

Using a machine learning framework to understand intelligent behaviour in humans and non-human primates

Michał Józef Wójcik

New College
University of Oxford
Trinity Term 2023

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Experimental Psychology

Using a machine learning framework to understand intelligent behaviour in humans and non-human primates

Michał Józef Wójcik, New College, Oxford Trinity Term, 2023

Thesis abstract

This thesis explores the intersection of cognitive neuroscience and machine learning to shed light on the dynamic nature of neural processes supporting intelligent behaviour. Drawing upon research proposing that the human brain evolved in a generalist-specialist ecological landscape, I focus on neural processes that support adaptability at different time scales. Using a machine learning perspective, these generalist and specialist processes were recast as computational motifs, akin to those found in reservoir computing and task-optimised networks, respectively. The thesis posits the existence of a temporal evolution in neural strategies, from high-dimensional, generalist representations to low-dimensional, task-optimised ones, in the prefrontal cortex (PFC) over learning. This hypothesis is corroborated empirically in non-human primates, revealing that the PFC transitioned from initially coding task features uniformly to creating a task-optimised representation. Furthermore, the thesis confirms the capacity of the PFC to capture abstract environmental structures, thus facilitating the learning of novel but related tasks. Additionally, applying a machine-learning lens, it reconceptualises working memory as a dynamic computational resource. Human EEG data supports this view, showing working memory's role in filtering and maintaining task-relevant information, with efficient, low-dimensional neural codes emerging at decision time. The studies presented here, while rooted in a machine learning framework, underscore the significance of understanding the inherent nuances and limitations of this paradigm in comprehending the complexity and richness of neural processes.

This thesis was generously funded by the Clarendon Fund and the Saven European Fund.

Acknowledgements

First and foremost, my deep gratitude goes to my supervisors, the late Prof Mark Stokes, Prof Laurence Hunt, and Prof Nick Myers. This thesis stands as a testament to their unwavering, and perhaps more crucially, enthusiastic support. Mark, in particular, was a beacon of inspiration. The countless hours of intense discussions with him have shaped the conceptual heart of this work. He will be sorely missed not only as a mentor but also as a friend. Laurence graciously welcomed me as his student during one of the most challenging phases of my DPhil. His uncanny instinct for truth and precision has profoundly influenced the quality of this research. And to Nick, I am indebted for the vast ocean of knowledge he possesses, an expanse which he generously shared.

In my academic journey, I was fortunate to cross paths with Dr Dante Wasmuht and Dr Jake Stroud. Dante, with his infectious enthusiasm, unveiled the complex world of dynamical systems and the state-space perspective to me. Jake reshaped my understanding of brain function through the lens of machine learning, thus changing my trajectory as a scientist. Their contributions to my growth cannot be overstated, and it has been an honour to work alongside them.

I also wanted to thank my friend Dr Emilia Piwek who is truly one of the brightest people I know and probably will ever meet. My heartfelt thanks for the countless hours of listening to my scientific rambling, often at the most unconventional times.

I owe an immense debt of gratitude to my housemates and friends, Kevin, Alessandro, Mrinalini, and Vinita. The late-night discourses and shared culinary adventures have been a source of continuous emotional support and will remain etched in my heart forever. Equally, my gratitude extends to my dearest friends Maurits and Lauren, my saviours during the writing phase, who ensured I was well-fed and focused.

Lastly, the pillars of my life—my parents, my sister and my friend Aleksandra—deserve the most heartfelt appreciation. Their unwavering belief and support made this journey not only possible but also immensely rewarding.

To all mentioned and to those whose support silently underpinned this journey, I extend my deepest appreciation.

Contributions

In this thesis, each section represents my independent work. I received invaluable guidance on the theoretical framework, experimental design, and analysis for all empirical chapters from Prof Mark Stokes, Prof Laurence Hunt, Prof Nicholas Myers, Dr Jake Stroud, and Dr Dante Wasmuht.

The data described in Chapter 2 and Chapter 3 was collected in collaboration with the Prof John Duncan, Makiko Kadohisa and Makoto Kusunoki and analysed in collaboration with Jake Stroud. John Duncan led the experimental recordings, Makiko Kadohisa and Makoto Kusunoki performed the experimental recordings. I and Makoto Kisunoki programmed the task. I and Mikiko Kadohisa performed data pre-processing.

