlations therein (Bondy et al., 2018; Chicharro et al., 2017; Cohen and Maunsell, 2009; Cohen and Newsome, 2008; Ecker et al., 2016; Gu et al., 2011; Haefner et al., 2013; Mitchell et al., 2009; Ni et al., 2018; Wimmer et al., 2015). The timescale of fluctuations of such top-down cognitive factors is estimated to be comparably long. For example, fluctuations in attention range from hundreds of milliseconds to seconds (Cohen and Maunsell, 2011, 2010), but not less than 100-200ms (Nienborg et al., 2012).

To determine the signature of different sources of correlations, Wimmer et al. (2015) explored a hierarchical network model designed and trained to solve a direction discrimination task with random motion stimuli. Their model consisted of a sensory layer (representing V5/MT) which was recurrently connected to a higher order integration layer (representing LIP or FEF). It made it possible to disambiguate the correlation signatures in the sensory layer that were injected by the feedback signal (termed top-down correlations) versus those accounted for by the local fluctuation in sensory layer and the stimulus (termed bottom-up correlations). Their modelling found that bottom-up correlations depended on interactions at short time-lags and disappeared at longer time-lags (>125ms). Conversely, top-down correlations were weak for short time-lags but stronger for longer time-lags (>125ms), especially during the second half of the trial. Interestingly, they were overall negligible at the beginning of the trial but

rose towards the end. The structure of noise correlations observed in that model matched those of real V5/MT neurons from monkeys performing the motion discrimination task (Wimmer et al., 2015). In short, lagged correlations rose throughout the trial while instantaneous correlations were flat.

In our data, however, the correlation time course for the ambiguous cylinder showed a prominent bump in the middle of the trial. Like in Wimmer et al. (2015), our correlation time-course at longer time-lags showed a rising phase - but then peaked in the middle - unlike in Wimmer et al. (2015) where the lagged correlation time-course continued to rise until the end of the trial. Importantly, a great deal of the interpretation offered by that model is valid only under the assumption that the animal takes the decision about how to respond at the end of the trial. Some evidence has been put forward that this might be the case using different paradigms (Kiani et al., 2008). Furthermore, the random motion stimuli used in the model are noisy, do not evoke a consistent percept of motion direction and require prolonged integration of evidence. In contrast, the ambiguous version of the cylinder stimulus might evoke a clear 'winner take all' percept earlier in the trial.

If there were a feedback signal biasing perception towards one of the two possible configurations, this could act relatively early in the trial and might explain the observed bump in the rSC timecourse for the ambiguous cylinder stimulus. Intriguingly, this bump is not present in the instantaneous rSC timecourse for the unambiguous cylinder or random motion stimuli, which seem more similar to the flat instantaneous time-courses observed in Wimmer et al. (2015). A similar flat time course was also observed in a more recent study by Bondy et al. (2018), who used unambiguous and noisy stimuli –

unlike the ambiguous cylinder. Additionally, the observed rise and decay in correlations causing the distinct bump might mark the engagement and disengagement of attentional processes quenching the correlations in V5/MT (Cohen and Maunsell, 2009; Ecker et al., 2016; Ruff and Cohen, 2014b), which could act from the beginning of the trial (Hupé et al., 2001).

More generally, a top-down component dynamically changing correlations between relevant neuronal pools (Bondy et al., 2018; Cohen and Newsome, 2008) is also in accordance with pooling models for the cylinder discrimination task (Krug, 2004; Krug et al., 2016; Parker, 2013), that suggest that neuronal pools must be structured flexibly and independently for successful discrimination of stimuli. We note that feedback onto V5/MT need not be of a general 'cognitive nature'. The cylinder stimulus requires the specific conjoint read-out of motion and disparity cues (Krug et al., 2013). The medial superior temporal area (area MST), which lies upstream of V5/MT, shows disparity-dependent direction selectivity potentially related to the perception of rotation (Roy et al., 1992). Area MST might send feedback signals related to the perception of rotation to V5/MT and increasing the correlation between groups of neurons signalling the same rotation.

Here, our experimental results show a positive correlation between the rSC of a single unit with its surrounding multi-unit activity and that single unit's CP. Previously, theoretical models have proposed a relationship between the rSC of a given neuron and its CP (Shadlen et al., 1996). Haefner et al. (2013) derived analytically that a neuron's CP should be determined by the mean correlation between that neuron and all other neurons, weighted according to a pooling weight for each neuron. The direct correlation

we find between CP and rSC might be related to our estimation of rSC, the selection criteria and sampling of our recorded neurons: Our calculation of rSC is not pair-wise but for a given neuron with a number of its direct neighbours, probably situated in a column with similar tuning properties. In general, SU and MU showed very similar tuning (we only found 4 neuronal sites showing a clear tuning mismatch). Also, our selection criteria required a minimum for stimulus selectivity for both single and multiunit. Thus, our sampling/selection would prefer similarly tuned, spatially close neurons with presumably 'higher weights' because of their task relevance. However, the columnar architecture in V5/MT means that correlations between neural signals that are spatially very close to each other are a theoretically important part of the population's overall correlation structure. Nevertheless, a limitation of the current study is the small sample size and recording technique which prevent us from estimating the full correlation matrix, necessary to validate the exact predictions made by Haefner et al. (2013).

A recent study recording from multi-electrode arrays in V1 in the macaque monkey did provide this missing link and predicted CP accordingly (Bondy et al., 2018). Interestingly, they also found that the largest variability in the rSC structure seemed to be determined by a task-dependent feedback component, instead of a purely feed-forward component considered in Haefner et al. (2013). This observation and the direct rSC/CP correlation in our data may have a similar explanation. The correlation could be driven by a dynamic feedback signal commonly affecting the measured rSC and CP, which could also explain our observed positive correlation between rSC and CP after taking away the effect of stimulus tuning (Cicmil et al., 2015). Supporting this notion, we also find some evidence that CP is better predicted by the long timescale component

of rSC. More specifically, a rising lagged component of the rSC during the first half of the trial predicted larger CPs during the second half of the trial. To some extent, this is in accordance with previous results for the random motion discrimination by Wimmer et al. (2015), mentioned previously, where rising lagged correlations (possibly representing top-down feedback) correlated with larger CPs towards the end of the trial (Wimmer et al., 2015). In our, case lagged correlations decayed towards the second half of the trial while CPs remained constant. This might indicate differences in decision or attentional timescales as mentioned before. In short, a decision about the ambiguous cylinder might be conveyed during the first part of the trial, without prolonged need for a top-down feedback towards the end of the trial. Of course, we cannot fully determine the source of the long timescale correlations without simultaneous recordings in other brain areas. But our observations seem in accordance with a rising body of experimental (Bondy et al., 2018; Kwon et al., 2016; Nienborg and Cumming, 2009) and theoretical studies (Chicharro et al., 2017; Haefner et al., 2016, 2014) suggesting a top-down component conveying decision-related signals in sensory areas and specifically V5/MT (Cicmil et al., 2016, 2015; Krug, 2004; Maunsell, 2015; Nienborg et al., 2012).

Lastly, a recent study re-evaluating common measures of CP during 2AFC tasks in ventral parietal area (VIP) found that CP magnitude can be overestimated when computing stimulus tuning curves from trials collected during the actual decision-making experiment (Zaidel et al., 2017). This is because part of CP is decoupled from sensory input and rather reflects a purely choice dependent top-down signal that distorts the 'true' sensory tuning curves, which are then used to compute a possibly inflated CP. However, the tuning to cylinder disparity in visual area V5/MT is independent of task performance and the magnitude of CP for the cylinder stimulus is correlated with

neurometric threshold (Dodd et al., 2001). Together with micro-stimulation experiments (Krug et al., 2013), it is clear that V5/MT neurons contribute directly to perceptual decisions about SFM cylinders.

In summary, we show that bistable SFM cylinders associated with significant choice-related activity (CP), evoke highly correlated neuronal firing (rSC) between nearby neurons in visual area V5/MT. Much of the interneuronal correlation setting apart the bistable cylinder percept from unambiguous versions as well as classic random motion stimulus arises on relatively long timescales. It is these correlations on longer timescales that predict the particularly large choice-related component of neuronal activity observed for the bistable cylinder. Here, we argue that the relationship between interneuronal correlations and choice-related firing might be the neural signature of a feedback signal acting upon V5/MT circuits, stabilizing the perceptual appearance of the SFM cylinder. Bistable, ambiguous stimuli like the SFM provide an interesting probe for models of perceptual decision-making, in particular to investigate differential contributions of bottom-up and top-down influences during perceptual decisions.

Chapter 5. General Discussion

In the previous chapters I have presented the results from three experimental studies. Together, these results contribute towards the characterization and understanding of the dynamical neural processes underlying information representation during higher cognitive function. Specifically, I have shown that neural dynamics in PFC are high dimensional, allowing for different formats of information to be processed in parallel. The information processing capacities of cells in PFC can be directly predicted by the temporal statistics of their baseline firing, which distinguishes between dynamic or stable processing of working memory content. Finally, the temporal dynamics of correlations between neurons are directly related to and influence decision signals about ambiguous stimuli that might require top-down control. In the following, I will summarize the main findings and contributions of each experimental chapter. I then discuss some limitations and outstanding questions regarding the individual chapters that had not been addressed in the dedicated discussions. Finally, I will suggest future lines of research that build on the results presented in this thesis.

5.1 Chapter summaries & contributions

In **chapter 2** we aimed to characterize the neural dynamics underlying a highly demanding cognitive process, that of performing several tasks simultaneously and/or in close temporal succession. Such dual- or multi-task performance taxes working memory systems because several pieces of information have to be kept in mind for different use and without interference. Furthermore, it involves attention and selection mechanisms as well as other factors typically ascribed to the general concept of cognitive control.

Our main question was how representations in neural populations avoid interference during dual tasking. We found that monkeys learned to perform a complex dual task with almost perfect accuracy. That is, performance on dual task trials matched performance on each of the component tasks done separately (usually > 90% accuracy). We then examined neural responses in lPFC, a brain region that has been suggested to play a key role during dual- and multi-tasking, and found that prefrontal neurons represented feature related information throughout the various stages of the task including stimulus encoding, maintenance and response periods. Importantly, many neurons multiplexed feature information from both tasks, displaying linear and non-linear mixed selectivity, and thus indicating a potential for interference. Population level analysis revealed that despite mixed selectivity, feature information from component tasks transitioned into distinct (i.e. orthogonal) low dimensional subspaces of the high dimensional population activity. This observation was consistent with the monkeys' behaviour. Furthermore, information about the dual task features was present in a format that allowed for generalization of one task across the conditions of the other task. However, the ability to cross-generalize across conditions of the respective other task was compromised by the presence of non-linear mixed selectivity.

Our results make various contributions to the characterization and interpretation of prefrontal population responses during higher cognitive function. As I have argued in **chapter 1**, many cognitive processes are better understood by considering the distributed nature of neural codes. Here, we have provided evidence that interference between various streams of information within the same network can be resolved by restricting computations to orthogonal subspaces on the level of the population. Of course, this requires a high dimensional representation, which is typically associated

with non-linear mixed selectivity (Rigotti et al., 2013). Indeed, orthogonal task subspaces can exist in a population purely consisting of non-linear mixed selectivity. However, during the dual task, non-linear mixed selectivity could be problematic because it prevents generalization, another coding mechanism that has been observed in lPFC (Bernardi et al., 2019). In other words, a high dimensionality comes at the cost of generalizability, at least in some scenarios (see Bernardi et al. 2019 Supplementary Material for a more in-depth discussion, and Barak et al., 2013). Borrowing from the concept of adaptive coding (Duncan, 2001), it might be that lPFC adapts the dimensionality of neural representations according to current task demands i.e. whether behaviour needs to be generalized or differentiated. Our results speak to this hypothesis as the neural representation of the dual task seems to strike a compromise: Orthogonal task representations together with generalizable component task representations create a format that allows for stable task-specific readout.

As described in **chapter 1**, neuronal dynamics in PFC are heterogeneous, complicating interpretation of their functional roles during cognitive tasks. This is often reconciled by considering single neuron responses as snapshots of an underlying, distributed code. However, even a distributed code relies on the computations and dynamics that can be carried out by its neurons, whether those are inherent or the result of network architecture. For example, a population entirely comprised of persistently firing neurons with pure selectivity is bound to exhibit relatively low dimensional task representations. Thus, certain properties of single neuron responses constrain and can predict the representational and computational capacity displayed by a circuit. The inherent temporal variability in a neuron's firing rate is such a property. In fact, a striking phenomenon of the functional organization of cortex is the hierarchy of intrinsic

temporal receptive fields (i.e. intrinsic timescales) (Murray et al., 2014) with neurons in PFC expressing the longest timescales, possibly reflecting extended integration over time. In **chapter 3** we investigated whether the intrinsic timescales of neurons in various cortical brain areas predict complex population coding dynamics during WM. Indeed, we found that in three tested brain areas (lPFC, FEF and LIP) cells with long intrinsic timescales carried information more stably during WM. In lPFC, however, we observed an interesting dissociation. Short timescale cells encoded information earlier and preferentially during stimulus encoding, whereas long timescale cells became active later in the trial but robustly maintained their selectivities across the entire delay period. Furthermore, short timescale cells in lPFC specifically signalled transitions between task states (e.g. encoding to delay) and carried feature information in their dynamic firing pattern. Although the long timescale cells generally carried more feature related information per time point, their lower dimensional temporal dynamics carried less information. We also found a link between functional connectivity at rest and intrinsic timescale in FEF and LIP, but not lPFC pointing to potentially different underlying network mechanisms.

In summary, our results in **chapter 3** confirm that the hierarchical distribution of baseline timescales throughout cortex indeed relates to the cognitive processes those regions are involved in. Furthermore, the results from **chapter 3** contribute to the current debate about WM coding mechanisms (Constantinidis et al., 2018; Lundqvist et al., 2018). They are particularly in line with recent studies that report interactions between stable and dynamic coding schemes during WM (Murray et al., 2017; Chaisangmongkon et al., 2017), which have also been discussed in **chapter 1**. Our findings show that the diversity of intrinsic timescales among prefrontal neurons enables such division of

labour. Depending on the task at hand, prefrontal cells might be dynamically recruited based on their intrinsic timescale to ensure maintenance of some variables but concurrent dynamic processing of others (or the same).

In C**hapter 3** we focused on intrinsic timescales measured as temporal autocorrelation, and how these predicted task related processing over time. At the end of the chapter we attempted to relate temporal autocorrelation to activity covariance between neurons i.e. interneuronal correlations. While otherwise conceptually unrelated, **chapter 4** investigated how these interneuronal correlations - and the timescales on which they evolve - relate to decision signals in sensory cortex (area V5/MT). Specifically, we found that an ambiguous 3D moving cylinder stimulus evoked relatively high interneuronal correlations, when compared to unambiguous versions of the same stimulus or 2D random motion stimuli. Interneuronal correlations also correlated with choice signals (measured as CP) for the ambiguous cylinder stimulus. We then applied a method to dissociate the timescales on which interneuronal correlations arise. We found that the magnitude of short timescale correlations was similar for all tested stimuli. However, long timescale correlations were highest for the ambiguous cylinder stimulus. Interestingly, the timecourse of the long timescale (but not the instantaneous component of the interneuronal correlations) predicted choice signals later on in the trial.

The results in **chapter 4** contribute to the current research on the formation and nature of choice related activity in sensory cortex. Specifically, they link choice signals represented in individual neurons to timescales of correlations between neurons, a relationship which had been suggested theoretically (Shadlen et al., 1996; Haefner et al., 2013; Wimmer et al., 2014) and which is increasingly being explored experimentally

(Zaidel et al., 2017; Bondy et a., 2018). Because the cylinder stimulus is ambiguous, it might recruit a top-down component to stabilize the percept, whose influence on V5/MT could be reflected in the form of long timescale correlations between neurons (Wimmer et al., 2015). Given the focus of this thesis it is tempting to argue that this might involve PFC. However, the nature of our data and structure of the task did not enable us to explicitly distinguish the origin of the observed interneuronal correlations (see **chapter 4** discussion), and to implicate PFC in the process would be mere speculation, but see a recent modelling study (Engel et al., 2015).

5.2 Limitations and outstanding questions

The data presented in **Chapter 2** requires several additional analyses. For example, the fact that we did not find an obvious reduction in accuracy scores between single and dual task conditions does not exclude a subtler form of dual task interference. It could be that specific combinations of attention and WM locations e.g. trials with increased or decreased proximity between locations cause different error rates. Furthermore, differences in delay length typically influence performance (Watanabe and Funashashi, 2014), which could be more pronounced during dual tasking (Watanabe and Funshashi, 2014). Thus, a more sophisticated analysis might reveal subtle interference between component tasks on the behavioural level. Another behavioural measure of task performance are RTs, which in monkey T were clearly longer for lever releases during dual tasking (Supp. Fig. 2.1). A more detailed analysis of RTs as described above for accuracy, or even analysing saccade trajectories might provide a more consistent signature of dual task interference. Focusing on RT as behavioural measure for the dual task would also allow for testing the behavioural relevance of the observed neural

representations and dynamics. Specifically, a central question is whether the observed geometry of the neural representation (see scheme in Fig. 2.8b) in PFC is indeed important for - or related to - performing the dual task fast and accurately. Would overlap in task subspaces, and/or an increase in non-linear mixed selectivity, lead to greater dual task interference on the behavioural level? While there are to few error trials to test those predictions in terms of accuracy, we could test whether our metrics of subspace overlap and cross-instance generalization vary with RT e.g. increased subspace overlap results in slower RTs or vice versa. Another possibility is that RT does not vary with any of those metrics but instead correlates with general decoding performance. For example, stronger attention task decoding leads to faster lever release RTs and so on. It might be that after learning the task over many months or even years, representations and trajectories in PFC are relatively fixed and behaviour simply depends on whether the monkey has seen or paid attention to the stimuli. In this case, neural coding along the axes that provoke subspaces overlap or failure to generalize should not vary with RT and one would need to look at how the representation changes over the course of learning in order to determine whether they are relevant to behaviour (which will be further discussed in the *Future perspectives* in this chapter).

A particularly consistent observation across **chapters 2 & 3** was that activity patterns changed between initial encoding and delay epochs, which seems to be common during task processing in PFC (Stokes et al., 2013; Spaak et al., 2017; Parthsarathy et al., 2017) but also other higher level brain areas (Elsayed et al., 2016). Our results from **chapter 3** suggest that the initial encoding phase might be more task general i.e. dominated by sensory coding as the attention and WM task subspaces overlapped. While we can not completely rule out task general coding during the delay period (see **chapter 3**

Discussion) our data suggest task specific delay coding because attention and WM delay subspaces did not overlap. Cognitive processing in PFC in general and during WM in particular entails more than passively storing task content (see **chapter 1**). Instead, information is very often transformed in order suit current task demands (Rainer et al., 1999). As we have seen, this can include different codes for the same task features, depending on future use (Warden and Miller, 2010; Mante et al., 2013; Stokes et al., 2013; see Nobre and Stokes, 2019 for review). Our results align well with the idea of such use-dependent processing in PFC (saccade for WM task versus lever release for attention task), which is also supported by recent theoretical results. Orhan and Ma (2019) trained recurrent neural networks on a range of WM tasks and found that the same features of visual stimuli (e.g. orientation) were stored in an entirely different format depending on what had to be done with them at the end of the delay (Orhan and Ma, 2019). During the dual task, use-dependent processing could be beneficial for several reasons. As observed, different task formats can be orthogonal and therefore resolve interference at the neural level. On the other hand, the distinct patterns of network activities for different task but also across epochs may interact with incoming stimuli differently (i.e. attention target or memory cue/target). This state dependent interaction might allow for computations that generate fast and context dependent behaviour (Buonomano and Maass, 2009). Another way to probe the use-dependent format of the WM and attention tasks could be to compare them to the neural activity during performance of the single tasks. Are subspaces between the WM and attention tasks inherently orthogonal because of their use-dependence or are they only orthogonal if both tasks are being performed concurrently?

One of the major limitations in **chapters 3 & 4** was our inability to expand the correlational analyses to neuronal pairs recorded at different contacts. In **chapter 4** this would have allowed us to characterize interneuronal correlations (i.e. noise correlations) between all kinds of differently tuned neurons, which allows for the construction of full signal and noise covariance matrices (see Discussion in **chapter 4**). That is, an accurate description of how the variability in the activity of a neuron tuned to e.g. rightward motion co-varies with that of a neuron tuned to e.g. far disparities in response to the cylinder stimulus (for all pairings of tunings). Such matrices are needed to fully quantify the effect of noise correlations on the processing of the cylinder stimulus and also to accurately test the very specific predictions of various models explaining bottom-up and top-down component of CP in sensory areas (Haefner et al., 2013). In **chapter 3,** more closely spaced electrodes would have allowed us to investigate the local circuit mechanisms and their relation to intrinsic timescales in much greater detail. We are currently expanding this line of research by analysing a dataset consisting of linear-array (U-probes) recordings in multiple areas of macaque cortex during a range of WM tasks. The dataset was collected by our collaborators André M. Bastos and Earl K. Miller (see Bastos et al., 2018, which includes a part of the to be analysed dataset). Recorded areas include various sub-regions of PFC including lPFC as well as visual and parietal areas. Furthermore, laminar recordings were preferentially made perpendicular to the cortical surface and thus simultaneously capture neurons across cortical layers. This dataset allows us to test several hypothesis regarding the emergence of long timescale cells in PFC and other brain regions. First, we can more explicitly test whether and how intrinsic timescales relate to local functional connectivity (this includes applying the Methods developed in **Chapter 4**). Preliminary results suggest that neurons in superficial and intermediate layers in PFC show particularly stable delay firing as well as increased

functional connectivity (i.e. higher noise correlations). Long timescale cells seem to preferentially cluster within intermediate layers and show high interneuronal correlations with neurons in deep as well as superficial layers. Other than local connectivity properties, long intrinsic timescales can also arise as a result of long range inter-areal connections (Chaudhuri et al., 2015; Mejias and Wang, 2019). This prediction can also be tested in our dataset by analysing local field potentials (LFPs) and using Granger causality analysis to determine functional connectivity between brain areas. This could then be compared between contacts that record from long or short timescale cells. Finally, although we refer to intrinsic time scales, it is possible that this parameter can vary under specific circumstances. In the data presented in **chapter 3**, a cell's 'intrinsic' timescale was relatively stable across the duration of a recording session (analysis not shown). However, it is not known whether this is also true across different tasks that might require a shift in coding strategy i.e. longer delay, different stimuli, distinct goals and so on. This refers to the more general question of tuning stability in PFC. We know that prefrontal cells flexibly switch their coding preferences and become active at different time points within a trial (see **Chapter 1**); however, does a cell that fundamentally exhibits stable firing, or a cell that preferentially fires to spatial cues in one context show the same profile across many tasks and situations? (see e.g. Kollmorgen et al. 2019). Recording from the same prefrontal population during different tasks might help in disambiguating some of these questions.

Lastly, a shortcoming of our analysis in **chapter 4** is the lack of a thorough quantification of microsaccadic eye movements and their relationship to the measured interneuronal correlations. Microsaccades have been shown to modulate the activity of MT neurons (Bair and O'Keefe, 1998; Herrington et al., 2009) and can account for a

small part of the measured CPs through suppression of neural activity (Herrington et al., 1998). In a previous publication based on the same data presented here, Dodd et al. (2001) ruled out a relationship between CPs for the ambiguous cylinder stimulus and microsaccades (as well as vergence eye movements). However, this does not mean that microsaccades do not differ between - and/or do not impact interneuronal correlations differently - for distinct stimuli and tasks. More Specifically, microsaccades have been linked to neural synchrony of cells (Bosman et al., 2009; Ito et al., 2013; Bellet et al., 2017; Lowet et al., 2016) and stimulus specific firing rate modulations during covert attention (Lowet et al., 2018). It therefore remains to be tested whether the high interneuronal correlations for the ambiguous cylinder stimulus might be caused by microsaccade related activity.