Dr Jake Stroud performed analytical derivations and constructed the optimised networks supporting the finding presented in Chapter 2. This work is included in this thesis in the supplementary materials for reference.

The data described in Chapter 4 was collected in collaboration with Dr Dante Wasmuht and Amy Li. Dante Wasmuht provided input to conceptualisation and the experimental design and Amy Li assisted with data collection.

This work was funded by the Wellcome Trust (award 101092/Z/13/Z to M.K., M.K., and J.D.), the Strategic Longer and Larger grant (awarded to L.H; BB/W003392/1), the Medical Research Council UK Program (MC_UU_00030/7 to M.K., M.K., and J.D.), the Biotechnology and Biological Sciences Research Council (award BB/M010732/1 to M.G.S.), Clarendon Fund and Saven European Scholarship (M.J.W), and the James S. Mc-Donnell Foundation (award 220020405 to M.G.S.).

CHAPTER 1: INTRODUCTION **8**

What does it mean to be intelligent? **8**
The evolutionary success of a flexible species 8
Bigger means better 10
Is intelligent behaviour synonymous with flexible behaviour? 14
A short story about a flexible agent 18
The prefrontal cortex and flexibility 21

A brave new world for brains and machines **24**
A machine learning framework of brain function 24
Adaptive coding and task-optimised representations 29
Random mixed selectivity 33

Large theories and how to test them **37**
What is feedback control and why do we need it? 37
System-level consolidation 39
Compositionality 41
Working memory is a computational resource 43
Learning shapes PFC representations 46

CHAPTER 2: LEARNING SHAPES NEURAL DIMENSIONALITY IN THE PREFRONTAL CORTEX. **49**

Abstract **49**

Introduction **50**

Methods **54**
Animals and task 54
Data acquisition and pre-processing 55
Constantinidis et al. 2016 dataset 56
Multiple linear regression 57
Generative models 58
Optimised feedforward networks 59
Decoding 60
Cross-generalised decoding 62
Measuring similarity between selectivity distributions 63
Principal Component Analysis 65
Statistical testing for decoding and cross. gen. decoding 66
Statistical testing for selectivity measures 66

Results **67**
Generative models of random and minimal selectivity 67
A low-dimensional task relevant geometry emerges over learning 73

Discussion **79**

CHAPTER 3: LOW-DIMENSIONAL REPRESENTATIONS SUPPORT CROSS-GENERALISATION **83**

Abstract **83**

Introduction **84**

Methods **88**
Animals and task 88
Data acquisition and pre-processing 90
Multiple regression 91
Decoding 92
Uniform Manifold Approximation and Projection 92
Principle Component Analysis 93
Cross-stimulus set generalised decoding 93
Statistical testing for selectivity measures 94
Statistical testing for decoding and cross-set generalised decoding 94

Results **95**
A dynamical solution for the XOR operation 95
The PFC constructs an abstract representation of context over learning 100
The PFC uses one XOR readout to classify both new and old stimuli early in learning 102

Discussion **106**

CHAPTER 4: WORKING MEMORY SHAPES NEURAL GEOMETRY IN HUMAN EEG **112**

Abstract **112**

Introduction **113**

Methods **117**
Participants 117
Procedure 117
Task 118
Stimuli 119

EEG acquisition and pre-processing 119
Behavioural analysis 121
Decoding 121
Neural dimensionality 123
Cross-colour generalised decoding 123
Full cross-generalised decoding 124
Significance testing 124

Results 125
Human behaviour exhibiting evidence of employing a hierarchical strategy 125
Humans learn to maintain the context information and construct the XOR feature over learning 128
The neural format of the context signal evolves over learning 132
Analysis of neural geometry reveals evidence of using minimal representations at the time of the decision 135

Discussion 143

CHAPTER 5: GENERAL DISCUSSION 150

Chapter summaries and contributions 150

Limitations of the machine-learning perspective 153

Future directions 157

The integration of neuroscience and machine learning 159

Conclusions 160

References 162

Supplementary information 175
Section 1: How to set selectivity parameters for optimal XOR decoding 175
Section 2: An energy cost on unnecessary neural activity 176
Section 3: Supplementary figures 179

Chapter 1: Introduction

What does it mean to be intelligent?

The evolutionary success of a flexible species

A staggering number of research papers in neuroscience commences with a declaration akin to the following: “The unique morphological and computational features of the human brain allowed our species to achieve unparalleled evolutionary success”. While this assertion might appear self-evident, it merits a closer examination. Evolutionary sciences produced various metrics of evolutionary success, or fitness over the years (Losos, 2017). One of the most popular, reproductive success, gauges a species success based on their ability to produce more offspring as compared to their competitors, within a specific timeframe. Given this particular criterion, a myriad of species demonstrates superior fitness compared to *homo sapiens*, or primates in general. For example, in the microworld of prokaryotic organisms, bacteria like *E. coli* double in their population size every 20 minutes, in optimal conditions (Allen & Waclaw, 2018). Another popular metric operationalises fitness as establishing ecological dominance through cumulative biomass (Bar-On et al., 2018). These analyses demonstrate that, although humans contribute ~0.06 Gt C (gigatons of carbon) to the global biomass, species like the artic krill or termites can rival these estimates (Bar-On et al., 2018). Interestingly, humans constitute a surprisingly small

proportion of the currently observed biomass on Earth (~0.0001%) when compared to the kingdom of plants (~82%) or bacteria (~13%). An alternative measure conceptualises evolutionary success as the survivability or longevity of a species, or even an individual. Long individual lifespans and the ability to persist across vast geological epochs on a species level are the result of finding and dominating an ecological niche where a species can achieve stable reproduction. A perfect example of animal species exploiting this strategy are sharks (*Elasmobranchii*). Their origin dates back 450 millions of years (Hara et al., 2018) and genome analysis of some species revealed multiple physiological mechanism supporting DNA preservation and repair (Marra et al., 2018). These mutations could potentially elucidate the observed longevity of sharks, which in some species has been observed to be 500 years (Nielsen et al., 2016).

These insights highlight the intricacies involved in evaluating the evolutionary success of *homo sapiens*. Yet while we may not dominate across every measure of evolutionary fitness, there exists one criterion that distinctly positions us as unparalleled victors in the evolutionary arena: *adaptability*. It has been recently proposed that *homo sapiens* defy the traditional division of species into generalists and specialists, by simultaneously exhibiting the ability to occupy a diversity of biomes and possessing the capacity to specialise in some of these environmental extremes (Roberts & Stewart, 2018). This unique ecological plasticity is further supported by the propensity of the early human species to actively seek out a mosaic of ecological landscapes instead of exploiting one particular niche (Zeller et al., 2023). It has been hypothesised that our ability to flexibly switch from a

generalist to a specialist niche within one generation is related to a weaker reliance on the final stages of the Baldwin effect (Varki et al., 2008). That is, many species are able to transform learned experiences into genetic changes that will be then passed onto their offspring. In these next generations, behaviours that were initially learned by their ancestors become automatic and habit-like. It has been proposed that within these fast time scales, hominids (especially *homo sapiens*), adapt to their current ecological niche by propagating crucial behaviours through social transmission, rather than genetic hardwiring (Varki et al., 2008). However, this mechanism of fast adaptation requires a neural system large and powerful enough to support the construction, maintenance, and social transmission of complex models of the environment.

Bigger means better

The hominin brain expanded rapidly over the last 3 million years (Buckner & Krienen, 2013). What distinguishes the human brain from other hominin brains is its notably extended growth period. This unique characteristic stems from the primate brain's *late means large* developmental pattern. In this dynamic, the neocortical progenitor cells divide longer than any other neural progenitor cells, leading to a disproportionately larger cerebral cortex in humans than other neural structures (Buckner & Krienen, 2013). The prolonging of the time window of embryonic brain development has been proposed as one of the possible physiological mechanisms responsible for the increase in the size of the human cerebral cortex, as compared to other primates (Buckner & Krienen, 2013). Consequently, the size of the human brain, after taking bodyweight into account,

is estimated to be approximately 5 times larger than the size observed in an average mammal (Buckner & Krienen, 2013).