5.3 Future directions

I would like to use this section of the thesis to briefly outline an ongoing project that has built on some of the major topics outlined in **chapter 1** but also relates to the work presented in **chapter 2**. The main question of this project was how task specific representations are formed in PFC during the course of learning a complex cognitive task. Typically, neural activity is recorded after animals have learned to perform a given task to perfection. However, it is largely unclear how PFC represents 'unlearned' information and how neuronal responses and distributed representations change to enable goal directed behaviour.

A striking property of higher level association cortices, including PFC, is their recurrent local connectivity. Taken together with the fact that PFC receives inputs from most

cortical and subcortical structures (see **Chapter 1**) it is possible that pre-learning representations in PFC are high dimensional, akin to those in randomly connected reservoir networks (Enel et al., 2016). This leads to the empirical prediction that mixed selectivity and dynamic coding should be observed in PFC even before any task is learned. During learning, a high dimensional representation might be beneficial because behavior has to be continually adjusted (Tang et a., 2019) and new input-output mappings have to be found and tested. However, during the course of learning dimensionality might collapse to enable more robust generalization (Mack et al., 2019). That is, learning might be important in biasing selectivity profiles towards task relevant variables and their interactions (Lindsay et al., 2017). On the contrary, selectivity to task related features and contingencies as well as mixed selectivity and abstract responses could initially be absent from PFC, and only gradually emerge as the task is learned.

To test these possibilities, we designed a complex cognitive task (Supp. Fig. 5.1), containing an XOR problem (as described in **Chapter 1**). Additionally, certain features of objects in the task are irrelevant, and have to be generalized over. Together with our collaborator Kei Watanabe, we have been training monkeys on this task while simultaneously recording neural responses in lPFC using linear-arrays (U-probes). However, learning has not yet been completed and we have only started neural data analysis. Nonetheless, this project has formed a major part of my doctoral work and directly builds on the ideas I have set up during **chapter 1**.

5.4 Conclusion

In summary, this thesis has investigated single neuron responses as well as population

level representations underlying processes relating to complex, cognitive behavior. The format and dynamics of distributed neural representations can display emergent properties that allow for a better understanding of how the brain implements those behaviours. However, future work could focus on how higher level cortices acquire such representations in the first place, as most of the important things might happen when something new has to be learned.

References

Abbott, L. F., Rajan, K., Sompolinsky, H. Interactions between intrinsic and stimulus-evoked activity in recurrent neural networks. In: *Neuronal variability and its functional significance* (Oxford University Press, Oxford, 2010)

Aertsen, A. M., Gerstein, G. L., Habib, M. K. & Palm, G. Dynamics of neuronal firing correlation: modulation of 'effective connectivity'. *J. Neurophysiol.* **61**, 900–917 (1989).

Alemi, A. & Abbara, A. Exponential Capacity in an Autoencoder Neural Network with a Hidden Layer. *arXiv* 1705.07441v1 (2017).

Albert, D. What's on the mind of a jellyfish? A review of the biobehavioural observations on jellyfish. *Neurosci. and Biobehav. Reviews* **35**, 474-482 (2011).

Alexander, G. E., Delong, M. R., Strick, P. L. Parallel organization of functionally segregated
circuits linking basal ganglia and cortex. *Annu. Rev. Neurosci*. **9**, 357–81 (1986).

Alvarez, G.A., Cavanagh, P. Independent resources for attentional tracking in the left and right visual hemifields. *Psychol. Sci.* **16**, 637–643 (2005).

Amaral, D. G., Price, J. L. Amygdalo-cortical projections in the monkey (*Macaca fascicularis*). *J. Comp. Neurol.* **230**, 465–96 (1984).

Amit, D. J., and Brunel N. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cerebr. Cortex* **7**, 237–52 (1997).

Anderson, P. W. More is different. *Science* **177**, 393-396 (1972).

Asaad, W. F., Rainer, G. & Miller, E. K. Neural activity in the primate prefrontal cortex during associative learning. Neuron **21**, 1399–1407 (1998).

Baddeley, A., Logie, R., Bressi, S., Sala Della, S., Spinnler, H. Dementia and working memory. *Q J Exp Psychol A* **38**, 603–618 (1986).

Baddeley, A. Working memory. Science **255**, 556–559 (1992).

Baddeley, A., Della Sala, S., Papagno, C., Spinnler, H. Dual-task performance in dysexecutive and nondysexecutive patients with a frontal lesion. Neuropsychology **11**, 187–194 (1997).

Baeg, E. H. et al. Dynamics of population code for working memory in the prefrontal cortex. *Neuron* **40**, 177–188 (2003).

Bair, W., O'Keefe, L. P. The influence of fixational eye movements on the response of neurons in area MT of the macaque. *Vis. Neurosci.* **15**, 779–786 (1998).

Bair, W., Zohary, E. & Newsome, W. T. Correlated Firing in Macaque Visual Area MT: Time Scales and Relationship to Behavior. *J. Neurosci.* **21**, 1676–1697 (2001).

Balaguer-Ballester, E., Lapish, C. C., Seamans, J. K., and Durstewitz, D. At- tracting dynamics of frontal cortex ensembles during memory-guided decision-making. *PLoS computational biology* **7**, e1002057 (2011).

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

Barak, O., Sussillo, D., Romo, R., Tsodyks, M., Abbott, L. F. From fixed points to chaos: Three models of delayed discrimination. *Prog. Neurobiol.* **103**, 214–22 (2013).

Barak, O., Tsodyks, M. & Romo, R. *Neuronal population coding of parametric working memory*. *J. Neurosci.* **30**, 9424–9430 (2010).

Barak, O., Rigotti, M. & Fusi, S. *The sparseness of mixed selectivity neurons controls the generalization–discrimination trade-off*. *J. Neurosci.* **33**, 3844–3856 (2013)

Barbas, H., De Olmos J. Projections from the amygdala to basoventral and mediodorsal prefrontal regions in the rhesus monkey. *J. Comp. Neurol.* **300**, 549–71(1990).

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

Barlow, H. B. Single Units and Sensation: A Neuron Doctrine for Perceptual Psychology? *Perception* **1**, 371–394 (1972).

Bastos, A. M., Loonis, R., Kornblith, S., Lundqvist, M. & Miller, E. K. Laminar recordings in frontal cortex suggest distinct layers for maintenance and control of working memory. *PNAS* **115**, 1117–1122 (2018).

Bellet, J., Chen, C.-Y., and Hafed, Z.M. Sequential hemifield gating of alpha and beta behavioral performance oscillations after microsaccades. *J. Neurophysiol*. **118**, 2789–2805 (2017).

Bernacchia, A., Seo, H., Lee, D., Wang, X. J. A Reservoir of time constants for memory traces in cortical neurons. *Nat. Neurosci.* **14**, 366–72 (2013).

Bernardi, S., Benna, M. K., Rigotti M., Munuera, J., Fusi, S., Salzman, D. The geometry of abstraction in hippocampus and prefrontal cortex. *bioRxiv* 10.1101/408633 (2019).

Bisley, J. W., Krishna, B. S., & Goldberg, M. E. A rapid and precise on-response in posterior parietal cortex. *J. of Neurosci.* **24**, 1833–1838 (2004).

Blackman, R. K., Crowe, D. A., DeNicola, A. L., Sakellaridi, S., MacDonald, A. W., Chafee, M. V. Monkey Prefrontal neurons reflect logical operations for cognitive

control in a variant of the AX continuous performance task (AX-CPT). *J. Neurosci.* **36**, 4067–79 (2016).

Boch, R. A., & Goldberg, M. E. Participation of prefrontal neurons in the preparation of visually guided eye movements in the rhesus monkey. *J. of Neurophysiol.*, ***61***, 1064–1084 (1989).

Bondy, A. G., Haefner, R. M. & Cumming, B. G. Feedback determines the structure of correlated variability in primary visual cortex. *Nat. Neurosci.* **21**, 598–606 (2018).

Boran, E., Fedele, T., Klaver, P. et al. Persistent hippocampal neural firing and hippocampal–cortical coupling predict verbal working memory load. *Sci. Adv.* **5**, eaav3687 (2019).

Born, R. T. & Bradley, D. C. Structure and function of visual area MT. *Annu. Rev. Neurosci.* **28**, 157–189 (2005).

Bosman, C.A., Womelsdorf, T., Desimone, R., and Fries, P. A microsaccadic rhythm modulates gamma-band synchronization and behavior. *J. Neurosci*. **29**, 9471–9480 (2009).

Bouchacourt, F. & Buschman, T. J. A flexible model of working memory. *Neuron* **103**, 1–14 (2019).

Brascamp, J., Blake, R. & Knapen, T. Negligible fronto-parietal BOLD activity accompanying unreportable switches in bistable perception. *Nat. Neurosci.* **18**, 1672–1678 (2015).

Brascamp, J., Sterzer, P., Blake, R. & Knapen, T. Multistable Perception and the Role of the Frontoparietal Cortex in Perceptual Inference. *Annu. Rev. Psychol.* **69**, 77–103 (2018).

Brincat, S. L., and Miller, E. K. Frequency-specific hippocampal-prefrontal interactions during associative learning. *Nat. Neurosci.* **18**, 576–581 (2015).

Britten, K. H., Newsome, W. T., Shadlen, M. N., Celebrini, S., Movshon, J. A. A realtionship between behavioral choice and the visual responses of neurons in macaque MT. *Vis. Neurosci.* **13**, 87-100 (1996).

Brody, C. D., Hernández, A., Zainos, A., Romo, R. Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. *Cerebr. Cortex* **13**, 1196–1207 (2013).

Bruce, C. J., & Goldberg, M. E. Primate frontal eye fields. I. Single neurons discharging before saccades. *J. of Neurophysiol.* **53***,* 603–635 (1985).

Buckley, M.J., Mansouri, F.A., Hoda, H., Mahboubi, M., Browning, P.G., Kwok, S.C., Phillips, A., and Tanaka, K. Dissociable components of rule guided behavior depend on distinct medial and prefrontal regions. *Science* **325**, 52–58 (2009).

Bullock, K. R., Pieper, F., Sachs, A. J., Martinez-Trujillo, J. C. Visual and presaccadic activity in area 8Ar of the macaque monkey lateral prefrontal cortex. *J. Neurophysiol.* **118**, 15–28 (2017).

Buonomano, D. V, Maass, W. State-dependent computations: spatiotemporal processing in cortical networks. *Nat. Rev. Neurosci*. **10**, 113–25 (2009).

Buračas, G. T., Zador, A. M., DeWeese, M. R., Albright, T. D. Efficient discrimination of temporal patterns by motion-sensitive neurons in primate visual cortex. *Neuron* **20**, 959–69 (1998).

Buschman, T. J., Siegel, M., Roy, J. E., Miller, E. K. Neural substrates of cognitive capacity limitations. *PNAS* **10**, 11252–55 (2011).

Butters, N., and Pandya, D. Retention of delayed-alternation: Effect of selective lesions of sulcus principalis. *Science* **165**, 1271–73 (1969).

Cao, R., Pastukhov, A., Mattia, M. & Braun, J. Collective Activity of Many Bistable Assemblies Reproduces Characteristic Dynamics of Multistable Perception. *J. Neurosci.* **36**, 6957–6972 (2016).

Carandini, M., Heeger, D. J. Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* **13**, 51–62 (2012).

Catani, M., Dell'acqua, F., Vergani, F., Malik, F., Hodge, H., Roy, P., Valabregue, R., and Thiebaut de Schotten, M. Short frontal lobe connections of the human brain. *Cortex* **48**, 273–291 (2012).

Cavanagh, S. E., Wallis, J. D., Kennerley, S. W., Hunt, L. T. Autocorrelation structure at rest predicts value correlates of single neurons during reward-guided choice. *eLife* **5**, e18937 (2016).

Cavanagh, S. E., Towers, J. P., Wallis, J. D. et al. Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex. *Nat. Commun*. **9**, 3498 (2018).

Celebrini, S. & Newsome, W. T. Neuronal and psychophysical sensitivity to motion signals in extrastriate area MST of the macaque monkey. *J. Neurosci.* **14**, 4109–4124 (1994).

Chafee, M. V. and Goldman-Rakic, P. S. Matching patterns of activity in primate prefrontal area and parietal area neurons during a spatial working memory task. *J. of Neurophysiol.* **79**, 2919–40 (1998).

Chaudhuri, R., Knoblauch, K., Gariel, M. A., Kennedy, H., Wang, X. J. A large-scale circuit mechanism for hierarchical dynamical processing in the primate cortex. *Neuron* **88**, 419–31 (2015).

Chaudhuri, R., Gerçek, B., Pandey, B., Peyrache, A., Fiete, I. The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep. *Nat. Neurosci.* **22**, 1512-1520 (2019).

Chaisangmongkon, W., Swaminathan, S. K., Freedman, D. J., Wang X. J. Computing by Robust Transience: How the fronto-parietal network performs sequential, category-based decisions. *Neuron* **93**, 1504–1517 (2017).

Chang, L. & Tsao, D. Y. The code for facial identity in the primate brain. *Cell* **169**, 1013–1028 (2017).

Chelazzi, L., Duncan, J., Miller, E. K, Desimone, R. Responses of Neurons in Inferior Temporal Cortex During Memory-Guided Visual Search. *J. of Neurophysiol*. **80**, 2918–2940 (1998).

Chelazzi, L., Miller, E. K., Duncan, J., Desimone, R. Responses of neurons in macaque area V4 during memory-guided visual search. *Cerebral Cortex* **11**, 761–772 (2001).

Chen, J., Hasson, U., Honey, C. J. Hierarchical process memory: memory as an integral component of information processing. *Neuron* **88**, 244–46 (2014).

Chicharro, D., Panzeri, S. & Haefner, R. M. Decision-Related Signals In The Presence Of Nonzero Signal Stimuli, Internal Bias, And Feedback. *bioRxiv* 118398 (2017).

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Ryu, S. I., Shenoy, K. V. Cortical preparatory activity: representation of movement or first cog in a dynamical machine? *Neuron* **68**:387-400 (2010).

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I., Shenoy, K. V. Neural population dynamics during reaching. *Nature* 487-51 (2012).

Churchland, P. S. & Sejnowski, T. The Computational Brain (MIT Press, 1992).

Cicmil, N., Parker, A. & Krug, K. Changes in Variance of Neuronal Signals May Be Perceptually Relevant for Stereo Vision. *BICT.* doi:10.4108/eai.3-12-2015.2262431 (2016).

Cicmil, N., Cumming, B. G., Parker, A. J. & Krug, K. Reward modulates the effect of visual cortical microstimulation on perceptual decisions. *eLife* **4**, e07832 (2015).

Clark, K. L., Noudoost, B., and Moore, T. Persistent spatial information in the frontal eye field during object-based short-term memory. *J. Neurosci*. **32**, 10907–10914 (2012).

Cohen, J. D. Cognitive control: Core constructs and current considerations. *The Wiley handbook of cognitive control*. (Ed. Egner, T.) 2-28 (Wiley-Blackwell, Oxford, 2017).

Cohen, J. D., Dunbar, K., McClelland, J. L. On the control of automatic processes: a

parallel
distributed processing account of the Stroop effect. *Psychol. Rev.* **97**, 332–61 (1990).

Cohen, M. R. and Newsome, W. T. Context-dependent changes in functional circuitry in visual area MT. *Neuron* **60**, 162–73 (2008).

Cohen, M. R. & Maunsell, J. H. R. Attention improves performance primarily by reducing interneuronal correlations. *Nat. Neurosci.* **12**, 1594–1600 (2009).

Cohen, M. R. and Kohn A. Measuring and interpreting neuronal correlations. *Nat. Neurosci.* **14**, 811–19 (2011).

Cohen, M. R. & Maunsell, J. H. R. A Neuronal Population Measure of Attention Predicts Behavioral Performance on Individual Trials. *J. Neurosci.* **30**, 15241–15253 (2010).

Cohen, M. R. & Maunsell, J. H. R. Using Neuronal Populations to Study the Mechanisms Underlying Spatial and Feature Attention. *Neuron* **70**, 1192–1204 (2011).

Cohen, Y., Schneidman, E., Paz, R. A geometric representation unveils learning dynamics in primate neurons, *bioRxiv* 561670 (2019).

Compte, A., Brunel, N., Goldman-Rakic, P. S., Wang, X. J. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebr. Cortex* **10**, 910–23 (2000).

Constantinidis, C., Franowicz, M. N. & Goldman-Rakic, P. S., The sensory nature of mnemonic representation in the primate prefrontal cortex. *Nat. Neurosci.* 4, 311-316 (2001).

Constantinidis, C., Funahashi, S., Lee, D., Murray, J. D., Qi, X.-L., Wang, M., & Arnsten, A. F. T. Persistent Spiking Activity Underlies Working Memory. *J. Neurosci.*, ***38***(32), 7020–7028 (2018).

Constantinidis, C., & Qi, X.-L. Representation of Spatial and Feature Information in the Monkey Dorsal and Ventral Prefrontal Cortex. Frontiers in Integrative Neuroscience, **12** (2018).

Constantinidis, C. & Steinmetz, M. A. Neuronal activity in posterior parietal area 7a during the delay periods of a spatial memory task. *J. Neurophysiol.* **76**, 1352–1355 (1996).

Cook, E. P. & Maunsell, J. H. R. Dynamics of neuronal responses in macaque MT and VIP during motion detection. *Nat. Neurosci.* **5**, 985–994 (2002).

Cromer, J. A., Roy, J. E., Miller, E. K. Representation of multiple, independent categories in the primate prefrontal cortex. *Neuron* **66**, 796–807 (2010).

Croxson, P.L. et al. Quantitative investigation of connections of the prefrontal cortex in

the human and macaque using probabilistic diffusion tractography. *J. Neurosci.* **25**, 8854-8866 (2005).

Cueva, J. C., Marcos, E., Saez, A., Genovesio, A., Jazayeri, M., Romo, R., Salzman, D. C., Shadlen, M. N., Fusi, S. Low dimensional dynamics for working memory and time encoding. *bioRxiv* 10.1101/504936 (2019).

Cumming, B. G. & Nienborg, H. Feedforward and feedback sources of choice probability in neural population responses. *Curr. Opin. Neurobiol.* **37**, 126–132 (2016).

Cunningham, J. P., and Yu, B. M. Dimensionality Reduction for Large-Scale Neural Recordings. *Nat. Neurosci.* **17**, 1500–1509 (2014).

Curtis, C. E., D'Esposito, M. Persistent activity in the prefrontal cortex during working memory. *Trends in Cogn. Sci.* **7**, 415-423 (2003).

D'Esposito, M., Detre, J. A., Alsop, D. C., Shin, R. K., Atlas, S., Grossman, M. The neural basis of the central executive system of working memory. *Nature* **378**, 279–281 (1995).

De la Rocha, J., Doiron, B., Shea-Brown, E., Josic, K., Reyes, A. Correlation between neural spike trains increases with firing rate. *Nature* **448**, 802–6 (2007).

Diamond, A. Normal development of prefrontal cortex from birth to young adulthood: Cognitive functions, anatomy, and biochemistry. *In Principles of Frontal Lobe Function*, (Eds. Stuss D. T. and Knight, R. T.) 466-503 (Oxford University Press, Oxford, 2002).

Dias, R., Robbins, T. W., and Roberts, A. C. Primate analogue of the Wisconsin Card Sorting Test: effects of excitotoxic lesions of the prefrontal cortex in the marmoset. *Behav. Neurosci.* **110**, 872–886 (1996b).

DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).

Di Pellegrino, G. & Wise, S. P. Visuospatial versus visuomotor activity in the premotor and prefrontal cortex of a primate. *J. Neurosci.* **13**, 1227–1243 (1993).

Dodd, J. V., Krug, K., Cumming, B. G. & Parker, A. J. Perceptually Bistable Three-Dimensional Figures Evoke High Choice Probabilities in Cortical Area MT. *J. Neurosci.* **21**, 4809–4821 (2001).

Doiron, B., Litwin-Kumar, A., Rosenbaum, R., Ocker, G. K. & Josić, K. The mechanics of state-dependent neural correlations. *Nat. Neurosci.* **19**, 383–393 (2016)

Donahue, C. J., Glasser, M. F., Preuss, T. M, Rilling, J. K., Van Essen, D. C. Quantitative assessment of prefrontal cortex in humans relative to nonhuman primates. *PNAS*, 201721653 (2018).

Donahue, C.H. & Lee, D. Dynamic routing of task-relevant signals for decision making in dorsolateral prefrontal cortex. *Nat. Neurosci.***18**, 295–301 (2015).

Dreher, J. C., Koechlin, E., Tierney, M., and Grafman, J. Damage to the fronto-polar cortex is associated with impaired multitasking. *PLoS ONE* 3, e3227 (2008).

Druckmann, S., and Chklovskii, D. B. Neuronal circuits underlying persistent representations despite time varying activity. *Curr. Biol.* **22**, 2095–2103 (2012).

Dubner, R. & Zeki, S. M. Response properties and receptive fields of cells in an anatomically defined region of the superior temporal sulcus in the monkey. *Brain Res.* **35**, 528–532 (1971).

Duncan, J. The demonstration of capacity limitation. *Cognit. Psychol.* **12**, 75–96 (1980).

Duncan, J. Disorganisation of behaviour after frontal lobe damage. *Cogn. Neuropsychol.* **37**, 271–290 (1986).

Duncan, J. The adaptive coding model of neural function in prefrontal cortex. *Nat. Rev. Neurosci.* **2**, 820-829 (2001).

Duncan, J., and Miller, E. K. Cognitive focus through adaptive neural coding in the primate prefrontal cortex. *Principles of Frontal Lobe Function*, 278–91 (Oxford University Press, Oxford, 2002).

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

Duncan, J., Mitchell, D. J. Training refines brain representations for multitasking, *PNAS* **112**, 14127-14128 (2015).

Duncan, J., Emslie, H., Williams, P., Johnson, R., & Freer, C. Intelligence and the frontal lobe: the organization of goal-directed behavior. *Cognitive Psychology*, **30**, 257–303 (1996).

Duncan, J., Johnson, R., & Swale, M. J. Frontal Lobe Deficits after Head Injury: Unity and Diversity of Function. *Cognitive Neuropsychology, **14***, 713–741 (1997).

Duong, L., Leavitt, M., Pieper, F., Sachs, A. & Martinez-Trujillo, J. A Normalization Circuit Underlying Coding of Spatial Attention in Primate Lateral Prefrontal Cortex. *eneuro* **6**, 1081 (2019).

Durstewitz, D., Vittoz, N. M., Floresco, S. B., and Seamans, J. K. Abrupt tran- sitions between prefrontal neural ensemble states accompany behavioral transitions during rule learning. *Neuron* **66**, 438–48 (2010).

Ecker, A. S. *et al.* Decorrelated Neuronal Firing in Cortical Microcircuits. *Science* **327**, 584–587 (2010).

Ecker, A. S., Denfield, G. H., Bethge, M. & Tolias, A. S. On the Structure of Neuronal Population Activity under Fluctuations in Attentional State. *J. Neurosci.* **36**, 1775–1789 (2016).

Egorov, A. V., Hamam, B. N., Fransén, E. et al. Graded persistent activity in entorhinal cortex neurons. *Nature* **420**, 173–178 (2002).

Eiselt, A. K. & Nieder, A. Representation of abstract quantitative rules applied to spatial and numerical magnitudes in primate prefrontal cortex. *J. Neurosci.* **33**, 7526–7534 (2013).

Eichenbaum, H. Barlow versus Hebb: when is it time to abandon the notion of feature detectors and adopt the cell assembly as the unit of cognition? *Neurosci Lett.* (2017).

Elsayed, G. F., Lara, A. H., Kaufman, M. T., Churchland, M. M., Cunningham, J. P. Reorganization between preparatory and movement population responses in motor cortex. *Nat. Commun.* **7**, 13239 (2016).

Eggermont, J. J. & Smith, G. M. Neural connectivity only accounts for a small part of neural correlation in auditory cortex. *Exp. Brain Res.* **110**, 379–391 (1996).

Eggermont, J. J. & Smith, G. M. Separating local from global effects in neural pair correlograms. *Neuroreport* **6**, 2121–2124 (1995).

Enel, P. Procyk, E., Quilodran, R., Dominey, P. F. Reservoir computing properties of neural dynamics in prefrontal cortex. *PLOS Comput. Biol.* **12**, e1004967 (2016).

Engel, T.A., Chaisangmongkon, W., Freedman, D.J. & Wang, X.J. Choice-correlated activity fluctuations underlie learning of neuronal category representation. *Nat. Commun.* **6**, 6454 (2015).

Felleman, D. J., and Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cerebr. Cortex* **1**, 1–47 (1991).

Finkelstein, A., Ulanovsky, N., Tsodyks, M. & Aljadeff, J. Optimal dynamic coding by mixed-dimensionality neurons in the head-direction system of bats. *Nature communications* **9**, 3590 (2018).

Flechsig, P. Anatomie des Menschlichen Gehirns und Rückenmarks auf Myelogenetischer Grundlage. (Thieme, Leipzig, 1920).

Fransén, E., Tahvildari, B., Egorov, A. V., Hasselmo, M. E., Alonso, A. A. Mechanism of graded persistent cellular activity of entorhinal cortex layer V neurons. *Neuron* **49**, 735–46 (2006).

Freedman, D. J., Riesenhuber, M., Poggio, T. & Miller, E. K. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* **291**, 312–316 (2001).

Funahashi, S., Bruce, C. J., Goldman-Rakic, P. S. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J. Neurophysiol.* **61**, 331–49 (1989).

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. Visuospatial coding in primate prefrontal neurons revealed by oculomotor paradigms. *J. of Neurophysiol.*, **63**, 814–831 (1990).

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. Neuronal activity related to saccadic eye movements in the monkey's dorsolateral prefrontal cortex. *J. of Neurophysiol.*, **65**, 1464–1483 (1991).

Funahashi, S., Bruce, C. J., and Goldman-Rakic, P. S. Dorsolateral prefrontal lesions and oculomotor delayed-response performance: evidence for mnemonic ''scotomas''. *J. Neurosci.* **13**, 1479–1497 (1993).

Fusi, S., Miller, E. K., & Rigotti, M. Why neurons mix: high dimensionality for higher cognition. Current opinion in neurobiology **37**, 66-74 (2016).

Fuster, J. M. Network memory. *Trends Neurosci.* **20**, 451-459 (1997).

Fuster, J. M., and Alexander, G. E. Neuron activity related to short-term memory. *Science* **173**, 652–54, (1971).

Fuster, J. M. & Jervey, J. P. Neuronal firing in the inferotemporal cortex of the monkey in a visual memory task. *J. Neurosci.* **2**, 361–375 (1982).

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

Fuster, J. M. Frontal lobe and cognitive development. *J. Neurocytol.* **31**, 373–385 (2002).

Fuster, J. M. The Prefrontal Cortex 4th edn (*Academic*, 2008).

Fuster, J. M., Bodner, M. & Kroger, J. K. Cross-modal and cross-temporal association in neurons of frontal cortex. *Nature* **405**, 347–351 (2000).

Gallego, J. A., Perich, M. G., Naufel, S. N., Ethier, C., Solla, S. A., Miller, L.E. Cortical population activity within a preserved neural manifold underlies multiple motor behaviors. *Nat. Commun.* **9**, 4233 (2018).

Ganguli, S., Latham, P. Feedforward to the past: the relation between neuronal connectivity, amplification, and short-term memory. *Neuron* **61**, 499–501 (2009).

Gao, P., & Ganguli, S. On simplicity and complexity in the brave new world of largescale neuroscience. *Current opinion in neurobiology*, **32**, 148-155 (2015).

Garner, K. G., Dux, P. E. Training conquers multitasking costs by dividing task representations in the frontoparietal-subcortical system. *PNAS* **112**, 14372–77 (2015).

Gauthier, B., Eger, E., Hesselmann, G., Giraud, A. L., Kleinschmidt, A. Temporal tuning properties along the human ventral visual stream. *J. Neurosci.* **32**, 14433–41 (2012).

Genovesio, A., Tsujimoto, S., Wise, S.P. Encoding goals but not abstract magnitude in the primate prefrontal cortex**.** *Neuron*, **74**, 656-662 (2012).

Genovesio, A., Brasted, P. J., Mitz, A. R., Wise, S. P. Prefrontal cortex activity related to abstract response strategies. *Neuron* **47**, 307–320 (2005).

Gnadt, J. W. & Andersen, R. A. Memory related motor planning activity in posterior parietal cortex of macaque. *Exp. Brain Res.* **70**, 216–220 (1988).

Godfrey-Smith, P. Environmental complexity and the evolution of cognition. *The Evolution of Intelligence* (Eds. Sternberg, R. and Kaufman, J.). (Lawrence and Erlbaum Associates, Mahwah, New Jersey, 2001)

Goel, V., and Grafman, J. Are the frontal lobes implicated in 'planning' functions? Interpreting data from the Tower of Hanoi. *Neuropsychologia* **33**, 623–642 (1995).

Goldman, M. S., Compte, A., Wang, X. J. *Encyclopedia of Neuroscience* (Ed. Squire L.R.). 165-178 (Academic Press, Oxford, 2008).

Goldman-Rakic, P. S. Cellular basis of working memory. *Neuron* **14**, 477–85 (1995).

Goldman-Rakic, P. S., Selemon, L. D., Schwartz, M. L. Dual pathways connecting the dorsolateral prefrontal cortex with the hippocampal formation and parahippocampal cortex in the rhesus monkey. *Neuroscience* **12**, 719–43 (1984).

Goldman-Rakic, P. S. Circuitry of Primate Prefrontal Cortex and Regulation of Behavior by Representational Memory. (Hoboken, NJ, USA: John Wiley & Sons, 1987).

Goldman-Rakic, P. S., Bates, J. F., Chafee, M. V. The prefrontal cortex and internally generated motor acts. *Curr. Op. in Neurobiol.* **2**, 830–835 (1992).

Goldman, P. S., Rosvold, H. E., Vest, B., Galkin, T. W. Analysis of the delayed-alternation deficit produced by dorsolateral prefrontal lesions in the rhesus monkey. *J Comp Physiol Psychol* **77**, 212–220 (1971).

Goldman, M. S. Memory without feedback in a neural network. *Neuron* **61**, 621–34 (2009).

Grafman, J. Plans, actions, and mental states: managerial knowledge units in the frontal lobes. *In Integrating Theory and Practice in Clinical Neurpsychology*, (Ed. Perecman, E.) 93-138 (Hillsdale, NJ: Erlbaum, 1989).

Green, D. M. & Swets, J. A. *Signal detection theory and psychophysics*. (Wiley, 1966).

Gu, Y., Angelaki, D. E. & DeAngelis, G. C. Neural correlates of multisensory cue integration in macaque MSTd. *Nat. Neurosci.* **11**, 1201–1210 (2008).

Gu, Y. *et al.* Perceptual learning reduces interneuronal correlations in macaque visual cortex. *Neuron* **71**, 750–761 (2011).

Gutnisky, D. A. & Dragoi, V. Adaptive coding of visual information in neural populations. *Nature* **452**, 220–224 (2008).

Haefner, R. M., Gerwinn, S., Macke, J. H. & Bethge, M. Inferring decoding strategies from choice probabilities in the presence of correlated variability. *Nat. Neurosci.* **16**, 235–242 (2013).

Haefner, R. M., Berkes, P. & Fiser, J. The implications of perception as probabilistic inference for correlated neural variability during behavior. *ArXiv*14090257 *Q-Bio* (2014).

Haefner, R. M., Berkes, P. & Fiser, J. Perceptual Decision-Making as Probabilistic Inference by Neural Sampling. *Neuron* **90**, 649–660 (2016).

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., Rubin, N. A hierarchy of temporal receptive windows in human cortex. *J. Neurosci.* **28**, 2539–50 (2008).

Hebb, D. O. The Organization Of Behaviour. (Wiley, 1949).

Hebb, D. O. The Organization of Behavior: A Neuropsychological Theory. (Psychology Press; 2005).

Helias, M., Tetzlaff, T. & Diesmann, M. The correlation structure of local neuronal networks intrinsically results from recurrent dynamics. *PLoS Comput. Biol.* **10**, e1003428 (2014).

Herrington, T. M. *et al.* The Effect of Microsaccades on the Correlation between Neural Activity and Behavior in Middle Temporal, Ventral Intraparietal, and Lateral Intraparietal Areas. *J. Neurosci.* **29**, 5793–5805 (2009).

Honey, C. J., Thesen, T., Donner, T. H., Silbert, L. J., Carlson, C. E., Devinsky, O., Doyle, W. K., Rubin, N., Heeger, D. J., Hasson, U. Slow cortical dynamics and the accumulation of information over long timescales. *Neuron* **76**, 423–34 (2012).

Hopfield, J.J. Neural networks and physical systems with emergent collective computational abilities. *PNAS* **79**, 2554–2558 (1982).

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

Huang, X. & Lisberger, S. G. Noise Correlations in Cortical Area MT and Their Potential Impact on Trial-by-Trial Variation in the Direction and Speed of Smooth-Pursuit Eye Movements. *J. Neurophysiol.* **101**, 3012–3030 (2009).

Hubel, D. H., and Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. of Physiol.* **160**, 106–154 (1962).

Hunt, L. T. et al. Triple dissociation of attention and decision computations across prefrontal cortex. *Nat. Neurosci.* **21**, 1471–1481 (2018).

Hupé, J. M. et al. Feedback Connections Act on the Early Part of the Responses in Monkey Visual Cortex. *J. Neurophysiol.* **85**, 134–145 (2001).

Huttenlocher, P.R. Morphometric study of human cerebral cortex development. *Neuropsychologia* **28**, 517–527 (1990).

Ilinsky, I. A., Jouandet, M. L., and Goldman-Rakic, P. S. Organization of the nigrothalamocortical system in the rhesus monkey. *J. Comp. Neurol*. **236**, 315–330 (1985).

Inagaki, H.K., Fontolan, L., Romani, S. et al. Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature* **566**, 212–217 (2019).

Ito, J., Maldonado, P., and Grün, S. Cross-frequency interaction of the eye-movement related LFP signals in V1 of freely viewing monkeys. *Front. Syst. Neurosci*. **7**, 1 (2013).

Jacob, S. N. & Nieder, A. Complementary roles for primate frontal and parietal cortex in guarding working memory from distractor stimuli. *Neuron* **83**, 226–237 (2014).

Jacobsen, C.F. Functions of frontal association area in primates. *Arch NeurPsych* **33**, 558–569 (1935).

Jaeger, H. & Haas, H. Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication. *Science* **304**, 78–80 (2004).

Jerison, H.J. Evolution of the brain. *Neuropsychology* (Ed. Zaidel, D. W.) 53-81 (Academic Press, Inc., San Diego, 1994).

Johnston, W. J., Palmer, S. E., Freedman, D. J. Nonlinear mixed selectivity supports reliable neural computation. *bioRxiv* 577288 (2019).

Kadohisa, M., Kusunoki, M., Petrov, P., Sigala, N., Buckley, M. J., Gaffan, D., et al. Spatial and temporal distribution of visual information coding in lateral prefrontal cortex. *Eur. J. Neurosci.* **41**, 89–96 (2015).

Kahneman, D. Attention and Effort. (Prentice Hall, New Jersey, 1973).

Kalaska, J. F. Emerging ideas and tools to study the emergent properties of the cortical neural circuits for voluntary motor control in non-human primates. *F1000Res* **8**, (2019).

Kamiński, J., Sullivan, S., Chung, J.M. et al. Persistently active neurons in human

medial frontal and medial temporal lobe support working memory. *Nat. Neurosci.* **20**, 590-601 (2017).

Kornblith, S., Quian Quiroga, R., Koch, C. et al. Persistent single-neuron activity during working memory in the human medial temporal lobe. *Curr. Biol*. **27**, 1026–1032 (2017).

Kaufman, M.T., Churchland, M. M., Ryu, S. I., Shenoy, K. V. Cortical activity in the null space: permitting preparation without movement. *Nat. Neurosci.* **17**: 440-448 (2014).

Keemink, S. W., Machens, C.K. Decoding and Encoding (de) mixed Populations Responses. *Curr. Op. Neurobiol.* **58**, 112-121 (2019).

Kennerley, S. W., Walton, M. E., Behrens, T. E., Buckley, M. J., and Rushworth, M.F. Optimal decision making and the anterior cingulate cortex. *Nat. Neurosci.* **9**, 940–947 (2006).

Kiani, R., Hanks, T. D. & Shadlen, M. N. Bounded Integration in Parietal Cortex Underlies Decisions Even When Viewing Duration Is Dictated by the Environment. *J. Neurosci.* **28**, 3017–3029 (2008).

Kline, R. B. Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research. (American Psychological Association, Washington DC, 2004)

Klingberg, T. & Roland, P. E. Interference between two concurrent tasks is associated with activation of overlapping fields in the cortex. *Brain Res. Cogn. Brain Res.* **6**, 1–8 (1997).

Knutsen, P. M. & Ahissar, E. Orthogonal coding of object location. *Trends Neurosci.* **32**, 101–108 (2009).

Kobak, D., Brendel, W., Constantinidis, C., Feierstein, C. E., Kepecs, A., Mainen, Z.F., Qi X-L., Romo, R., Uchida, N., Machens, C. K. Demixed Principal component analysis of neural population data. *eLife,* **5**: e10989 (2016).

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

Kohn, A., Coen-Cagli, R., Kanitscheider, I. & Pouget, A. Correlations and Neuronal Population Information. *Annu. Rev. Neurosci.* **39**, null (2016).

Kohn, A. & Smith, M. A. Stimulus Dependence of Neuronal Correlation in Primary Visual Cortex of the Macaque. *J. Neurosci.* **25**, 3661–3673 (2005).

Kollmorgen, S., Newsome, W. T., Mante, V. Spatial and temporal structure of choice representations in primate prefrontal cortex. *bioRxiv* 595520 (2019).

Koyluoglu, O. O., Pertzov, Y., Manohar, S., Husain, M., Fiete, I. R. Fundamental bound on the persistence and capacity of short-term memory stored as graded persistent activity. *eLife*, doi:10.7554 (2017).

Kreiman, G., Koch, C. & Fried, I. Category-specific visual responses of single neurons in the human medial temporal lobe. *Nat. Neurosci.* 946–953 (2000).

Krug, K. & Parker, A. J. Neurons in Dorsal Visual Area V5/MT Signal Relative Disparity. *J. Neurosci.* **31**, 17892–17904 (2011).

Krug, K., Cicmil, N., Parker, A. J. & Cumming, B. G. A Causal Role for V5/MT Neurons Coding Motion-Disparity Conjunctions in Resolving Perceptual Ambiguity. *Curr. Biol.* **23**, 1454–1459 (2013).

Krug, K. A common neuronal code for perceptual processes in visual cortex? Comparing choice and attentional correlates in V5/MT. *Philos. Trans. R. Soc. B Biol. Sci.* **359**, 929–941 (2004).

Krug, K., Curnow, T. L. & Parker, A. J. Defining the V5/MT neuronal pool for perceptual decisions in a visual stereo-motion task. *Philos. Trans. R. Soc. B Biol. Sci.* **371**, (2016).

Krug, K., Brunskill, E., Scarna, A., Goodwin, G. M. & Parker, A. J. Perceptual switch rates with ambiguous structure-from-motion figures in bipolar disorder. *Proc. R. Soc. B Biol. Sci.* **275**, 1839–1848 (2008).

Krug, K., Cumming, B. G. & Parker, A. J. Comparing Perceptual Signals of Single V5/MT Neurons in Two Binocular Depth Tasks. *J. Neurophysiol.* **92**, 1586–1596 (2004).

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

Kwon, S. E., Yang, H., Minamisawa, G. & O'Connor, D. H. Sensory and decision-related activity propagate in a cortical feedback loop during touch perception. *Nat. Neurosci.* **19**, 1243–1249 (2016).

Lara, A. H. & Wallis, J. D. Executive control processes underlying multi-item working memory. *Nat. Neurosci.* **17**, 876–883 (2014).

Lara, A. H., Kennerley, S. W., Wallis, J. D. Encoding of gustatory working memory by orbitofrontal neurons. *J. Neurosci.* **29**, 765-774 (2009).

Laurent, G. Olfactory network dynamics and the coding of multidimensional signals. *Nature Rev. Neurosci*. **3**, 884–896 (2002).

Lebedev, M.A., Messinger, A., Kralik, J.D., Wise, S.P. Representation of attended versus remembered locations in prefrontal cortex. *PLoS. Biol*. **2**, e365 (2004).

Lennert, T., and Martinez-Trujillo, J. Strength of response suppression to distracter

stimuli determines attentional-filtering performance in primate prefrontal neurons. *Neuron* **70**, 141–152 (2011).

Lindsay, G. W., Rigotti, M., Warden, M. R., Miller, E. K., and Fusi, S. Hebbian learning in a random network captures selectivity properties of the prefrontal cortex. *J. Neurosci*. **37**, 11021–11036 (2017).

Litwin-Kumar, A., Harris, K. D., Axel, R., Sompolinsky, H. & Abbott, L. Optimal Degrees of Synaptic Connectivity. *Neuron* **7**, 1153–1164 (2017).

Liu, S., Gu, Y., DeAngelis, G. C. & Angelaki, D. E. Choice-related activity and correlated noise in subcortical vestibular neurons. *Nat. Neurosci.* **16**, 89–97 (2013)

Loewenstein, Y., and Sompolinsky, H. Temporal integration by calcium dynamics in a model neuron. *Nat. Neurosci.* **6**, 961–67 (2003).

Lowet, E., Roberts, M. J., Fries, P., De Weerd, P. Areas V1 and V2 show microsaccade-related 3-5Hz covariation in gamma power and frequency. *Eur. J. Neurosci.* **43**, 1286-1296.

Lowet, E., Gomes, B., Srinivasan, K., Zhou, H., Schafer, R. J., Desimone, R. Enhanced neural processing by covert attention only during microsaccades directed toward the attended stimulus. *Neuron* **99**, 207-2014 (2018).

Lumer, E. D., Friston, K. J. & Rees, G. Neural Correlates of Perceptual Rivalry in the Human Brain. *Science* **280**, 1930–1934 (1998).

Lundqvist, M., Rose, J., Herman, P., Brincat, S. L., Buschman, T. J., Miller, E. K. Gamma and beta bursts underlie working memory. *Neuron* **90**, 152–64 (2016).

Lundqvist, M., Herman, P., Miller, E. K. Working memory: Delay activity, Yes! Persistent activity? Maybe not. *J. Neurosci.* **38**(32), 7013-7019. (2018).

Luo, T.Z., Maunsell, J. H. R. Attentional changes in either criterion or sensitivity are associated with robust modulations in lateral prefrontal cortex. *Neuron* **97***,* 1382–1393 *(*2018*)*.

Luria, A.R. Higher Cortical Functions in Man. (New York: Basic Books, 1966).

Luria, A.R. The frontal lobes and the regulation of behavior. *In Psychophysiology of the frontal lobes* (Pibram, K. H.) (3-26 Oxford, UK: Academic Press, 1973)

Maass, W., Natschläger, T., Markram, H. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Comput.* **14**, 2531–60 (2002).

Machens, C. K., Romo, R., Brody, C. D. Functional, but not anatomical, separation of "what" and "when" in prefrontal cortex. *J. Neurosci.* **30**: 350-360 (2010).

Machens, C. K., Romo, R., Brody, C. D. Flexible control of mutual inhibition: A neural model of two-interval discrimination. *Science* **307**, 1121–24 (2005).

Mack, M. L., Preston, A. R., Love, B. Ventromedial prefrontal cortex compression during concept learning. *bioRxiv* 178145 (2019).

Manes, F., Sahakian, B., Clark, L., Rogers, R., Antoun, N., Aitken, M., and Robbins, T. Decision-making processes following damage to the prefrontal cortex. *Brain* **125**, 624–639 (2002).

Manohar, S. G., Zokaei, N., Fallon, S. J., Vogels, T., Husain, M. Neural mechanisms of attending to items in working memory. *Neuroscience & Biobehavioral Reviews* **101**, 1-12 (2019).

Mansouri, F. A., Matsumoto, K. & Tanaka, K. Prefrontal cell activities related to monkeys' success and failure in adapting to rule changes in a Wisconsin card sorting test analog. *J. Neurosci.* **26**, 2745–2756 (2006).

Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).

Maris, E., and Oostenveld, R. Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* **164**. 177–90 (2007).

Marcos, E., Tsujimoto, S., Genovesio, A. Independent coding of absolute duration and distance magnitudes in the prefrontal cortex. *J. Neurophysiol.* **117**, 195–203 (2017).

Marcos, E., Tsujimoto, S., Mattia, M. & Genovesio, A. A Network Activity Reconfiguration Underlies the Transition from Goal to Action. *Cell Rep.* **27**, 2909-2920 (2019).

Markov, N. T. et al. A weighted and directed interareal connectivity matrix for macaque cerebral cortex. Cereb. *Cortex* **24**, 17–36 (2014).

Markowitz, D. A., Curtis, C. E., Pesaran, B. Multiple component networks support working memory in prefrontal cortex. *PNAS* **112**, 11084–89 (2015).

Marr, D. A theory of cerebellar cortex. *J. Physiol.* **202**, 437-470 (1969).

Masse, N. Y., Hodnefield, J. M., Freedman, D. J. Mnemonic Encoding and Cortical Organization in Parietal and Prefrontal Cortices. *J. Neurosci.* **37**, 6098–6112 (2017).

Masse, N. Y., Yang, G. R., Song, H. F., Wang, X.-J. & Freedman, D. J. Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nat. Neurosci.* **22**, 1159–1167 (2019).

Mastronarde, D. N. Correlated firing of cat retinal ganglion cells. I. Spontaneously active inputs to X- and Y-cells. *J. Neurophysiol.* **49**, 303–324 (1983).

Maunsell, J. H. & Essen, D. van. The connections of the middle temporal visual area (MT) and their relationship to a cortical hierarchy in the macaque monkey. *J. Neurosci.* **3**, 2563–2586 (1983).

Maunsell, J. H. R. Neuronal Mechanisms of Visual Attention. *Annu. Rev. Vis. Sci.* **1**, 373–391 (2015).

Matsushima, A., Tanaka, M. Different neuronal computations of spatial working memory for multiple locations within versus across visual hemifields. *J. Neurosci.* **34**, 5621–5626 (2014).

Mays, L. E., & Sparks, D. L. Dissociation of visual and saccade-related responses in superior colliculus neurons. *J. of Neurophysiol.* **43**, 207–232 (1980).

Mazor, O. & Laurent, G. Transient dynamics versus fixed points in odor representations by locust antennal lobe projection neurons. *Neuron* **48**, 661–673 (2005).

McCulloch, W.S., Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys*, **5**, 115-133 (1943).

Meister ML, Hennig JA, Huk AC: Signal multiplexing and single neuron computations in lateral intraparietal area during decision-making. *J. Neurosci.* **33**, 2254-2267 (2013).

Mejias, J. F., Wang, X-J. Mechanisms of distributed working memory in a large-scale model of the macaque neocortex. *bioRxiv* 760231 (2019).