This raises a fundamental question concerning the environmental pressures that led to this extreme expansion of cerebral tissue. As described above, the need for a fast adaptation mechanism that was not based on lasting genetic changes required additional computation capacity for the processing of extensive social communication, and thus could have led to cortical expansion (Varki et al., 2008). A large body of evidence, including positive correlations between the size of the neocortex and group size as well as the complexity of social interactions, supports this *social brain hypothesis* (Dunbar, 2003; Herrmann et al., 2007). Subsequent studies have, however, challenged this assertion on the grounds of two observations: (1) the lack of a relation between social organisation and brain size in some species (i.e., *grade shifts*), as well as (2) the correlation between brain size and other factors such as domain-general cognitive abilities (van Schaik et al., 2012).

Expanding upon the first criticism, the observation of species such as lemurs, spotted hyenas, and Bechstein bats showed that they live in structured social organisations that rival the complexity of primate groups, while remaining relatively small-brained (van Schaik et al., 2012). Furthermore, the relationship between social dynamics and brain morphology found in orangutans as well as in aye-ayes also argues against the social brain hypothesis, but in the opposite direction. Specifically, even though these species exhibit large brains, the social groups in

which they are embedded are relatively simple, as compared to related primates that live in highly structured groups (van Schaik et al., 2012). The small brain sizes of highly social animals like lemurs can be explained by life-history filters, such as the abundance of predatory species or severe seasonal shifts (i.e., whereby the high rate of early mortality results in a contracted brain maturity period and thus a small neocortex size) and thus are compatible with the *social brain hypothesis*. Yet the observations that some species have large neocortices in the absence of complex social organisation cannot be readily explained by this framework.

Detailed examination of these large-brained species revealed that they have highly evolved domain general intelligence capabilities (van Schaik et al., 2012). Interestingly, empirical evidence suggests that domain-general intelligence is not a uniquely human attribute but rather a phylogenetically old phenomenon that evolved independently in numerous species (van Schaik et al., 2012). These two observations led to the proposal that volatile environments encouraged the development of domain general cognitive skills which consequently drove neocortical expansion, particularly in primates (Reader & Laland, 2002). The standalone impact of this feature is not sufficient, however, to explain the increase of neocortex size observed in humans as compared to other primate species (van Schaik et al., 2012). Specifically, although domain-general cognitive skills allow the generation of creative and complex solutions, they provide only a modest fitness advantage. This is related to the fact that without social transmission these solutions are limited to the lifespan of the inventing individual. Additionally, the

generation of most useful strategies is highly probabilistic and thus they rarely arise independently in the next generations.

This critique resulted in the expansion of the *social brain hypothesis* into a format that accounts for grade shifts and the existence of large-brained animals with poor social skills. The new and improved *cultural intelligence hypothesis* states that the unparalleled expansion of neocortical tissue observed in humans is a result of the combined impact of environmental pressure for flexible, intelligent behaviour, complex social organisation and a third, developmental, factor (Herrmann et al., 2007; van Schaik et al., 2012). Specifically, it states that the unique solutions generated thanks to the general-domain capabilities can be rapidly transmitted in a specie that exhibits complex social dynamics. Building on the work of Vygotsky (Vygotsky, 2012), it also argues that the bulk of learning in primates takes place in the early development, under the structured supervision of adult role models (Moll & Tomasello, 2007). The interaction of these three factors most likely created a positive feedback loop that gave rise to the abilities to construct, maintain and, most importantly, transmit unbelievably complex models of the environment. This in turn led to the rapid expansion of the neocortex over the anthropogenesis of our species. Importantly, the fact that neural representations of the currently occupied ecological niche were socially (and not genetically) transmitted allowed the circumvention of the trade-off between generalisation and specialisation. In a scenario of forced migration to a different ecological landscape, a putative specie would retain the advantages of general-domain cognitive capabilities, as they are genetically predetermined, but would discard or reconfigure the culturally

transmitted model of the previous environment rapidly over several generations. The cultural intelligence hypothesis is thus compatible with the proposition that humans occupy a generalist-specialist ecological niche.

Is intelligent behaviour synonymous with flexible behaviour?