Mendoza-Halliday, D., and Martinez-Trujillo, J. C. Neuronal population coding of perceived and memorized visual features in the lateral prefrontal cortex. *Nat. Commun.* **8**, doi:10.1038 (2017).

Messinger, A., Lebedev, M. A., Kralik, J. D., Wise, S. P. Multitasking of attention and memory functions in the primate prefrontal cortex. *J. Neurosci*. **29**, 5640–5653 (2009).

Meyer, T., Qi, X. L., and Constantinidis, C. Persistent discharges in the prefrontal cortex of monkeys naive to working memory tasks. *Cereb. Cortex* **17**, 70–76. (2007).

Meyers, E. M., Freedman, D. J., Kreiman, G., Miller, E. K., Poggio, T. Dynamic population coding of category information in inferior temporal and prefrontal cortex. *J. of Neurophysiol.* **100**, 1407–19 (2008).

Middleton, J. W., Omar, C., Doiron, B. & Simons, D. J. Neural Correlation Is Stimulus Modulated by Feedforward Inhibitory Circuitry. *J. Neurosci.* **32**, 506–518 (2012).

Miller, E. K. The prefrontal cortex and cognitive control. *Nature Rev. Neurosci.* **1**, 59–65 (2000).

Miller, E. K., Li, L., Desimone R. Activity of neurons in anterior inferior temporal cortex during a short-term memory task. *J. Neurosci.* 13, 1460–1478 (1993).

Miller, E. K., Desimone, R. Parallel neuronal mechanisms for short- term memory. *Science* **263**, 520–522 (1994).

Miller, E. K. & J. D. Cohen. Anintegrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**: 167–202 (2001).

Miller, E. K., Erickson, C. A., Desimone, R. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *J. Neurosci.* **16**, 5154–67 (1996).

Milner, B. Effects of different brain lesions on card sorting: the role of the frontal lobes. *Arch. Neurol.* **9**, 90–100, (1963).

Miri, A., Warriner, C. L., Seely, J. S., Elsayed, G. F., Cunningham, J. P., Churchland, M. M., Jessell, T. M. Behaviorally selective engagement of short-latency effector pathways by motor cortex. *Neuron* **95**:683-696 (2017).

Mitchell, J. F., Sundberg, K. A. & Reynolds, J. H. Spatial attention decorrelates intrinsic activity fluctuations in Macaque area V4. *Neuron* **63**, 879–888 (2009).

Mongillo, G., Barak, O., Tsodyks, M. Synaptic theory of working memory. *Science* **319**, 1543–46 (2008).

Moore, T. L., Schettler, S. P., Killiany, R. J., Rosene, D. L., and Moss, M. B. Effects on executive function following damage to the prefrontal cortex in the rhesus monkey (*Macaca mulatta*). *Behav. Neurosci.* **123**, 231–241 (2009).

Murray, J. D., Bernacchia, A., Roy, N. A., Constantinidis, C., Romo, R., Wang, X. J. Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *PNAS* **114**, 394–99 (2017).

Murray, J. D., Bernacchia, A., Freedman, D. J., Romo, R., Wallis, J. D., Cai, X., Padoa-Schioppa, C., Wang, X. J. A Hierarchy of intrinsic timescales across primate cortex. *Nat. Neurosci.* **17**, 1661–63 (2014).

Nakamura, K., & Colby, C. L. Visual, saccade-related, and cognitive activation of single neurons in monkey extrastriate area V3A. *J. of Neurophysiol.* **84**, 677–692 (2000).

Nawrot, M. & Blake, R. On the perceptual identity of dynamic stereopsis and kinetic depth. *Vision Res.* **33**, 1561–1571 (1993).

Naya, Y., Sakai, K., & Miyashita, Y. Activity of primate inferotemporal neurons related to a sought target in pair-association task. *PNAS*, **7**, 2664–2669 (1996).

Nelder, J. A. & Mead, R. A Simplex Method for Function Minimization. *Comput. J.* **7**, 308–313 (1965).

Newsome, W. T., Britten, K. H. & Movshon, J. A. Neuronal correlates of a perceptual decision. *Nature* **341**, 52–54 (1989).

Ni, A. M., Ruff, D. A., Alberts, J. J., Symmonds, J. & Cohen, M. R. Learning and attention reveal a general relationship between population activity and behavior. *Science* **359**, 463–465 (2018).

Nieder, A. & Miller, E. K. Coding of cognitive magnitude: compressed scaling of numerical information in the primate prefrontal cortex. *Neuron* **37**, 149–157 (2003).

Nienborg, H. & Cumming, B. G. Decision-related activity in sensory neurons reflects more than a neuron's causal effect. *Nature* **459**, 89–92 (2009).

Nienborg, H., Cohen, M. R. & Cumming, B. G. Decision-Related Activity in Sensory Neurons: Correlations Among Neurons and with Behavior. *Annu. Rev. Neurosci.* **35**, 463–483 (2012).

Nijboer, M., Borst, J., van Rijn, H., Taatgen, N. Single-task fMRI overlap predicts concurrent multitasking interference. *Neuroimage* **100**, 60–74 (2014).

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

Ninomiya, T., Dougherty, K., Godlove, D. C., Schall, J.D., Maier, A. Microcircuitry of agranular frontal cortex: Contrasting laminar connectivity between occipital and frontal areas. *J Neurophysiol* **113***,* 3242–3255 *(*2015*).*

Nishida, S., Tanaka, T., Shibata, T., Ikeda, K., Aso, T., Ogawa, T. Discharge-rate persistence of baseline activity during fixation reflects maintenance of memory-period activity in the macaque posterior parietal cortex. *Cerebr. Cortex* **24**, 1671–85 (2014).

Nobre, A. C. and Stokes, M. G. Premembering experience: A hierarchy of timescales for pro-active attention. *Neuron* **104,** 132-146 (2019).

Ó Scalaidhe, S. P., Wilson, F. A., and Goldman-Rakic, P. S. Areal segregation of face-processing neurons in prefrontal cortex. *Science* **278**, 1135–1138 (1997).

Ó Scalaidhe, S. P., Wilson, F. A., and Goldman-Rakic, P. S. Face-selective neurons during passive viewing and working memory performance of rhesus monkeys: evidence for intrinsic specialization of neuronal coding. *Cereb. Cortex* **9**, 459–475 (1999).

Ogawa, T., and Komatsu, H. Differential temporal storage capacity in the baseline activity of neurons in macaque frontal eye field and Area V4. *J. of Neurophysiol.* **103**, 2433–45 (2010).

Olejnik, S. & Algina, J. Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychol. Methods* **8**, 434–447 (2003).

Orhan, A. E. & Ma, W. J. A diverse range of factors affect the nature of neural representations underlying short-term memory. *Nat. Neurosci.* **22**, 275–283 (2019).

Owen, A. M., Downes, J. J., Sahakian, B. J., Polkey, C. E., and Robbins, T. W. Planning and spatial working memory following frontal lobe lesions in man. *Neuropsychologia* **28**, 1021–1034 (1990).

Pagan M, Urban LS, Wohl MP, Rust NC: Signals in inferotemporal and perirhinal cortex suggest an untangling of visual target information. *Nat. Neurosci.* **16**, 1132-1139 (2013).

Pan, X., Sawa K., Tsuda I., Tsukada M., Sakagami M. Reward prediction based on stimulus categorization in primate lateral prefrontal cortex. *Nat. Neurosci.* **11**, 703–712 (2008).

Pandya DN, Barnes CL. 1987. Architecture and connections of the frontal lobe. *In The Frontal Lobes Revisited*, (Ed. E Perecman, E.) 41–72. (New York, 1987).

Pang, K. C. H., Yoder, R. M., Olton, D. S. Neurons in the lateral agranular frontal cortex have divided attention correlates in a simultaneous temporal processing task. *Neuroscience* **103**, 615–628 (2001).

Park, I. M., Meister, M. L. R., Huk, A. C., & Pillow, J. W. Encoding and decoding in parietal cortex during sensorimotor decision-making. *Nat. Neurosci. **17***, 1395-1403 (2014)

Parker, A. J. & Newsome, W. T. Sense and the single Neuron: Probing the Physiology of Perception. *Annu. Rev. Neurosci.* **21**, 227–277 (1998).

Parker, A. J. A micro-pool model for decision-related signals in visual cortical areas. *Front. Comput. Neurosci.* **7**, 115 (2013).

Parthasarathy, A. et al. Mixed selectivity morphs population codes in prefrontal cortex. *Nat. Neurosci.* **20**, 1770–1779 (2017).

Parthasarathy, A. *et al.* Time-Invariant Working Memory Representations in the Presence of Code-Morphing in the Lateral Prefrontal Cortex. *bioRxiv* 563668 (2019).

Passingham, R.E. Attention to action. *Phil. Trans. R. Soc. Lond. B* **351**, 1473–1479 (1996).

Passingham, R. E., Wise, S. P. The Neurobiology of the Prefrontal Cortex: Anatomy, Evolution, and the Origin of Insight. *Oxford University Press*, Oxford, UK. (2012).

Passingham, R. E. Memory of monkeys (Macaca mulatta) with lesions in prefrontal cortex. *Behavioral Neuroscience* **99**, 3–21 (1985).

Pastalkova, E., Itskov, V., Amarasingham, A., Buzsaki, G. Internally generated cell assembly sequences in the rat hippocampus. *Science* **321**, 1322–27 (2008).

Paus, T., Zijdenbos, A., Worsley, K., Collins, D. L., Blumenthal, J., Giedd, J. N., Rapoport, J. L., and Evans, A. C. Structural maturation of neural pathways in children

and adolescents: In vivo study. *Science* **283**, 1908–1911 (1999).

Peng, X., Sereno, M. E., Silva, A. K., Lehky, S. R., and Sereno, A. B. Shape selectivity in primate frontal eye field. *J. Neurophysiol.* **100**, 796–814 (2008).

Pernice, V., Staude, B., Cardanobile, S. & Rotter, S. How Structure Determines Correlations in Neuronal Networks. *PLOS Comput. Biol.* **7**, e1002059 (2011).

Perkel, D. H., Gerstein, G. L. & Moore, G. P. Neuronal spike trains and stochastic point processes. II. Simultaneous spike trains. *Biophys. J.* **7**, 419–440 (1967).

Perret E. The left frontal lobe of man and the suppression of habitual responses in verbal categorical behaviour. *Neuropsychologia* **12**, 323–30 (1974).

Pinto, L., and Dan, Y. Cell-type-specific activity in prefrontal cortex during goal-directed behavior. *Neuron* **87**, 437–50 (2015).

Pitkow, X., Liu, S., Angelaki, D. E., DeAngelis, G. C. & Pouget, A. How Can Single Sensory Neurons Predict Behavior? *Neuron* **87**, 411–423 (2015).

Postle, B. R. Working memory as an emergent property of the mind and brain. *Neuroscience* **139**, 23-38 (2006).

Prince, S. J. D., Pointon, A. D., Cumming, B. G. & Parker, A. J. Quantitative Analysis of the Responses of V1 Neurons to Horizontal Disparity in Dynamic Random-Dot Stereograms. *J. Neurophysiol.* **87**, 191–208 (2002).

Qi, X.-L., Katsuki, F., Meyer, T., Rawley, J. B., Zhou, X., Douglas, K. L., and Constantinidis, C. Comparison of neural activity related to working memory in primate dorsolateral prefrontal and posterior parietal cortex. *Front. Syst. Neurosci*. **4**, 12 (2010).

Quintana, J. & Fuster, J. M. Mnemonic and predictive functions of cortical neurons in a memory task. *Neuroreport* **3**, 721–724 (1992).

Rainer, G., Asaad, W. F., Miller, E. K. Memory fields of neurons in the primate prefrontal cortex. *PNAS* **95**, 15008–13 (1998).

Rainer, G., Rao, S. C. & Miller, E. K. Prospective coding for objects in primate prefrontal cortex. *J. Neurosci.* **19**, 5493–5505 (1999).

Rao, S. C., Rainer, G., Miller, E. K. Integration of what and where in the primate prefrontal cortex. Science **276**, 821–824 (1997).

Raposo, D., Kaufman, M. T., Churchland, A. K. A Category-free neural population supports evolving demands during decision-making. *Nat. Neurosci.* **17**, 1784–92 (2014).

Remington, E. D., Narain, D., Hosseini, E. A. & Jazayeri, M. Flexible sensorimotor computations through rapid reconfiguration of cortical dynamics. *Neuron* **98**, 1005–1019 (2018).

Rigotti, M., Barak, O., Warden, M. R., Wang, X. J., Daw, N. D., Miller, E. K., Fusi, S. The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–90 (2013).

Riley, M. R., Qi, X. L., and Constantinidis, C. Functional specialization of areas along the anterior-posterior axis of the primate prefrontal cortex. *Cereb. Cortex* **27**, 3683–3697 (2017).

Rishel, C. A., Huang, G. & Freedman, D. J. Independent category and spatial encoding in parietal cortex. *Neuron* **77**, 969-979 (2013).

Roca, M., Torralva, T., Gleichgerrcht, E., Woolgar, A., Thompson, R., Duncan, J., and Manes, F. The role of Area 10 (BA10) in human multitasking and in social cognition: a lesion study. *Neuropsychologia* **49**, 3525–3531 (2011).

Roelfsema, P. R. & Spekreijse, H. The Representation of Erroneously Perceived Stimuli in the Primary Visual Cortex. *Neuron* **31**, 853–863 (2001).

Romo, R., Brody, C. D., Hernández, A., Lemus, L. Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature* **399**, 470–73 (1999).

Rose S. N., La Rocque, J. J., Riggall, A. C., Gosseries, O., Starrett, M. J., Meyering, E. E., Postle, B. R. Reactivation of latent working memories with transcranial magnetic stimulation. *Science* **354**, 1136-1139 (2016).

Rosenbaum, R., Smith, M. A., Kohn, A., Rubin, J. E. & Doiron, B. The spatial structure of correlated neuronal variability. *Nat. Neurosci.* **20**, 107-114 (2016).

Rosenbaum, R., Trousdale, J. & Josić, K. The Effects of Pooling on Spike Train Correlations. *Front. Neurosci.* **5**, (2011).

Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **65**, 386-408 (1958)

Roy, J. P., Komatsu, H. & Wurtz, R. H. Disparity sensitivity of neurons in monkey extrastriate area MST. *J. Neurosci.* **12**, 2478–2492 (1992).

Roy, J. E., Riesenhuber, M., Poggio, T.& Miller, E. K. Prefrontal cortex activity during
flexible categorization. *J. Neurosci.* **30**, 8519–8528 (2010).

Rudebeck, P. H., Behrens, T. E., Kennerley, S. W., Baxter, M. G., Buckley, M. J., Walton, M. E., and Rushworth, M. F. Frontal cortex subregions play distinct roles in choices between actions and stimuli. *J. Neurosci.* **28**, 13775–13785 (2008).

Ruff, D. A. & Cohen, M. R. Attention can increase or decrease spike count correlations between pairs of neurons depending on their role in a task. *Nat. Neurosci.* **17**, 1591–1597 (2014).

Runyan, C. A., Piasini, E., Panzeri, S., Harvey, C. D. Distinct timescales of population coding across cortex. *Nature* 548, 92–96 (2017).

Sakai, K., Rowe, J. B., Passingham, R. E. Active maintenance in prefrontal area 46 creates distractor-resistant memory. *Nat. Neurosci*. **5**, 479–484 (2002).

Salinas, E., Hernández, A., Zainos, A., Romo, R. Periodicity and firing rate as candidate neural codes for the frequency of vibrotactile stimuli. *J. Neurosci.* **20**, 5503–15 (2000).

Salzman, C. D., Britten, K. H. & Newsome, W. T. Cortical microstimulation influences perceptual judgements of motion direction. *Nature* **346**, 174–177 (1990).

Sarma, A., Masse, N. Y., Wang, X.-J. & Freedman, D. J. *Task-specific versus generalized mnemonic representations in parietal and prefrontal cortices*. *Nature Neurosci.* **19**, 143–149 (2016).

Saxena, S., Cunningham, J. P. Towards the neural population doctrine. *Curr Opin Neurobiol*, **55**, 103-111 (2019).

Schmahmann, J. D., Pandya, D. N. Anatomic organization of the basilar pontine projections from prefrontal cortices in rhesus monkey. *J. Neurosci.* **17**, 438–58 (1997).

Scott, B. B, Constantinople, C. M., Akrami, A., Hanks, T. D., Brody, C. D., Tank, D. W. Fronto-parietal cortical circuits encode accumulated evidence with a diversity of timescales. *Neuron* **95**, 385–398 (2017).

Selemon, L. D., and Goldman-Rakic, P. S. Common cortical and subcortical targets of the dorsolateral prefrontal and posterior parietal cortices in the rhesus monkey: evidence for a distributed neural network subserving spatially guided behavior. *J. Neurosci.* **8**, 4049–4068 (1988).

Shadlen, M. N., Britten, K. H., Newsome, W. T. & Movshon, J. A. A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *J. Neurosci.* **16**, 1486–1510 (1996).

Shadlen, M. N. & Newsome, W. T. The Variable Discharge of Cortical Neurons: Implications for Connectivity, Computation, and Information Coding. *J. Neurosci.* **18**, 3870–3896 (1998).

Shafi, M., Zhou, Y., Quintana, J., Chow, C., Fuster, J., Bodner, M. Variability in neuronal activity in primate cortex during working memory tasks. *Neurosci.* **146**, 1082–1108 (2007).

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

Shallice, T. Specific impairments of planning. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **298**, 199–209 (1982).

Shepherd, G. M. Foundations of the Neuron Doctrine. (Oxford Univ. Press, 1991).

Shima, K., Isoda, M., Mushiake, H., and Tanji, J. Categorization of behavioural sequences in the prefrontal cortex. *Nature* **445**, 315–318 (2007).

Shiozaki, H. M., Tanabe, S., Doi, T. & Fujita, I. Neural Activity in Cortical Area V4 Underlies Fine Disparity Discrimination. *J. Neurosci.* **32**, 3830–3841 (2012).

Sigala, N., Kusunoki, M., Nimmo-Smith, I., Gaffan, D., Duncan, J. Hierarchical Coding for Sequential Task Events in the Monkey Prefrontal Cortex. *PNAS* **105**, 11969–74 (2008).

Singh, R., & Eliasmith, C. Higher-dimensional neurons explain the tuning and dynamics of working memory cells. *J. Neurosci.* **26**, 3667–3678 (2006).

Smith, M. A. & Kohn, A. Spatial and Temporal Scales of Neuronal Correlation in Primary Visual Cortex. *J. Neurosci.* **28**, 12591–12603 (2008).

Sohn, H., Narain, D., Meirhaeghe, N. & Jazayeri, M. Bayesian computation through cortical latent dynamics. *Neuron* **103**, 934–947 (2019).

Solomon, S. S., Chen, S. C., Morley, J. W. & Solomon, S. G. Local and Global Correlations between Neurons in the Middle Temporal Area of Primate Visual Cortex. *Cereb. Cortex* **25**, 3182–3196 (2015).

Sowell, E. R., Thompson, P. M., Holmes, C. J., Jernigan, T. L., Toga, A.W. In vivo evidence for post-adolescent brain maturation in frontal and striatal regions. *Nat. Neurosci*. **2**, 859–861 (1999).

Spaak, E., Watanabe, E., Funahashi, S., Stokes, M. G. Stable and dynamic coding for working memory in primate prefrontal cortex. *J. of Neurosci.* **37**, 6503–16 (2017).

Stamm, J. S. Electrical stimulation of monkeys' prefrontal cortex during delayed-response performance. *J. Comp. Physiol. Psychol.* **67**, 535–46 (1969).

Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., Duncan, J. Dynamic coding for cognitive control in prefrontal cortex. *Neuron* **78**, 364–75 (2013).

Stokes, M. G. 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework. *Trends Cogn. Sci.* **19**, 394–405 (2015).

Stokes, M. G. *J. Neurosci.* **31**, 1167–69 (2011).

Stokes, M. G., Buschman, T. J., Miller, E. K., Dynamic coding for flexible cognitive control. *In The Wiley handbook of cognitive control*. (Ed. Egner, T.) 221-241 (Wiley-Blackwell, Oxford, 2017).

Stroop, J. R. Studies of interference in serial verbal reactions. *J. Exp. Psychol.* **18**, 643–62 (1935).

Super, H. A. Neural correlate of working memory in the monkey primary visual cortex. *Science* **293**, 120–124 (2001).

Suzuki, M., and Gottlieb, J. Distinct neural mechanisms of distractor suppression in the frontal and parietal lobe. *Nat. Neurosci*. **16**, 98–104 (2013).

Swaminathan, S.K. & Freedman, D.J. *Preferential encoding of visual categories in parietal cortex compared with prefrontal cortex*. *Nat. Neurosci.* **15**, 315–320 (2012)

Szatmáry, B., and Izhikevich, E. M. Spike-timing theory of working memory. *PLOS Comput. Biol.* **6**, e1000879 (2010).

Szczepanski, S. M., Knight, R. T. Insights into Human Behavior from Lesions to the Prefrontal Cortex. *Neuron* **83**, 1002-1018 (2014).

Tang, H., Qi, X.-L., Riley, M. R., Constantinidis, C. Working memory capacity is enhanced by distributed prefrontal activation and invariant temporal dynamics, *PNAS* (2019).

Tang, C., Herikstad, R., Parthasarathy, A., Libedinsky, C., Yen, S-C. Independent Activity Subspaces for Working Memory and Motor Preparation in the Lateral Prefrontal Cortex. *bioRxiv* 756072 (2019).

Tang, E., Mattar, M. G., Giusti, C. et al. Effective learning is accompanied by high-dimensional and efficient representations of neural activity. *Nat. Neurosci.* **22**, 1000 (2019).

Tolhurst, D. J., Movshon, J. A. & Thompson, I. D. The dependence of response amplitude and variance of cat visual cortical neurones on stimulus contrast. *Exp. Brain Res.* **41**, 414–419 (1981).

Townsend, J., Koep, N., Weichwald, S. Pymanopt: A Python Toolbox for Optimization over Manifolds using Automatic Differentiation. *J. of Mach. Learning Research.* **17**, 1-5 (2016).

Trübutschek, D., Marti, S., Ueberschär, H. & Dehaene, S. Probing the limits of activity-silent non-conscious working memory. *PNAS* **116**, 14358–14367 (2019).

Tsutsui, K., Hosokawa, T., Yamada, M. & Iijima, T. Representation of Functional Category in the Monkey Prefrontal Cortex and Its Rule-Dependent Use for Behavioral Selection. *J Neurosci* **36**, 3038–3048 (2016).

Uka, T. & DeAngelis, G. C. Contribution of Area MT to Stereoscopic Depth Perception: Choice-Related Response Modulations Reflect Task Strategy. *Neuron* **42**, 297–310 (2004).

Uka, T., Tanabe, S., Watanabe, M. & Fujita, I. Neural Correlates of Fine Depth Discrimination in Monkey Inferior Temporal Cortex. *J. Neurosci.* **25**, 10796–10802 (2005).

Vaadia, E., Benson, D. A., Hienz, R. D. & Goldstein, M. H. Jr. Unit study of monkey frontal cortex: active localization of auditory and of visual stimuli. *J. Neurophysiol*. **56**, 934–952
(1986).

Van Hoesen, G. W., Pandya, D. N., Butters, N. Cortical afferents to the entorhinal cortex of the rhesus monkey. *Science* **175**, 1471–73 (1972).

Vapnick, V. N. Statistical learning theory (Wiley, New York, 1998).

Vendrell, P., Junque, C., Pujol, J., Jurado, M. A., Molet, J., Grafman, J. The role of prefrontal regions in the Stroop task. Neuropsychologia **33**, 341–52 (1995).

Verhoef, B.-E., Vogels, R. & Janssen, P. Contribution of Inferior Temporal and Posterior Parietal Activity to Three-Dimensional Shape Perception. *Curr. Biol.* **20**, 909–913 (2010).