As discussed in the previous section, during the evolution of our species environmental pressures shaped neural mechanisms that supported adaptation on various scales. The generalist-specialist label that has been proposed as a description of our species' ecological niche also provides an accurate depiction of the computational abilities that are characteristic of contemporary humans. An average person is perfectly capable of solving a myriad of simple tasks they have never encountered before. Furthermore, if the environmental pressure is strong enough, they are not only able to specialise in a specific, highly complex task, but also to generate an optimal solution for this particular problem. The computational architecture, as well as the biological implementation of the breadth and depth of computational operations that make human cognition unique remain, however, a subject of ongoing research.

In psychology, domain-general cognitive capabilities were traditionally summarised by the general intelligence coefficient (Sternberg & Kaufman, 2011). It has been proposed that intelligent behaviour arises from a single computational principle that leads to the proficiency in a multitude of tasks (Spearman, 1923). Support for this statement was found in an overwhelming number of studies that showed that the performance in several distinct cognitive task seems to correlate (Sternberg &

Kaufman, 2011). The idea of a single unifying principle was later contested by theories arguing that every human exhibits a unique constellation of various types of intelligence (Gardner, 2011). Importantly, these theories focused on the educational, environmental and cultural impact, underscoring the fact that mental abilities are not fixed and can be developed over time through experience.

While the nature of intelligence is still a central theme in psychological research, the debate over its structure —i.e., whether it consists of one unifying factor or represents a constellation of distinct mental abilities — continues to elude a conclusive resolution. This ambiguity largely stems from the absence of a robust theoretical foundation, as the structure of intelligence is often contingent upon its measurement methodology, specifically the degree of internal similarity or dissimilarity between selected tests (Summerfield, 2022a). As a result, these psychometric methodologies primarily produce models that capture the covariance structure of a specific set of cognitive abilities rather than delineating cognitive models in their own right (Birney & Beckmann, 2022; Frischkorn et al., 2022).

The nuances of this issue were illustrated effectively by a classical study of frontal lobe patients performing everyday tasks (T. I. M. Shallice & Burgess, 1991; Summerfield, 2022a). Frontal areas of the cortex are often referred to as the hub for higher cognitive functions such as complex decision making or planning – i.e., processes vital for supporting intelligent behaviour (Miller & Cohen, 2001). However, previous accounts convincingly demonstrate that patients with lesions in these brain areas perform surprisingly well on intelligence tests, sometimes even

demonstrating enhanced performance after the injury (T. Shallice & Evans, 1978). In line with these accounts, the patients studied in Shallice and Burgess's experiment also performed well on an intelligence test. The main purpose of the study, however, was to relate the performance scores obtained from a neuropsychological test to the proficiency with which the patients solved everyday problems. After the laboratory phase of the experiment, the patients were thus given a list of errands and were required to complete them in a novel realistic environment, within a given timeframe. This list contained tasks like buying a loaf of bread, arriving in a specific place at a specific time, finding out the most expensive item that a shop sells, and inquiring about the weather. Moreover, these tasks needed to be completed under constraints such as minimising expenditure, accomplishing them swiftly, and ensuring that no shop is entered without making a purchase. Surprisingly, all participants exhibited extremely poor performance in this task, as compared to a control group that was matched with regards to age and intelligence scores. Some participants could not complete the experiment without committing a criminal offence. This study illustrates that operationalising intelligence through a predefined task-set performed in a laboratory setting may result in its narrower portrayal, failing to fully capture the behavioural flexibility inherent to our intuitive understanding of this concept.

One approach that allows to circumvent this problem is to decompose intelligence into basic cognitive processes (Frischkorn et al., 2022; Sternberg, 2022). According to these process-based conceptualisations, measuring and, more crucially, explaining the interpersonal differences in intelligence can be achieved