Walker, S. C., Robbins, T. W., and Roberts, A. C. Response disengagement on a spatial self-ordered sequencing task: effects of regionally selective excitotoxic lesions and serotonin depletion within the prefrontal cortex. *J. Neurosci.* **29**, 6033–6041 (2009).

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

Wang, B., Yin, L., Zou, X. et al. A subtype of inhibitory interneuron with intrinsic persistent activity in human and monkey neocortex. *Cell Rep*. **10**, 1450–1458 (2015).

Wang, X. J. The Prefrontal Cortex as a quintessential "cognitive-type" neural circuit. *In Principles of Frontal Lobe Function*, edited by Stuss D. T. and Knight, R. T. (Oxford University Press, Oxford, 2013).

Wang, X. J. Synaptic Reverberation underlying mnemonic persistent activity. *Trends Neurosci.* **24**, 455–63 (2001).

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

Wasmuht, D. F., Spaak, E., Buschman, T. J., Miller, E. K., Stokes, M. G. Intrinsic neuronal dynamics predict distinct functional roles during working memory. *Nat. Commun*. **9**: 3499 (2018).

Watanabe, K., Funahashi, S. Neuronal correlates of cognitive resource allocation revealed by a dual-task paradigm. *Neurosci. Res.* **71**, e382 (2011).

Watanabe, K., Funahashi, S. Neural mechanisms of dual-task interference and

cognitive capacity limitation in the prefrontal cortex. *Nat. Neurosci.* **17**, 601–611 (2014).

Watanabe, M. Prefrontal unit activity during delayed conditional go/no-go discrimination in the monkey. I. Relation to the stimulus. *Brain Res*. **382**, 1–14 (1986).

Watanabe, M. Prefrontal unit activity during associative learning in the monkey. *Exp. Brain Res*. **80**, 296–309 (1990).

Watanabe, M. Frontal units of the monkey coding the associative significance of visual and auditory stimuli. Exp. Brain Res. **89**, 233–247 (1992).

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

Weilnhammer, V. A., Ludwig, K., Hesselmann, G. & Sterzer, P. Frontoparietal Cortex Mediates Perceptual Transitions in Bistable Perception. *J. Neurosci.* **33**, 16009–16015 (2013).

White, I. M. & Wise, S. P. Rule-dependent neuronal activity in the prefrontal cortex. *Exp. Brain Res.* **126**, 315–335 (1999).

Wickens, C.D. The structure of attentional resources. in *Attention and Performance VIII* (ed. Nickerson, R. S.) 239-257 (Lawrence Erlbaum, Hillsdale, New Jersey, 1980)

Wickens, C.D. Multiple resources and performance prediction. *Theor. Issues Ergon.* **3**, 159-177 (2002).

Wilson, C. R., Gaffan, D., Browning, P. G., and Baxter, M. G. Functional localization within the prefrontal cortex: missing the forest for the trees? *Trends Neurosci*. **33**, 533–540 (2010).

Wilson, F. A., Scalaidhe, S. P. and Goldman-Rakic, P. S. Dissociation of object and spatial processing domains in primate prefrontal cortex. *Science* 260, 1955-1958 (1993).

Wimmer, K., Nykamp, D.Q., Constantinidis, C. et al. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nat. Neurosci*. **17**, 431–439 (2014).

Wimmer, K. *et al.* Sensory integration dynamics in a hierarchical network explains choice probabilities in cortical area MT. *Nat. Commun.* **6**, (2015).

Wong, K. F., and Wang, X. J. A Recurrent network mechanism of time integration in perceptual decisions. *J. of Neurosci.* **26**, 1314–28 (2006).

Wutz, A., Loonis, R., Roy, J. E., Donoghue, J. A. & Miller, E. K. Different levels of category abstraction by different dynamics in different prefrontal areas. *Neuron* **97**, 716–726 (2018).

Xie, J. & Padoa-Schioppa, C. Neural remapping and circuit persistence in economic decisions. *Nat. Neurosci.* **19**, 855–861 (2016).

Yeterian, E. H., Pandya, D. N., Tomaiuolo, F., and Petrides, M. The cortical connectivity of the prefrontal cortex in the monkey brain. *Cortex* **48**, 58–81 (2012).

Yuste, R. From the neuron doctrine to neural networks. *Nat Rev Neurosci*. **16**, 487-497 (2015).

Zaidel, A., DeAngelis, G. C. & Angelaki, D. E. Decoupled choice-driven and stimulus-related activity in parietal neurons may be misrepresented by choice probabilities. *Nat. Commun.* **8**, 715 (2017).

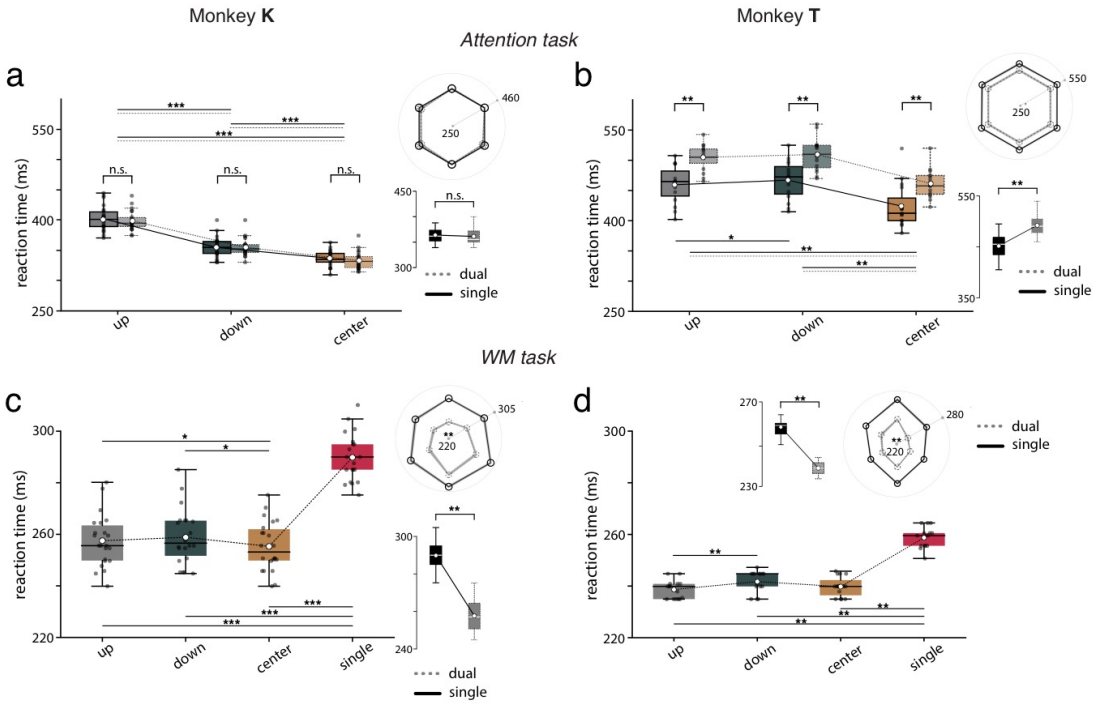
Zaksas, D. & Pasternak, T. Directional signals in the prefrontal cortex and in area MT during a working memory for visual motion task. *J. Neurosci.* **26**, 11726–11742 (2006).

Zhang, J. & Abbott, L. F. Gain modulation of recurrent networks. *Neurocomputing* **32-33,** 623-628 (2000).

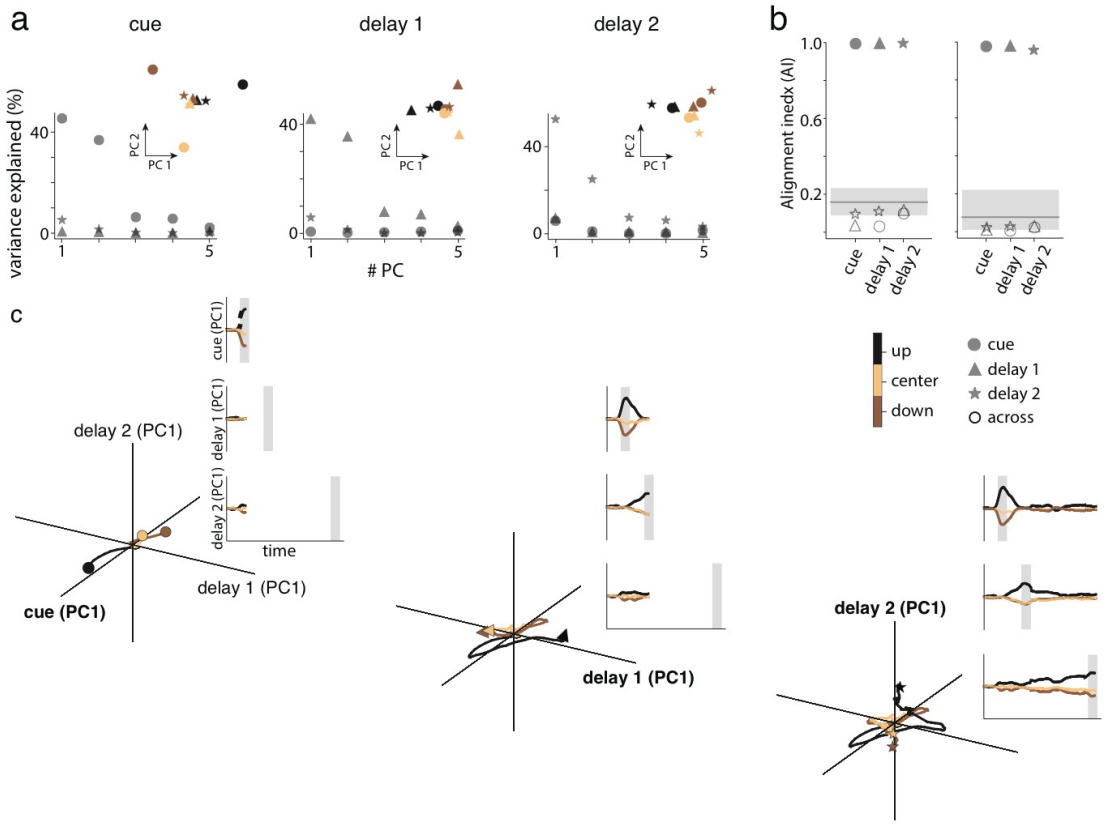
Zohary, E., Shadlen, M. N. & Newsome, W. T. Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature* **370**, 140–143 (1994).

Zylberberg, J., and Strowbridge, B. W. Mechanisms of persistent activity in Cortical circuits: Possible neural substrates for working memory. *Annu. Rev. Neurosci.* **40**, 603–27 (2017).

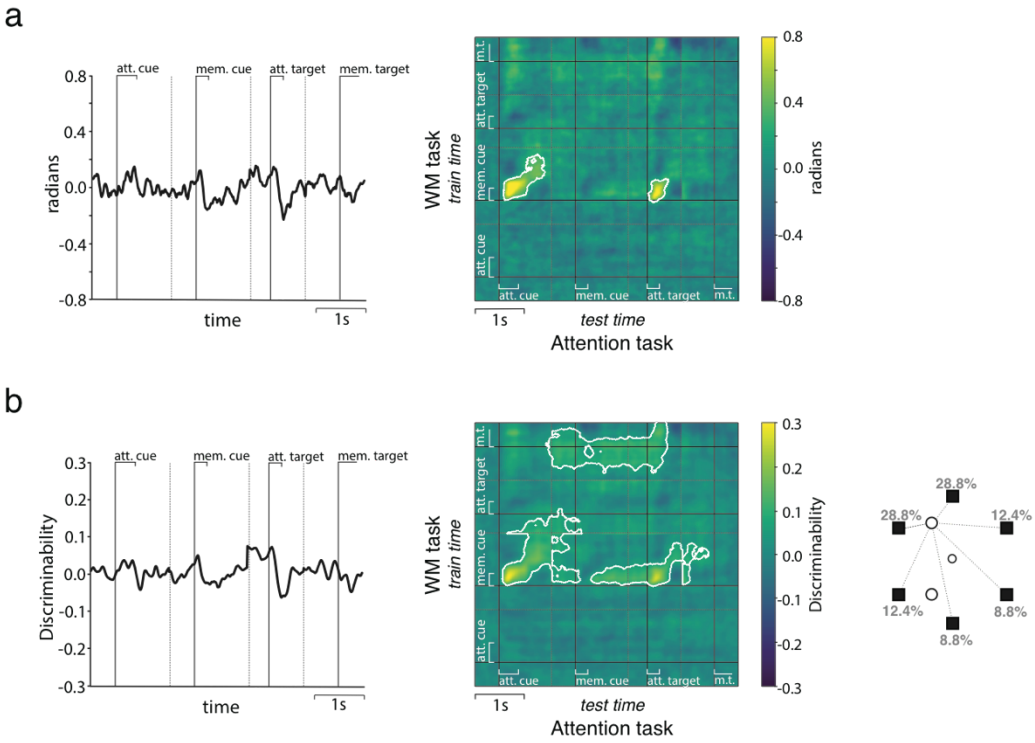
Appendix 1. Supplementary Figures to Chapter 2



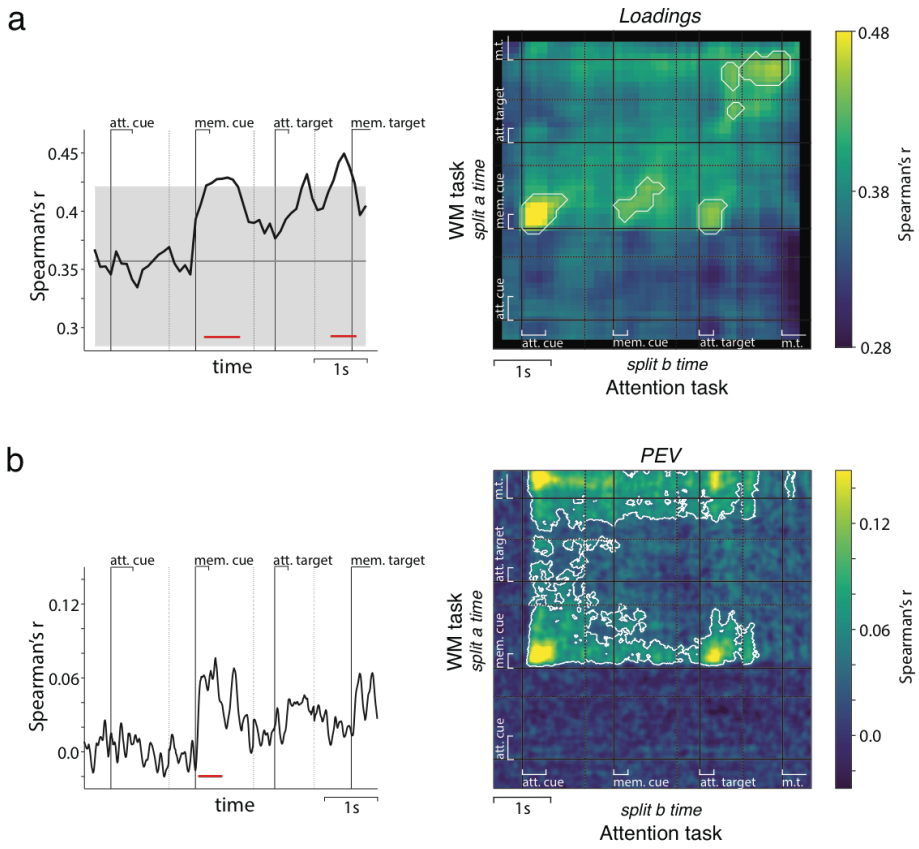
Supplementary Figure 2.1: Behavioural performance – reaction times. a) Boxplots show monkeys' median reaction times (RTs). Throughout the figure dots represent unique sessions (24 for monkey K; 15 for monkey T). Whiskers extend to the upper and lower quartiles of the respective distributions. Black or white lines within boxes indicate the median and white dots indicate the mean of the respective distributions. * $p<0.05$; ** $p<0.01$; *** $p<0.001$ (Wilcoxon signed-rank test; corrected for multiple comparisons by Holm's sequentially rejective Bonferroni procedure). **a) & b)** RTs for the dual attention task (light colours & dashed) and single attention task (dark colours & solid) as a function of attention cue location (up, down & center). Box plot insets compare overall RTs for the dual (grey; dashed) and single (black; solid) attention tasks, irrespective of cue locations. Hexagonal plot insets compare RTs for the dual attention task as a function of WM cue location (dashed lines). Note, solid lines indicate the RT for the single attention task, which is constant since there was no concurrent WM component, but displayed for comparison (shaded areas: s.e.m.). **c & d)** RTs for the dual WM task as a function of the concurrent attention cue conditions (up, down & center) and when performed on its own (single). Box plot insets compare overall RTs for the dual and single WM tasks, irrespective of cue locations. Hexagonal plot insets compare RTs for the dual (grey; dashed) and single WM tasks as a function of WM cue location (black; solid) (shaded area: s.e.m.).



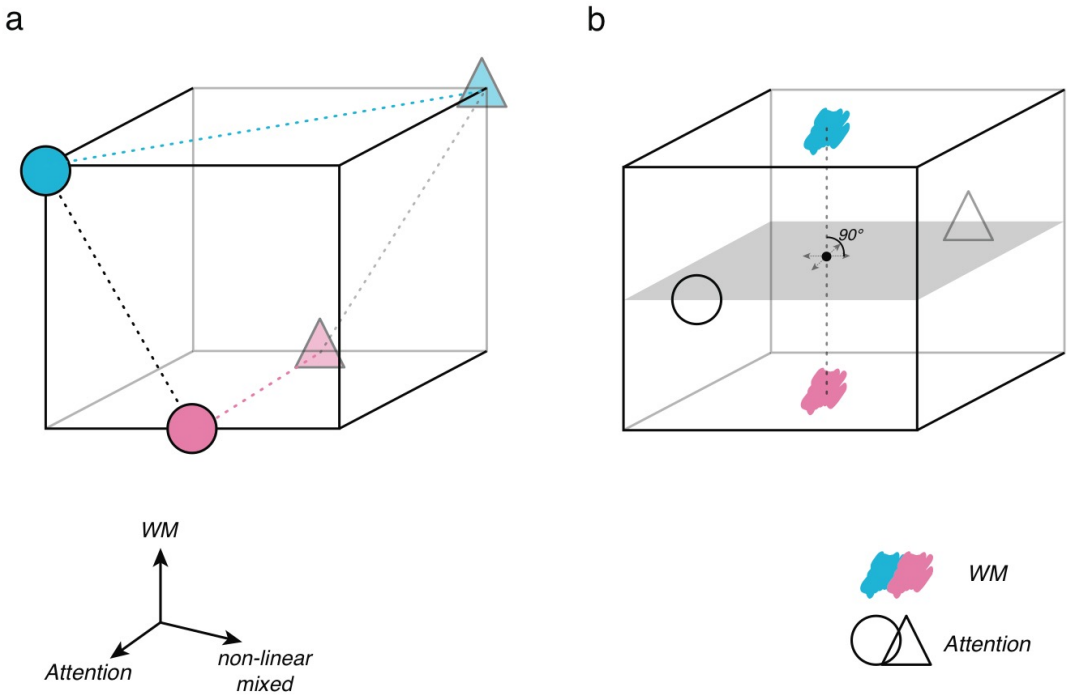
Supplementary Figure 2.2: Subspace analysis – Attention task. a) Condition averaged activity from three distinct task epochs was projected onto the top 5 principal components (PCs) i.e. 5 dimensional subspace computed at each task epoch. The variance in task epoch specific activity explained by each PC is displayed. Cue epoch (0.1s-0.4s after attention cue presentation; circle); delay 1 epoch (0.8ms-1.1ms after attention cue presentation; triangle); delay 2 epoch (-0.3s to 0.1s before attention target onset; star). Insets show condition averaged activity from each task epoch projected onto the top 2 PCs of each task epoch. **b**) Symbols indicate the AI (see chapter results and methods) between the subspace defined at one epoch with that defined at any other epoch (empty symbols) or the same epoch (solid symbols). The grey solid line and shaded area indicate the mean and 95% confidence interval of the null distribution of AIs computed between randomly sampled subspaces of the same dimensions (see Methods). The left panel shows the AI for the originally defined 5 dimensional subspaces. The right panel shows the AI for the same set of subspaces but reduced to 3 dimensions, and optimized to be explicitly orthogonal (see Methods) **c**) Condition averaged trajectories in the three dimensional subspace defined by the optimized first PCs for each epoch. Different panels show trajectories up to the cue (left), delay 1 (middle) and delay 2 (right) epochs. Insets show trajectories on the first PC alone for each epoch.



Supplementary Figure 2.3: Across task decoding. a) Across task decoding using a parametric (angular) classifier (see Methods). Left panel: Within time decoding performance, measured as deviation from chance (zero) in radians. Throughout the figure, black vertical lines mark the onset of relevant task events (solid) and splicing boundaries (dashed) (see Methods). Right panel: across-time decoding performance. Colours indicate magnitude of the deviation in radians. White contours indicate significant decoding ($p<0.05$; cluster based permutation test). **b)** Across task decoding using constructed attention locations i.e. fake attention location population responses were constructed as a weighted average of the memory location responses (see right most panel) and tested on the actual location specific responses in the attention task (see Methods). Left panel: Within time decoding measured as pattern discriminability score. Left panel: across-time pattern discriminability score. Colours indicate score magnitude. White contours indicate a significant score ($p<0.05$; cluster based permutation test).

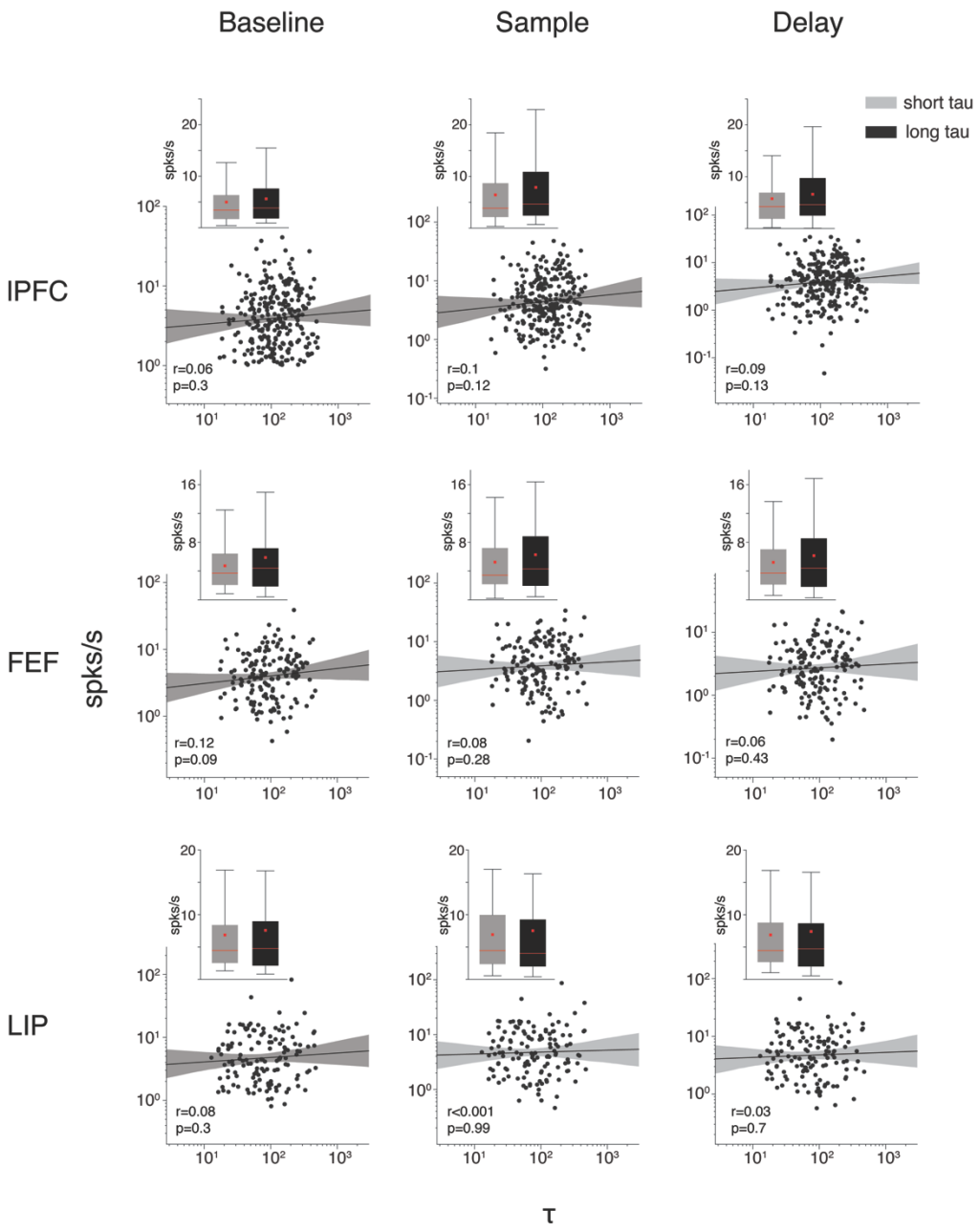


Supplementary Figure 2.4: Subspace loadings. a) Correlation between subspace loadings for the attention and WM tasks. **a)** Correlation between the loadings on subspaces calculated during the same sliding time windows (400ms) i.e. within time correlation. The grey solid line and shaded area indicate the mean and 95% confidence interval of the null distribution of the correlation between loadings on randomly sampled subspaces of the same dimensions as the original subspaces (see Methods). Red solid line indicates were the black line exceeded the null distribution ($p<0.05$). Right panel: Loading correlations for each pair of time windows. Colours indicate the correlation magnitude. White contours indicate a significant correlation ($p<0.05$) based on the null distribution shown in the left panel. **b)** Correlations between PEV values across neurons for both tasks. Left panel: correlation computed for each time point in the trial. Red solid line indicates a significant correlation ($p<0.05$; *cluster based permutation test*). Right panel: Correlation computed for each pair of time points. Colours indicate the magnitude of the correlation. White contours indicate a significant correlation ($p<0.05$; *cluster based permutation test*).

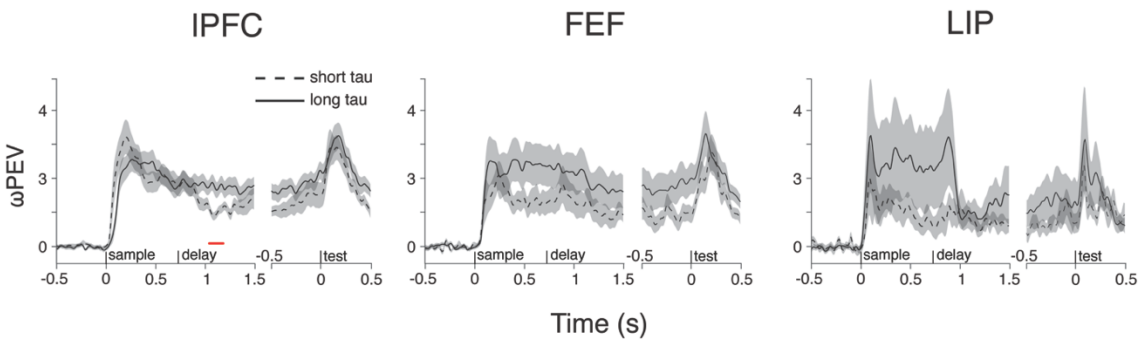


Supplementary Figure 2.5: Exemplary perturbations to the proposed task representation. a) Consider a hypothetical space defined by the responses of three neurons (or PCs). One neuron exclusively codes for attention location, a second neuron exclusively codes for WM location, and a third neuron responds to some locations in no predictable fashion (non-linear mixed). In such a scenario, a linear classifier could perfectly read out any of the displayed location stimuli and could readily distinguish to which task the belong. However, the pattern changes across instances. More specifically, one can not perfectly generalize the vector between the two blue objects (blue dashed line) to the vector between the to pink objects (pink line). That is, the pattern for the attention task changes as a function of the WM task, which can be problematic if, for example, the attention task would rely on a fixed readout. One can see that the same holds true when focusing on the WM task as a function of attention conditions. **b)** Averaging WM over attention conditions (colours) and attention over WM conditions (shapes) demonstrate that, in this specific case, attention and WM task would show no subspace overlap (on average), despite poor cross-instance generalization. Together the schemata in **a** & **b** exemplify how non-linear mixed selectivity could affect across instance generalization without necessarily leading to an overlap in subspaces. Note, this only holds true if subspace overlap is assessed on averages across instances, as was done in Figs. 2.5, 2.6 & 2.7 in the **chapter 3** results.

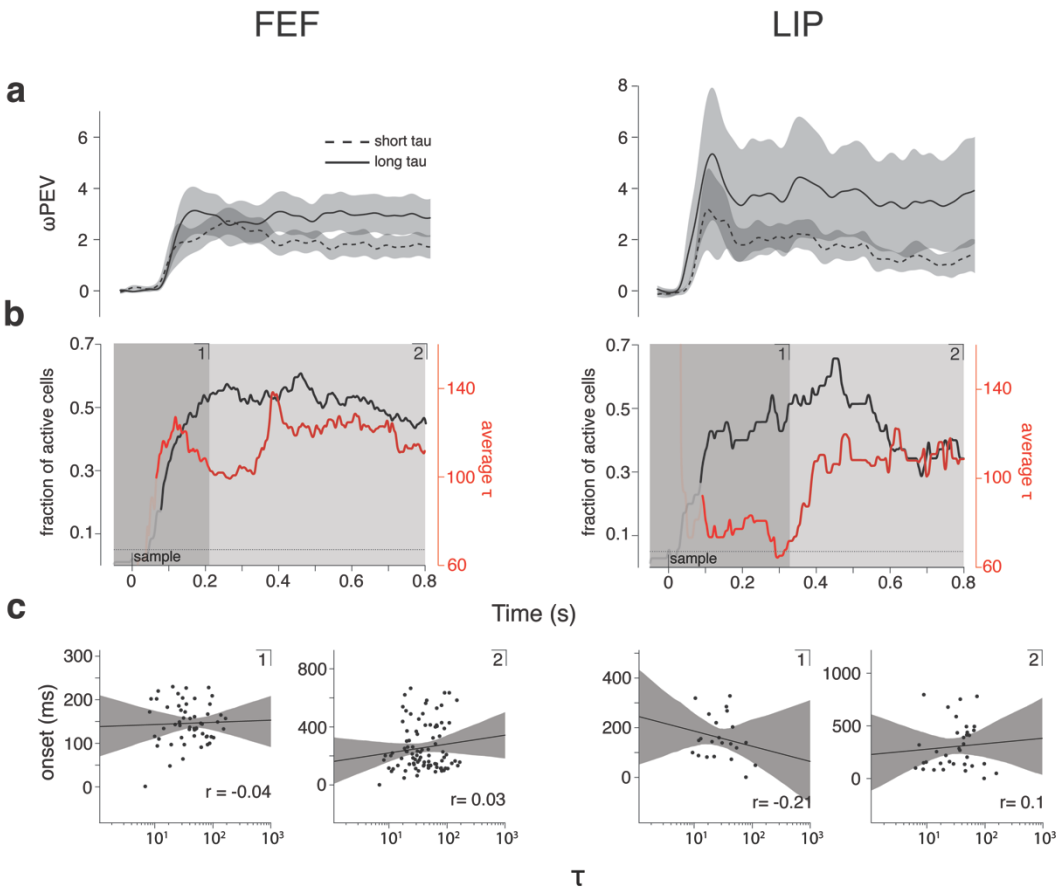
Appendix 2. Supplementary Figures to Chapter 3



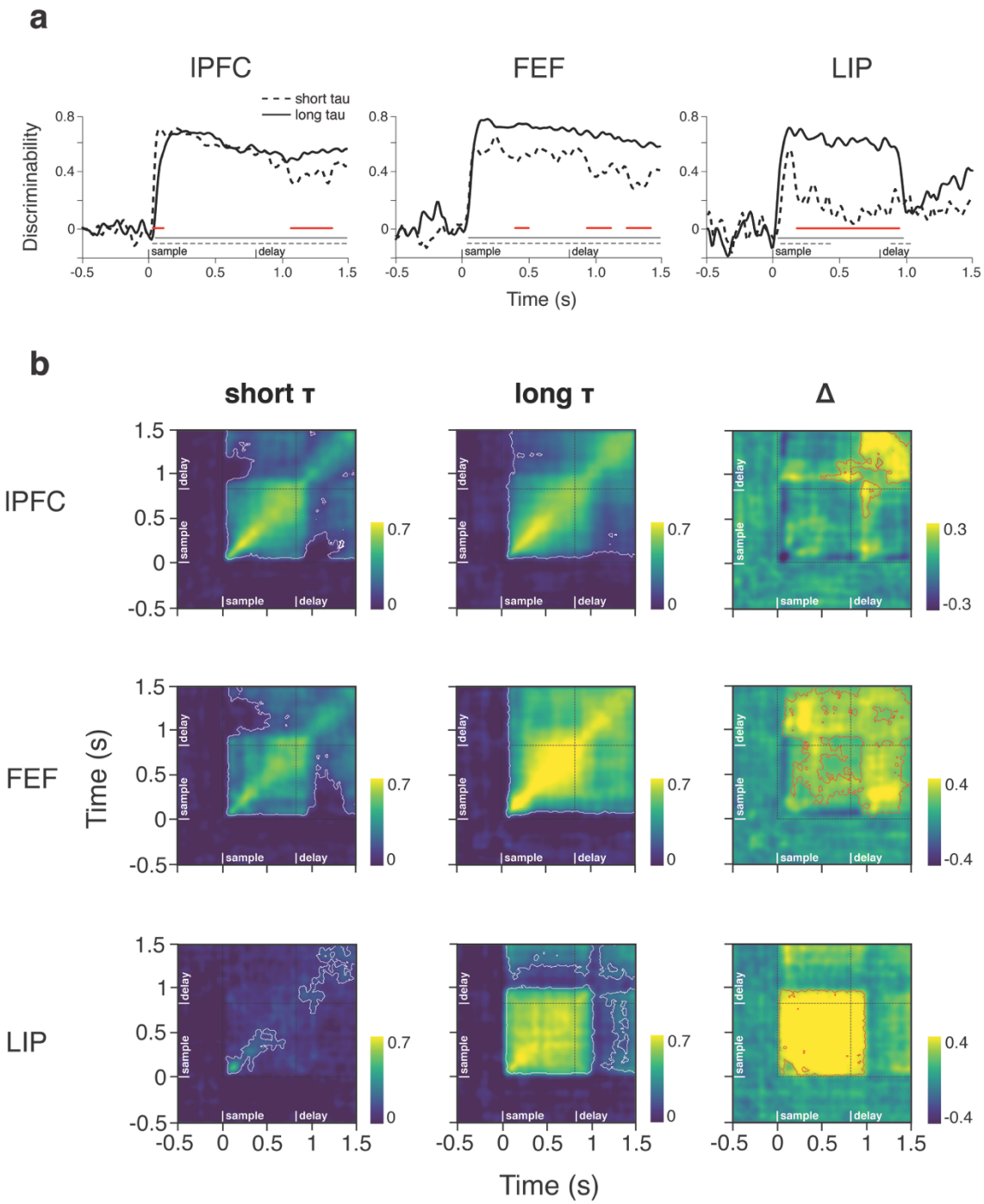
Supplementary Figure 3.1: Relationship between tau and firing rate. Log scaled scatter plots show tau (x-axis) against trial averaged firing rate (y-axis) for each task epoch (columns) and each brain region (rows). Every dot represents a neuron. (lPFC: n=265; FEF: n=168; LIP: n=136). Spearman rank correlation coefficients and associated p-values are displayed in the lower left corners. Note that none of the correlation coefficients was significant. Black solid lines depict linear fits to the log transformed data. Shaded area depicts 95% confidence intervals of the linear fits. Insets depict boxplots of epoch specific neuronal firing rates grouped according to associated short and long tau values (as determined by a median split). Red lines indicate median firing rate. Red squares indicate mean firing rate. Box edges represent top and bottom quartiles with whiskers indicating (1.5*interquartile range) in both directions. Outliers are not plotted. We performed a Wilcoxon rank sum test to evaluate whether the firing rates for short and long tau cells were drawn from the same distribution but found no evidence defying the null hypothesis: lPFC (Wilcoxon rank sum test, Baseline: p=0.21; Sample: p=0.13; Delay: p=0.75), FEF (Baseline: p=0.25; Sample: p=0.56; Delay: p=0.75), LIP (Baseline: p=0.87; Sample: p=0.53; Delay: p=0.7).



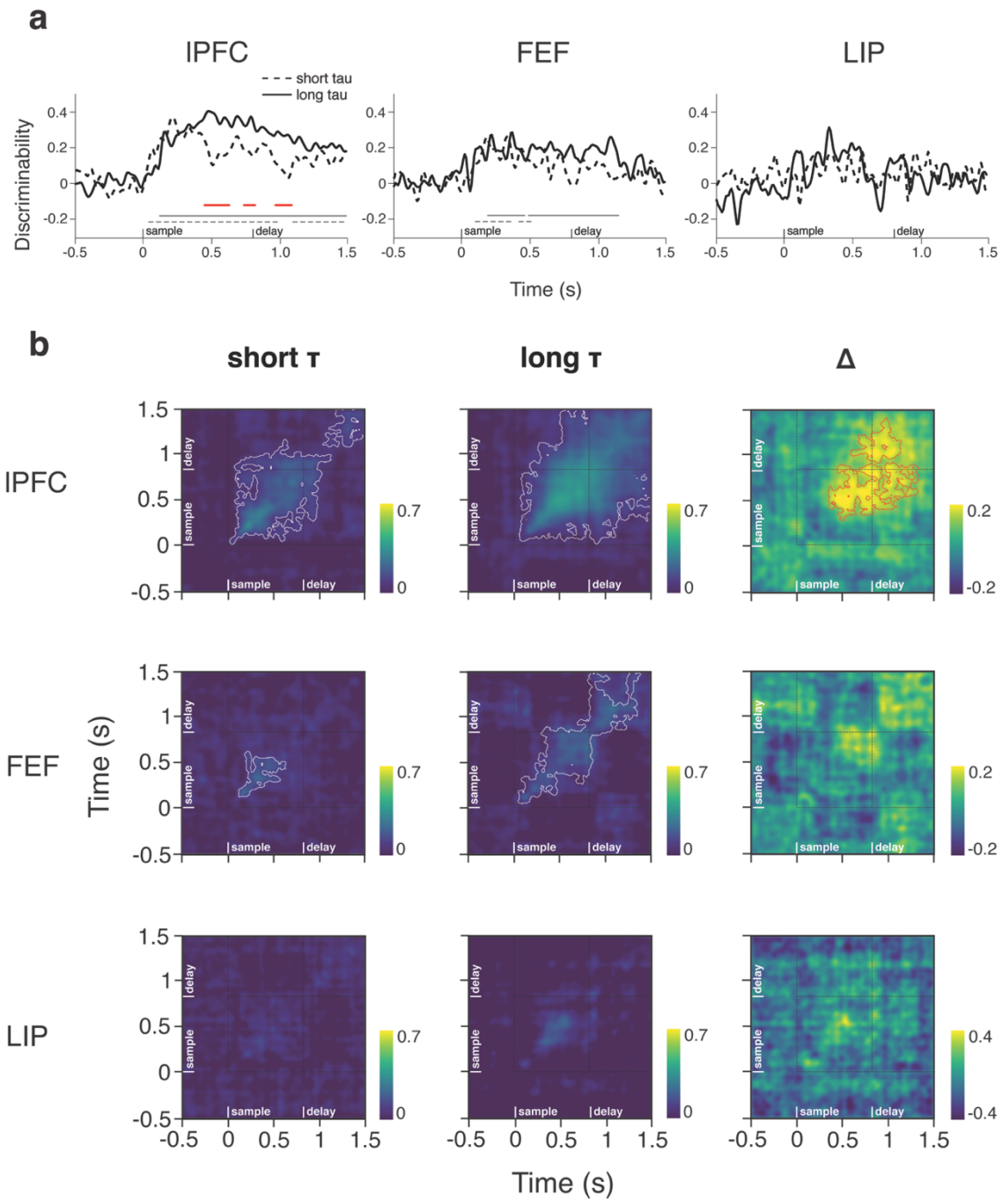
Supplementary Figure 3.2: Item information and intrinsic timescales for cells showing a significant ωPEV at any time point during the trial. Same conventions as in Fig. 3.2 hold true. Red line indicates a significant difference (cluster based permutation test, p=0.04). lPFC (n=203); FEF (n=121); LIP (n=53).



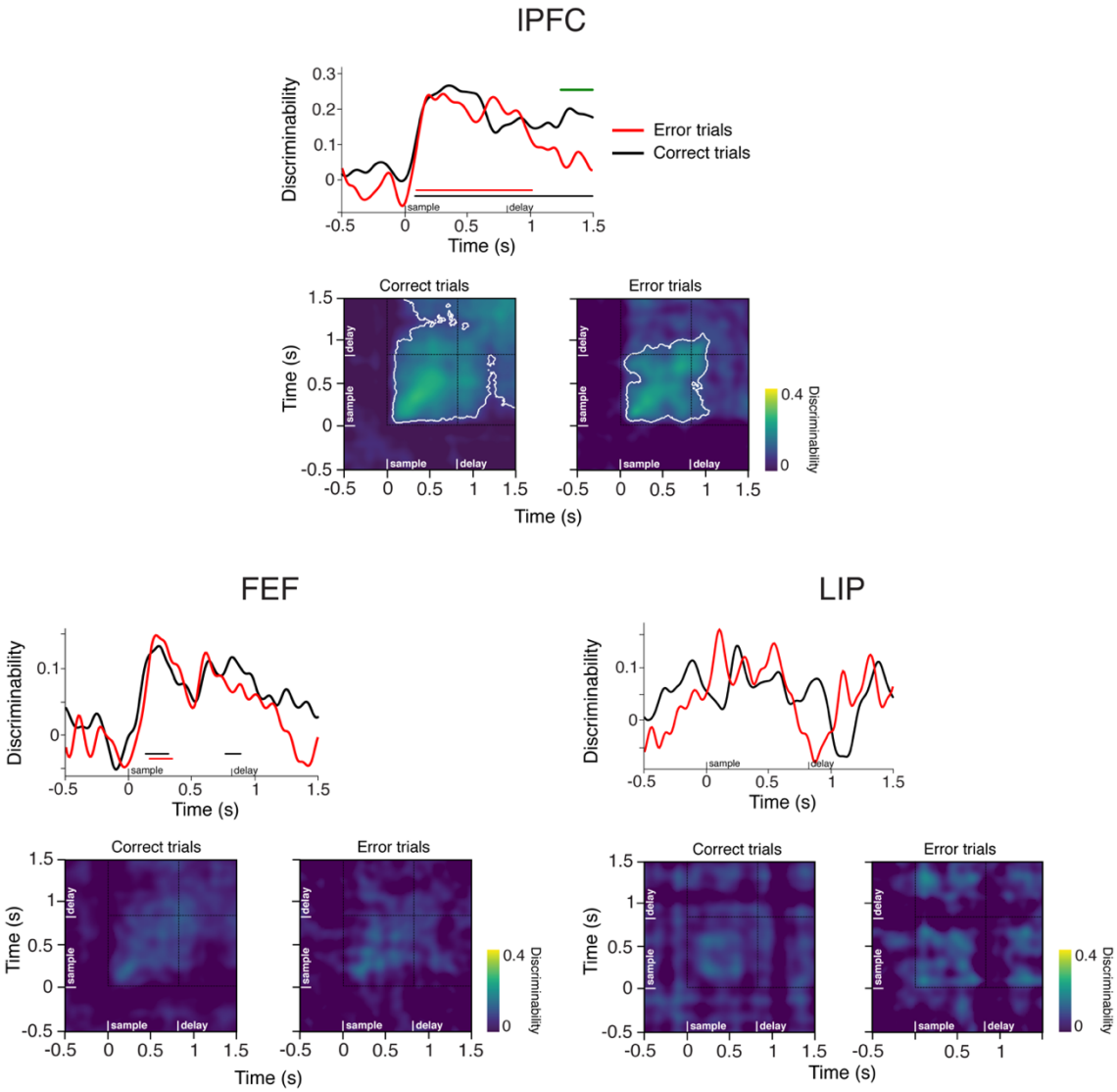
Supplementary Figure 3.3: **Evaluation of sample period onset times in relation to intrinsic timescales for FEF and LIP.** This figure shows the same plots as Fig. 3.3, for FEF (left column) and LIP (right column). **a)** Average ωPEV for cells with significant ωPEV at any time point during the sample period. **b)** Average tau for cells with significant ωPEV at time point on x-axis (red line) and fraction of cells with significant ωPEV at time point on x-axis out of all cells showing a significant ωPEV during the sample period. **c)** Scatter plots of encoding onsets (first point of significant ωPEV) versus tau for time periods 1 and 2. FEF: Time period **1** (Spearman rank correlation, r=-0.04, p=0.76, n=56); Time period **2** (r=0.04, p=0.71, n=96). LIP: Time Period **1**: (r=-0.21, p=0.34, n=22); Time Period **2**: (r=0.1, p=0.57, n=35).



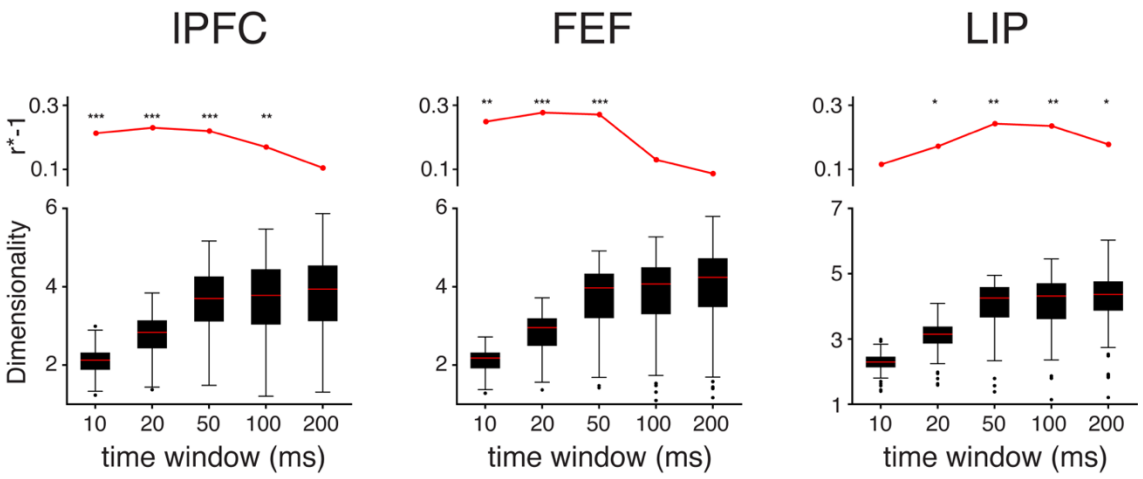
Supplementary Figure 3.4: Within-time and cross-temporal decoding for location information. Same as Fig. 3.4 but for location information only (see Methods). **a)** Within time decoding. **b)** Across time decoding. White contours: cluster based permutation test, $p<0.001$; Red contours: $p<0.01$. Color scale depicts decoding discriminability.



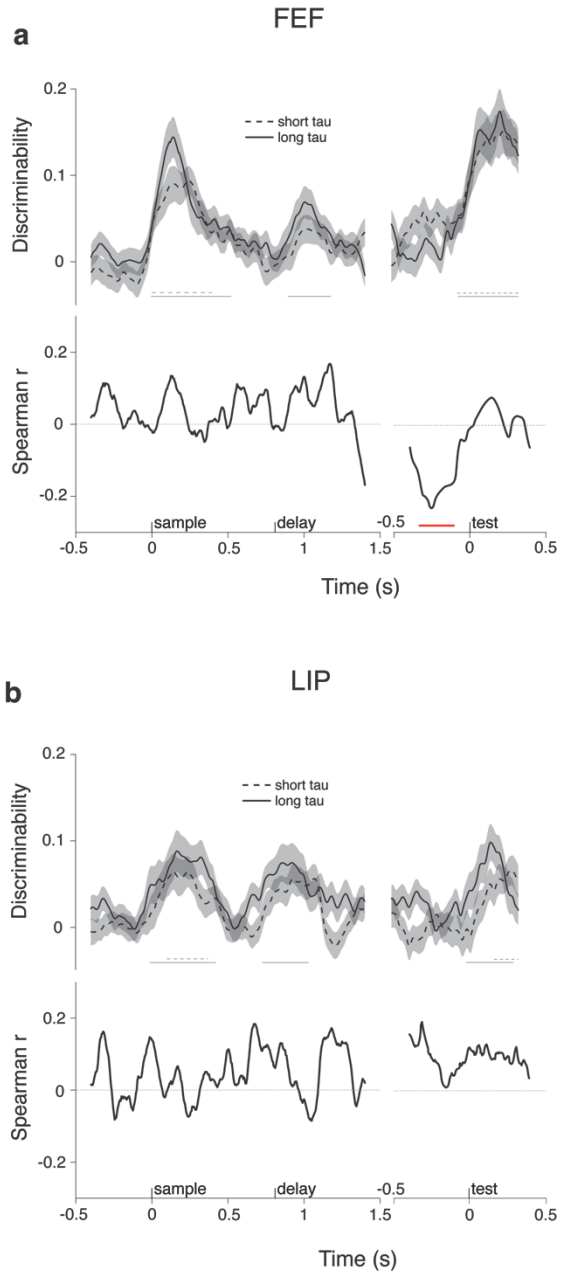
Supplementary Figure 3.5: Within-time and cross-temporal decoding for color information. Same as Fig. 3.4 but for color information only (see Methods). **a)** Within time decoding. **b)** Across time decoding. White contours: cluster based permutation test, $p<0.001$; Red contours: $p<0.01$. Color scale depicts decoding discriminability.



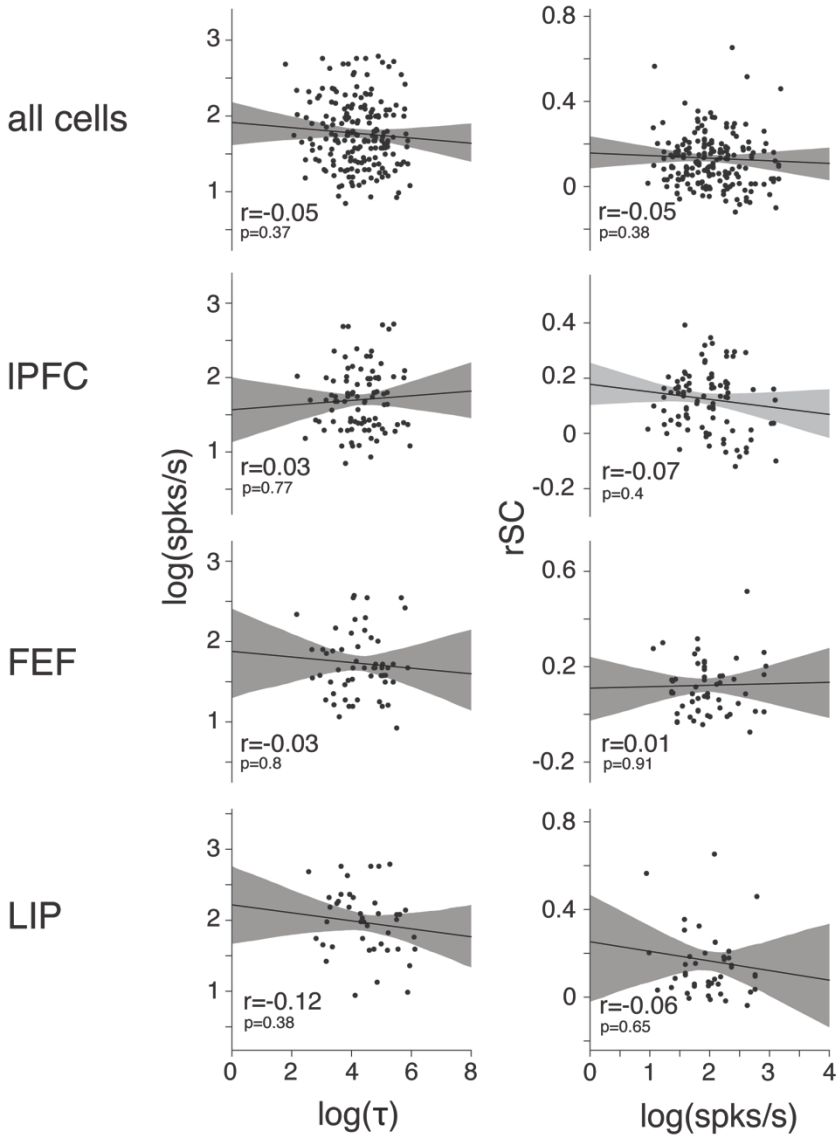
Supplementary Figure 3.6: Correct versus error trial target decoding. Within and across time target decoding within correct and error trials for each brain region. Within time target discriminability is depicted in the upper panels for each region. Black and red lines indicate discriminability for correct and error trial, respectively. The correspondingly colored horizontal lines indicate significant discriminability (cluster based permutation test, $p<0.05$). The green line indicates a significant difference between correct and error trials (cluster based permutation test, $p=0.035$). The lower two panels indicate the across time target decoding for each brain region. The discriminability score is color coded as indicated by the color bars. White contours indicate significant cross-temporal discriminability (cluster based permutation test, $p<0.05$). Black dashed lines distinguish task periods. lPFC (n=60); FEF (n=50); LIP (n=35).



Supplementary Figure 3.7: Effective temporal dimensionality estimated over different time windows and relationship with intrinsic timescales. Lower panels: Boxplots show N_{eff} estimated over differing time-windows (10ms*10, 20ms*10, 50ms*10, 100ms*10 and 200ms*10). Red lines denote median N_{eff} for each time-window over all cells, black boxes mark the upper and lower quartiles. Whiskers denote 15* interquartile range. Dots represent outliers. Upper panels: Each dot represents the Spearman rank correlation coefficient of N_{eff} and τ values, for each of the time window N_{eff} is estimated from. * represents $p<0.05$; ** represents $p<0.01$; *** represents $p<0.001$. lPFC (n=264); FEF (n=167); LIP (n=134).

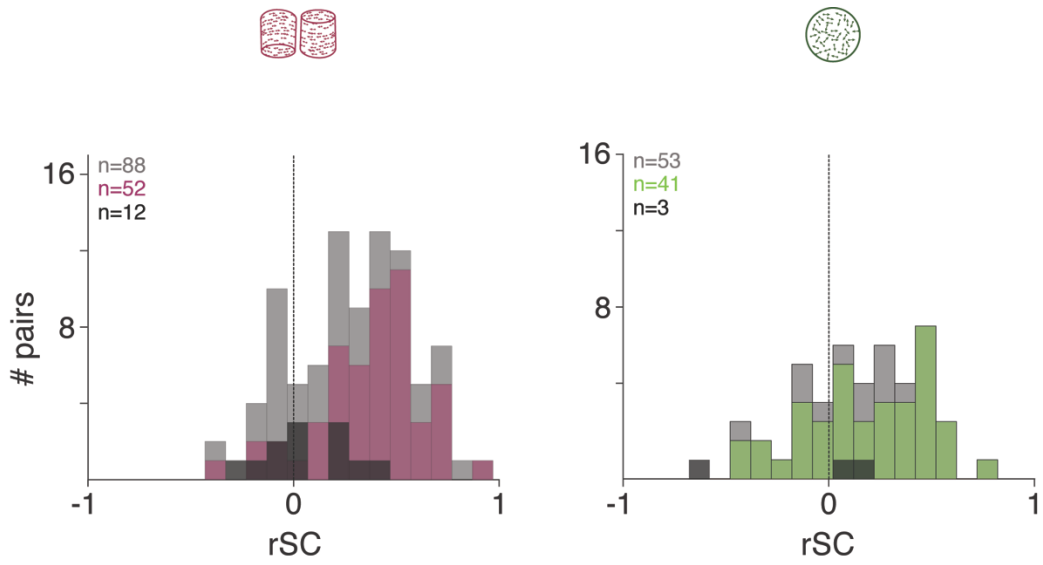


Supplementary Figure 3.8: Temporal discriminability in FEF and LIP. Same conventions as in Fig. 3.6. **a)** FEF; **b)** LIP. Red line indicated a significant difference (cluster based permutation test, $p<0.05$)

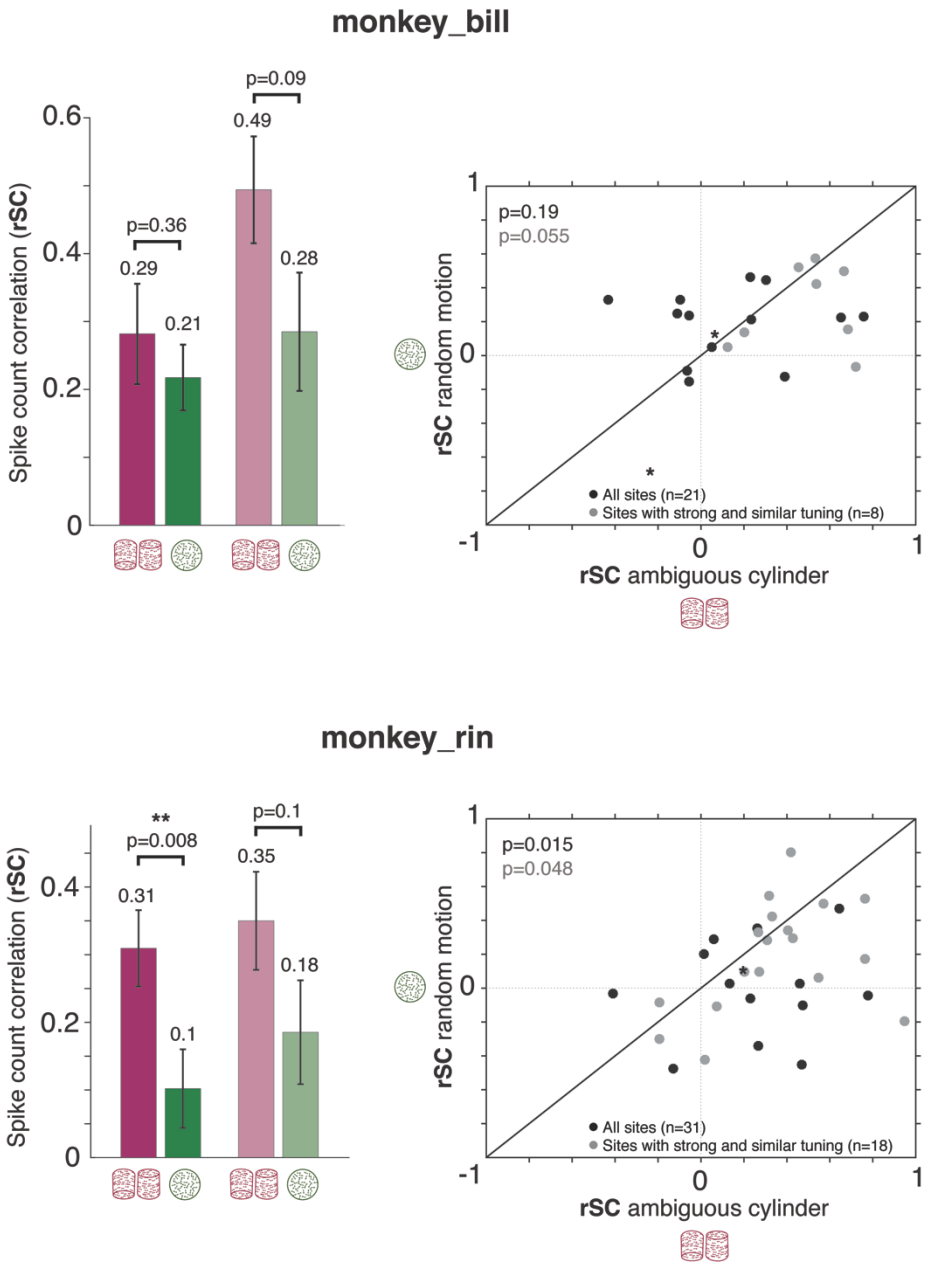


Supplementary Figure 3.9: Spike count correlations and intrinsic timescales. The first column shows scatter plots for the log transformed fixation period firing rate in spks*s^{-1} on the y-axis, versus the log transformed τ values on the x-axis. The r value within panels depicts the Spearman rank correlation coefficient for the two variables. Black solid lines show linear regression lines. Shaded region depicts 95% confidence interval. The second column shows scatter plots the log transformed fixation period firing rate (spks*s^{-1}) on the x-axis and rSC on the y-axis. The r value within panels depicts the Spearman rank correlation coefficient and associated p-value, between the two variables. All cells (n=255); lPFC (n=126); FEF (n=74); LIP (n=55).

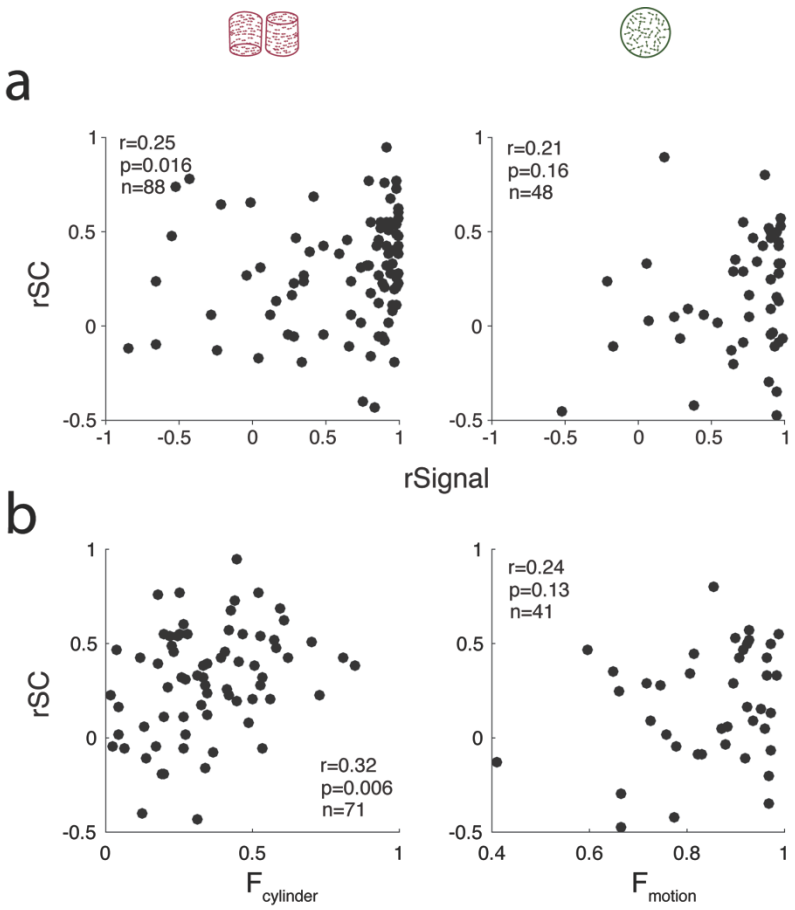
Appendix 3. Supplementary Figures to Chapter 4



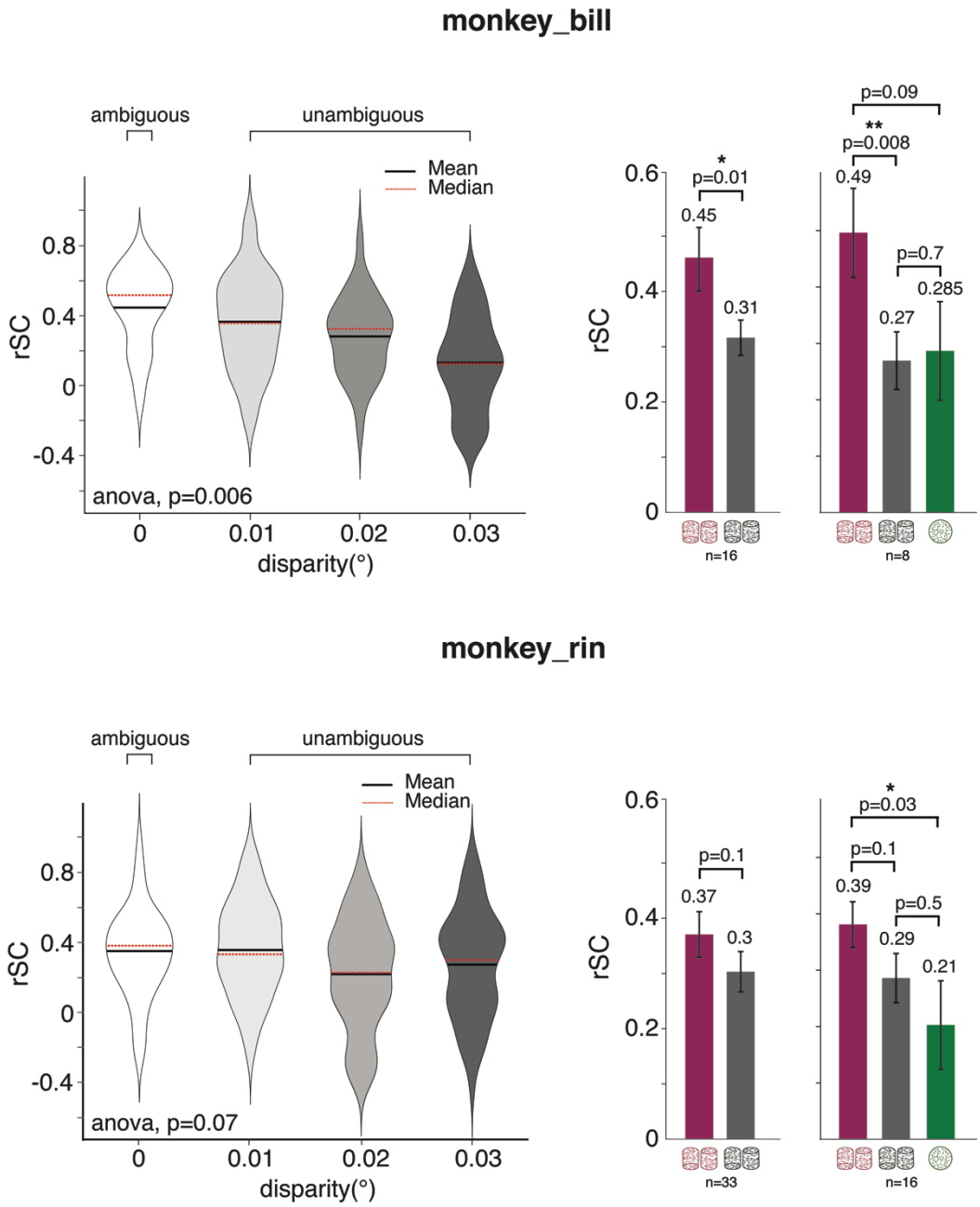
Supplementary Figure 4.1: Histograms of spike count correlations for the ambiguous cylinder and random dot motion. Grey coloured histograms show stimulus specific rSC distributions estimated from all pairs (i.e. SU/MU pairs) recorded with: the ambiguous cylinder (left panel); random dot motion (right panel). Purple and green coloured distributions indicate respective rSC sub-distributions for pairs with strong and matching SU and MU tuning for cylinder rotation (purple) and motion direction (green). Black coloured histograms represent distributions of stimulus specific rSC values for SU/SU pairs recorded from the same electrode. Dashed vertical lines indicate a rSC of zero.



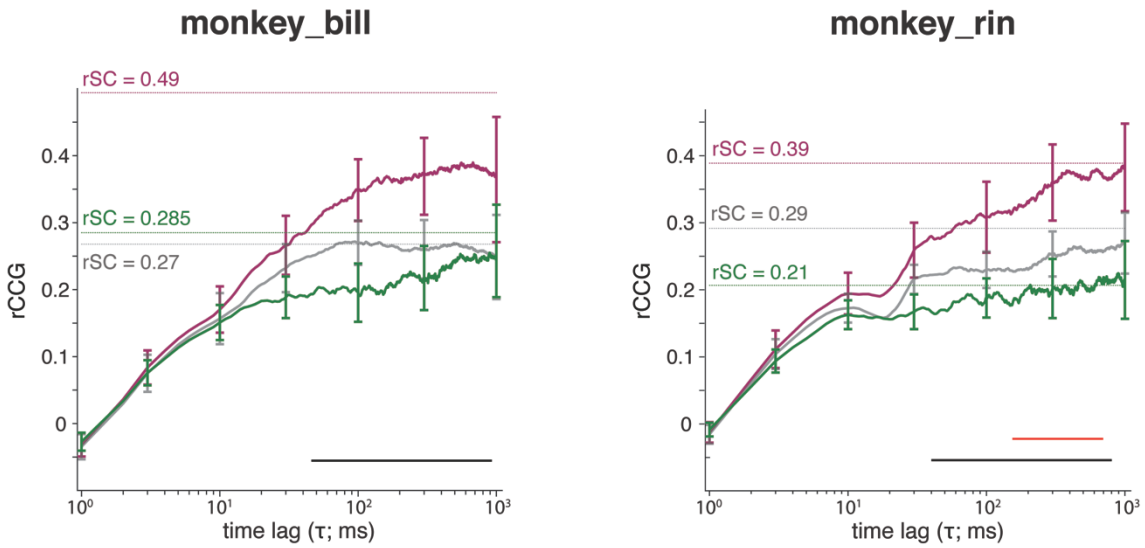
Supplementary Figure 4.2: Histograms of spike count correlations for the ambiguous cylinder and random dot motion (split by monkey). Conventions are the same as in Fig. 4.1 in the main text. Left column: Bar plot comparing the average rSC between the two stimuli; ambiguous cylinder (purple) and random motion (green). Bright bars show average values for all sites per animal; pale bars show average values for sites with strong and matching SU and MU tuning. P-values stem from t-tests on Fisher's z transformed rSC values. Error bars indicate s.e.m. Right column: The rSC values for the random motion stimulus (y-axis) are plotted against the rSC values for the ambiguous cylinder (x-axis) from the same sites, for each animal. P-values stem from sign tests.



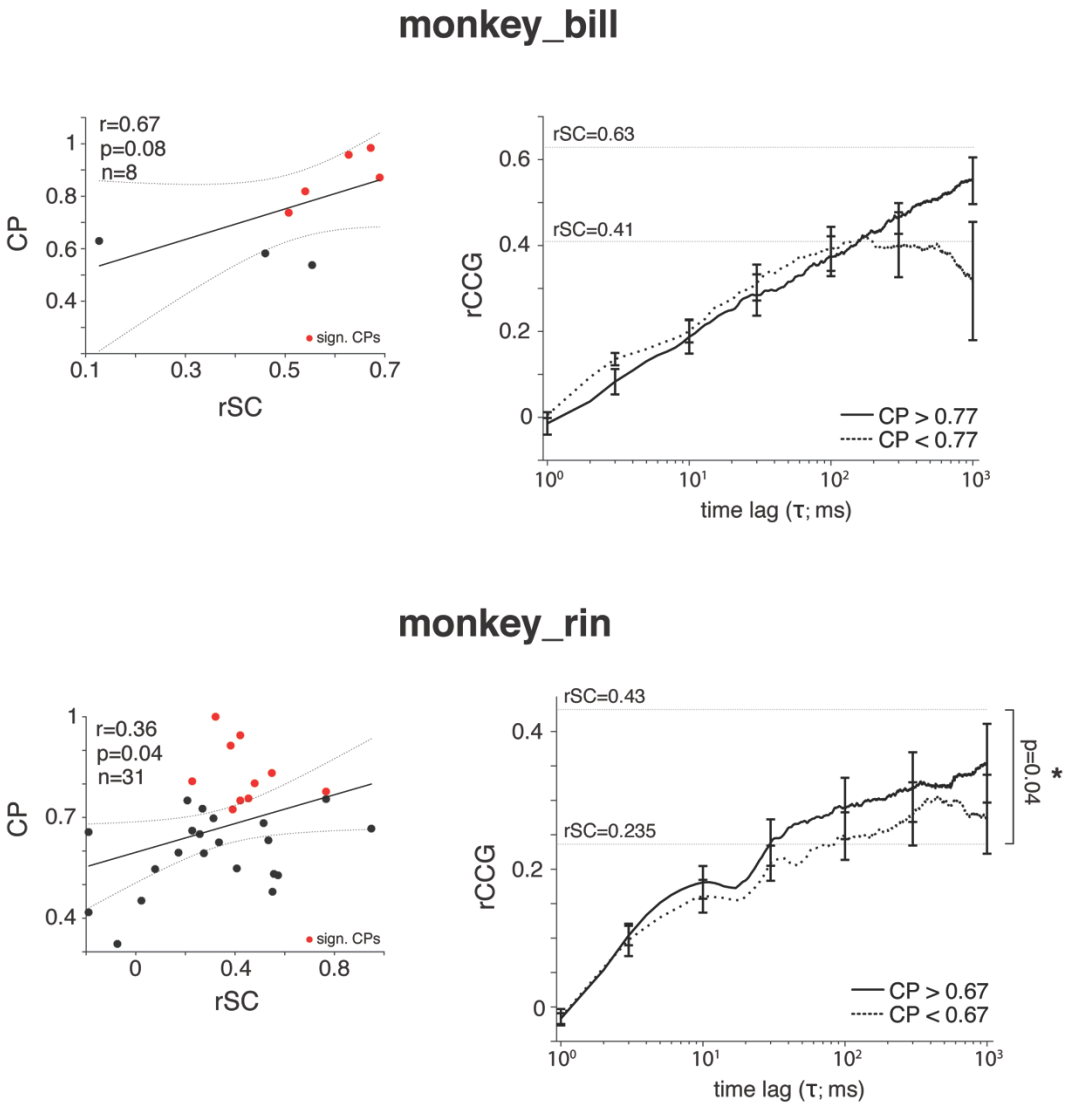
Supplementary Figure 4.3: Dependence of spike count correlations on tuning similarity and strength. **a)** Scatter plots showing rSC vs. r_{Signal} for all pairs recorded with ambiguous cylinder stimuli (left panel) and/or random motion stimuli (right panel). Each dot represents a recorded site (i.e. MU/SU pair). Correlation coefficients (Spearman rank, r) and associated p-values between rSC and r_{Signal} are displayed. **b)** Scatter plots of rSC for the ambiguous cylinder vs. $F_{cylinder}$ (left panel); rSC for random motion vs. F_{motion} (left panel).



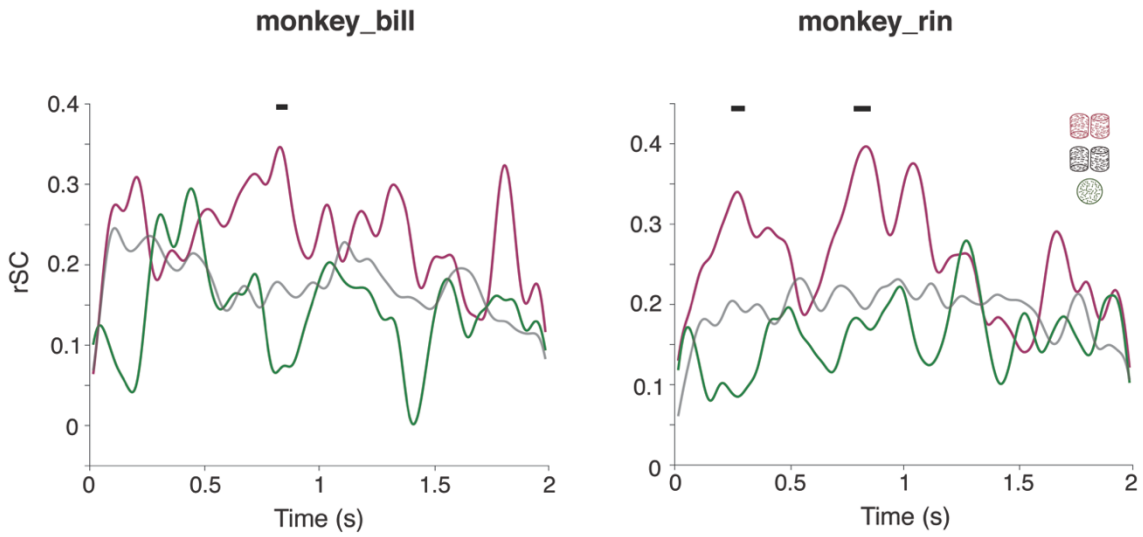
Supplementary Figure 4.4: Spike count correlations for intermediate cylinder disparities - for each monkey. Same conventions as in Fig. 4.2 in the main text. Violin plots in the left column show distributions of rSCs for cylinder stimuli with different disparity strengths. P-values stem from one-way ANOVAs on Fisher's-z transformed rSC values. Bar plots in the right column indicate mean rSC values for three stimulus conditions (purple: ambiguous cylinder; grey: unambiguous cylinder; green: random motion). Only sites with strong and matching MU and SU tuning (see Methods) were included. Left panel: Mean rSC values for all sites that were tested with the ambiguous cylinder (zero disparity) and at least one unambiguous cylinder stimulus (i.e. non-zero disparity). Right panel: Mean rSC values for sites that were tested with all three conditions i.e. stimuli. P-values stem from t-tests on Fisher's z transformed rSC values.



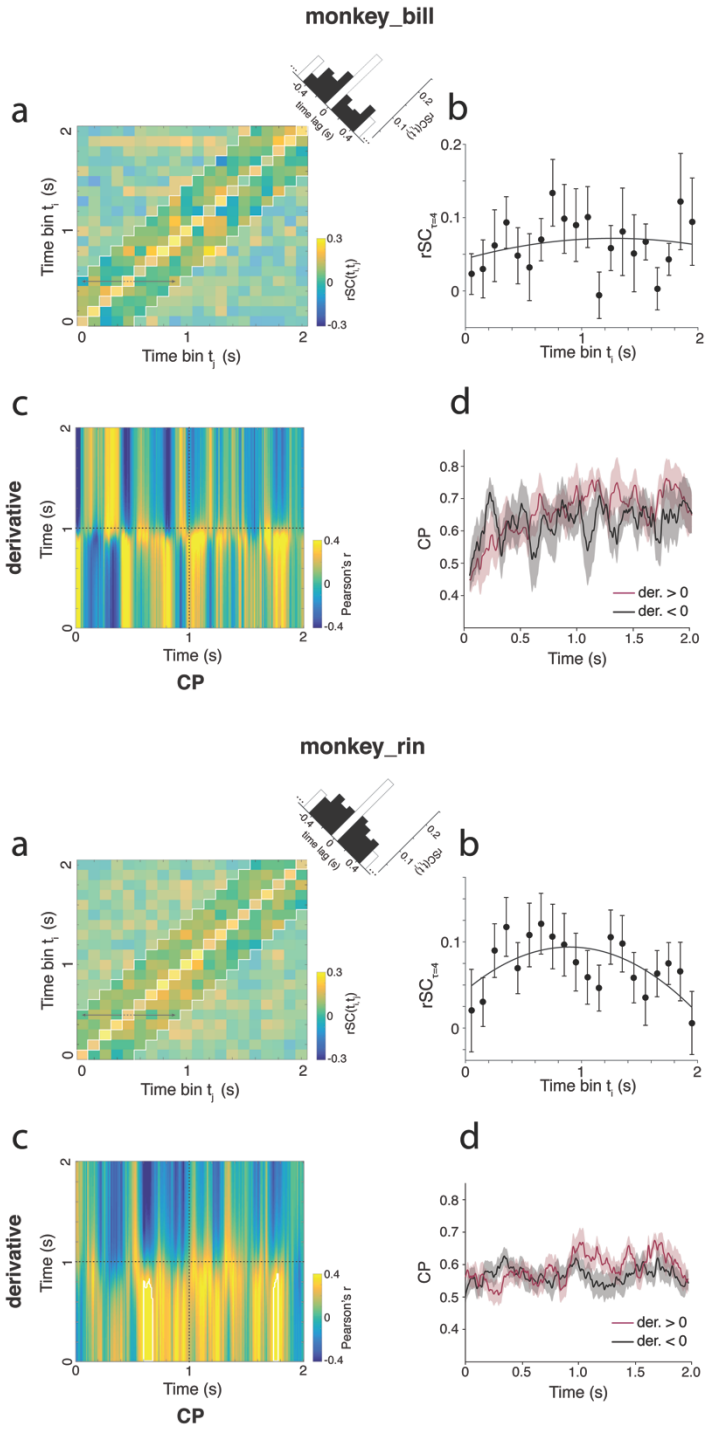
Supplementary Figure 4.5: Spike time correlations and timescales of correlations – data split by monkey. Same conventions as Fig. 4.3 in the main text. Average rCCG plotted as a function of the time-lag (τ) i.e. the integration window; for the ambiguous cylinder (purple), the unambiguous cylinder (grey) and the random motion (green) stimuli. Horizontal black and red lines indicate a significant difference between the purple and green and the purple and grey lines, respectively [$p<0.05$, *cluster-based permutation test,* (n=8 in monkey_bill; n=16 in monkey_rin)]. Horizontal dashed lines indicate the corresponding average rSC values. Error bars indicate the s.e.m.



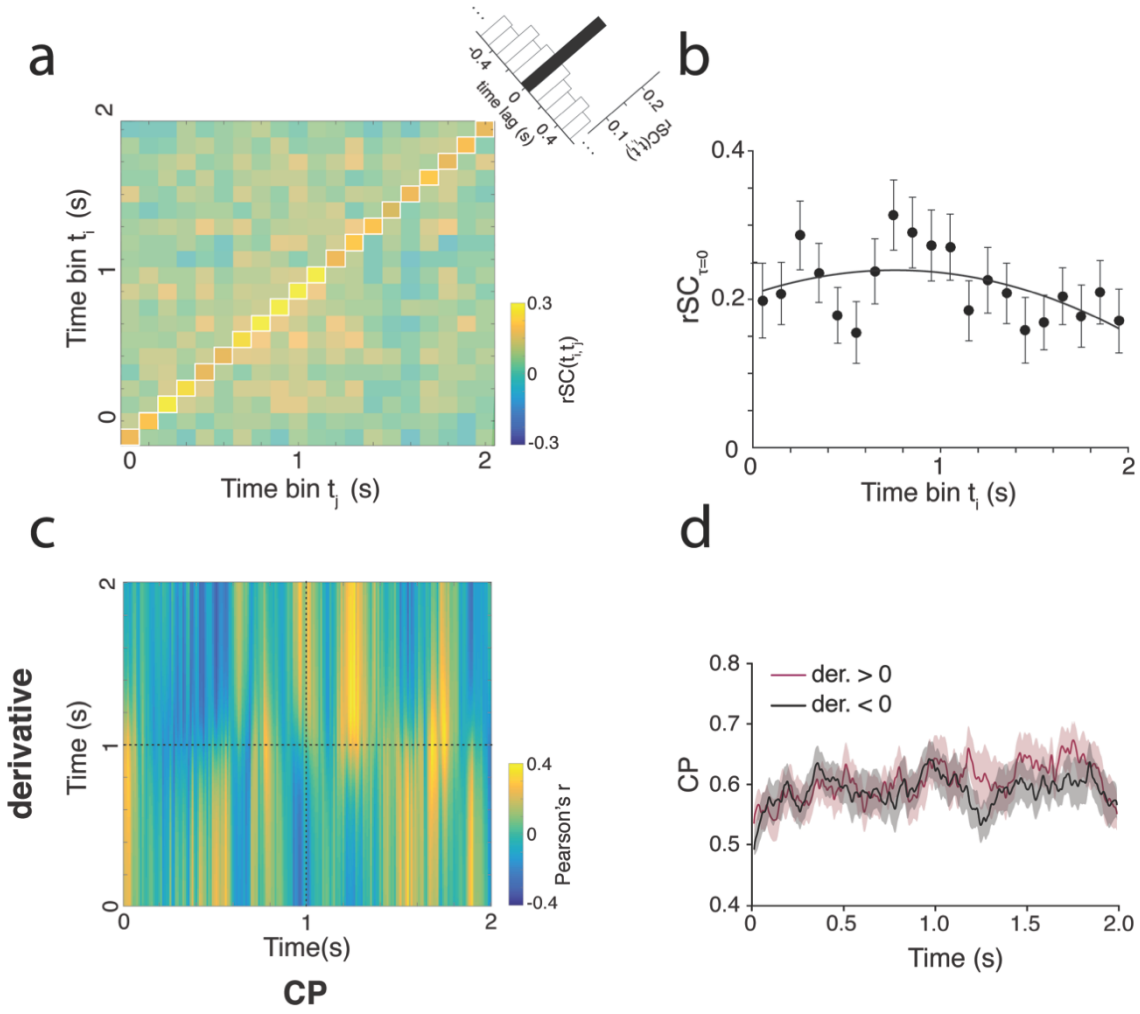
Supplementary Figure 4.6: Choice probabilities and spike count correlations – data split by monkey. Same conventions as in Fig. 4.4 in the main text. Left column: Scatter plots of SU CP and SU-MU rSC. Each dot represents a SU that showed strong and matching cylinder rotation tuning and for which a CP value could be measured. Red dots indicate SUs with a significant CP ($p<0.05$; cluster based permutation test). Black solid line indicates a linear fit to all data points with 95% confidence interval (dashed lines) and displayed correlation coefficient and associated p-value stem from the Pearson correlation. Right column: Average rCCG for subpopulations of sites split by the median CP associated with the SU. Error bars indicate the s.e.m. Horizontal lines indicate corresponding average rSC values. P-value stems from unpaired t-test on Fisher'z rSC values.



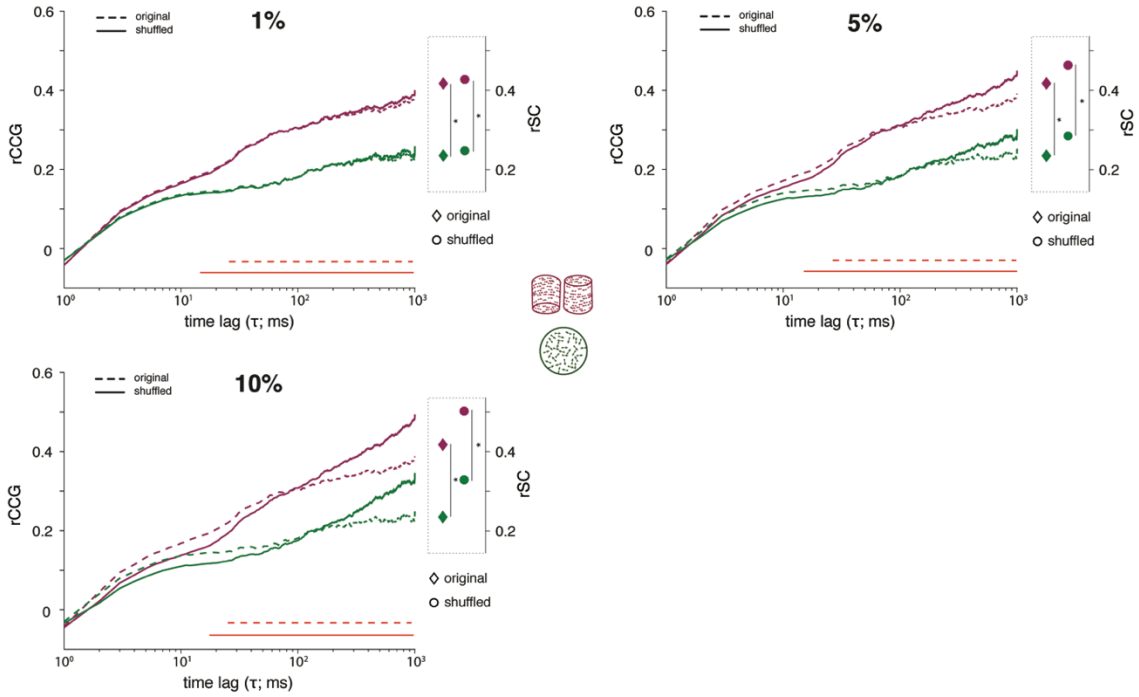
Supplementary Figure 4.7: Time-course of spike count correlations – data split by monkey. Same conventions as in Fig. 4.5 in the main text. Average rSC time-courses estimated from a sliding 100ms window. Purple, grey and green lines show the average rSC time-courses for the ambiguous cylinder, the unambiguous cylinder and random motion, respectively. Averages were computed over all sites recorded under all three stimulus conditions (n=8 in monkey_bill; n=16 in monkey_rin). The black horizontal bar indicates a time window with a significant difference between the ambiguous cylinder and the random motion condition ($p<0.05$; *cluster based permutation test*).



Supplementary Figure 4.8: Correlation time-course predicts choice probability - data split by monkey. Same conventions as in Fig. 4.6 in the main text. **a)** Spike count correlation matrices. **b)** rSC$\tau_{=4}$ averaged over all sites are plotted as a function of time-bins t_i. Error bars show s.e.m. The black line indicates a 2nd order polynomial fit to the rSC$\tau_{=4}$. **c)** The matrix for Pearson correlation coefficients between the derivative of the fitted polynomial in b) and CP across recording sites. White contours mark clusters with a significant Pearson correlation (p=0.01; cluster based permutation test). **d)** The CP time-courses averaged across sites with a negative derivative (black; n=5 monkey_bill; n=14 monkey_rin) and positive derivative (purple; n=4 monkey_bill; n=19 monkey_rin). Shaded area indicates s.e.m.

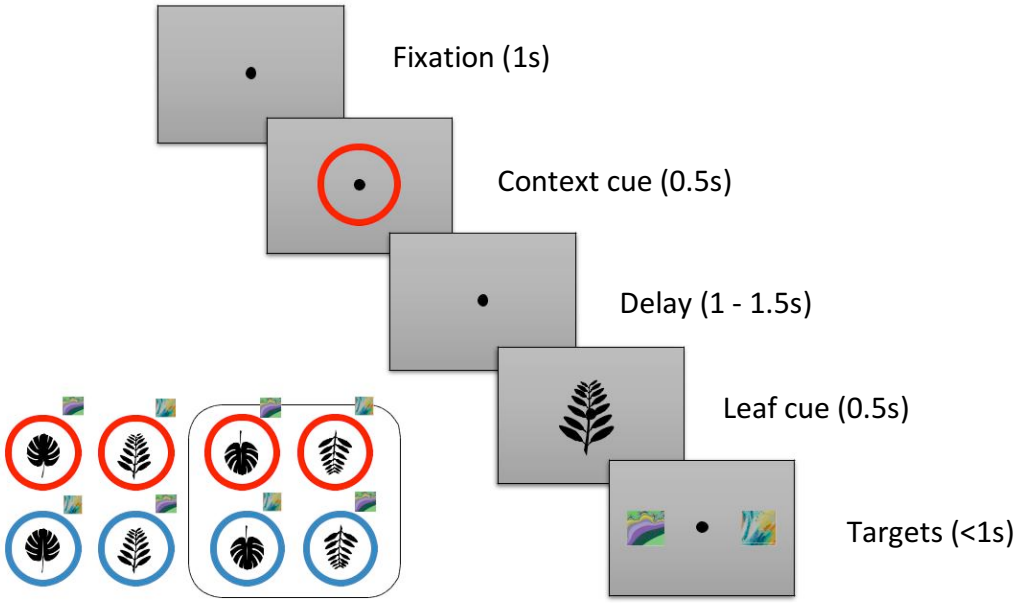


Supplementary Figure 4.9: Instantaneous correlation time-courses and choice probability. This figure is analogous to Fig. 4.6 in the text. Now, $rSC\tau_{=0}$ in **b** refers to the instantaneous rSC i.e. the main diagonal highlighted in **a**. Average CP time courses in **d** are comprised of n=24 sites with a derivative < 0 and n=15 sites with a derivate>0.



Supplementary Figure 4.10: Timescales of correlations for 'shuffled' spike trains. Each panel shows a replica of Fig. 4.3 in the main manuscript (without the unambiguous cylinder condition) for 'shuffled' spike trains with various fractions of misclassified spikes (as indicated by the percentages displayed). Misclassifications were repeated 1000 times i.e. on each repetition a fixed fraction randomly drawn spikes were exchanged between SU and MU. Panels show the mean rCCG and corresponding rSC values (insets) computed over all repetitions. Note, no confidence intervals are shown since they are barely visible on the panels' scales. Dashed lines and diamond shapes indicate rCCG and rSC values as computed from the original (without deliberate spike misclassifications). Solid lines and circles indicate rCCG and rSC values for data containing misclassified spikes (to various degrees). Purple colour indicates rCCG and rSC values for the ambiguous cylinder condition and green colour indicates the same for the random motion condition. Solid and dashed red lines show significant differences between the the differently coloured solid and dashed lines (cluster based permutation tests, $p<0.05$). Stars: *$p<0.05$; **$p<0.01$; t-test on Fisher's z.

Appendix 4. Supplementary Figures to Chapter 5



Supplementary Figure 5.1: Task design of learning experiment. An example trial: After fixating on a central dot, a context stimulus indicating the current rule is presented. Following a delay, a leaf stimulus is presented. Depending on the colour of the context cue and the identity of the leaf cue, NHPs have to saccade to one out of two presented target pictures to get a reward (target locations are randomized). Importantly, leaf orientations are irrelevant to the task. The inset depicts all possible correct context/leaf/target combinations (Note the XOR problem structure of the task, requiring non-linear mixed selectivity). Note, the task is complex e.g. humans need around 200 trials to learn this task. Over many months, we have trained monkeys to achieve an accuracy of about 65%, while recordings have started from day 1 of learning (we continuously train the animals offline while recordings are made every 1-2 weeks, using several linear probes in lPFC). Once the animals have fully acquired the task and stably perform around 80% accuracy, we plan to substitute old stimuli with new ones to test whether the acquired latent representation of the task, which we hope to find in PFC, allows for generalization to those new stimuli. We have also tested the same task in human participants using EEG. Humans show strong effects of transfer learning i.e. once the task structure has been learned, incorporating unseen stimuli and contingencies does affect performance only for a very short time. This suggest that humans use their knowledge of the structure of the task to generalize to new exemplars. Preliminary analyses of the EEG data suggest the formation of an abstract representation that could be used for generalizing to new stimuli. On the long term, the aim of this project is to characterize the emergence of a latent task structure i.e. representation that might be used as a scaffold to rapidly incorporate new information.