by building models of basic cognitive processes and relating them to behaviour that could be considered intelligent. Even the scientists who proposed the early psychometric approaches made efforts to identify computational operations that give rise to intelligence at the behavioural level (Sternberg, 2022). For example, Spearman (Spearman, 1923) proposed that intelligence is organised around the ability to extract abstract relations between experiences (*eduction of relations*) and then utilising this abstract representation to dissect a different experience *(eduction of correlates*). Thurstone (Thurstone, 1999) and Stenhouse (Stenhouse, 1974) theorised that intelligent behaviour originates from the ability to control automatic responses in favour of more voluntary and, using modern terminology, model-based behaviour. Subsequently, Donald Hebb (Hebb, 1942), influenced by his research on neural plasticity, split the general intelligence coefficient into two separate concepts: intelligence A and intelligence B. The former describes a raw intellectual potential, mostly determined by genetical factors and early life experiences. The later refers to the manifestation of intelligent behaviour after the innate intellectual potential was shaped by environmental factors such as education, cultural experiences, or socioeconomic status. Hebb's reformulation of the traditional definition of intelligence underlines the importance of measuring an individual's cognitive capacities in the context of their current and previous environment. Moreover, it stresses the fact that, even if there is an innate, genetically determined intellectual capacity, the correct way of measuring intelligence is through tasks that mirror an individual's everyday life and contain a dynamical, learning-oriented component.

Although some of these authors contributed to the operationalisation of intelligence as a static trait, their theoretical considerations about its basic computational components permit a more dynamic definition. In fact, integrating the aforementioned perspectives seems to reveal a definition of intelligent behaviour that is organised around the notions of adaptability and flexibility.

Interestingly, contemporary process-based approaches even go as far to equate general intelligence with cognitive flexibility (Birney & Beckmann, 2022). Such an operationalisation bears significant ramifications for the possible methods of measurement. Firstly, it requires the identification of the cognitive processes crucial for generating flexible behaviour. Secondly, to construct a mechanistic model of the relationship between these processes and behaviour, the physiological implementation of these functions needs to be explored. Finally, building on the conceptualisations that split intelligence into general task-agnostic and specific task-optimised modules, possible ways of understating how neural circuits give rise to flexibility should involve learning. More precisely, an emphasis should be placed on the dynamic transition from the state of the neural system at the beginning of learning, when a generalist strategy is likely to be employed, to the highly specialised state observed after extensive learning. This approach provides the means to delineate the computational operations attributed to generalist versus specialist adaptation and thus capture the temporal dynamic of intelligent behaviour.

A short story about a flexible agent

Before delving into the empirical and theoretical research on cognitive processes supporting generalist and specialist strategies, it is essential to reflect on the challenges an intelligent agent faces and their corresponding neural representations. The environment we live in is constantly changing, often on different timescales. Humans are required to solve a myriad of diverse task every single day, often in rapid succession, or even simultaneously. Furthermore, the order in which these tasks should be performed is often hard to predict, and distractions or interruptions are frequent. Sometimes, we are also required to learn a narrowly defined and complex task for an extended period of time, with the expectation that we will be able to perform it automatically and effortlessly. Finally, we are constantly bombarded with information which makes our environment complex and high-dimensional.

An agent that performs well in such a niche should have access to a general-purpose toolkit that can be employed to solve most problems without the necessity of lasting adaptive changes (i.e., domain-general cognitive abilities). One solution would be to represent *as many features of the environment as possible* and combine some of them dynamically to meet the current task demands (Enel et al., 2016; Rigotti et al., 2013). Such a maximalist strategy supports flexibility, as none of the dimensions of the environment are arbitrarily discarded or represented preferentially. The agent can select which of the features present in its rich world model are currently relevant for performance and use their combination to guide behaviour. Furthermore, it can quickly adapt to a new task by recombining a different set of already represented dimensions.

While a general-purpose toolkit is versatile enough to address a broad range of problems, there exist inherently specific tasks that are confined to narrow domains, necessitating an alternative strategy. Such challenges often require the acquisition of automatic response patterns that can be generated without the involvement of metabolically expensive computations. An agent using this strategy must also possess a capability for dimensionality reduction. That is, such an agent needs to be able to discard irrelevant information and favour only the dimensions that support performance. However, assigning saliency to specific features of the environment results in lasting changes to the neural system and consequently makes it less efficient when the agent is exposed to a different task.

The processes outlined above appear at odds with one another. On one hand, the agent needs to be able to quickly solve a myriad of tasks without specialising in them. Yet, there are moments when the agent must also demonstrate an ability to acquire lasting, adaptive changes to the environment. In certain situations, she must utilise adaptable representations to examine all potential combinations of environmental features. In contrast, in other contexts, she should concentrate solely on the most informative dimensions. It seems thus that depending on the environmental requirements, different types of representations should be employed. How can the primate brain sustain both capacities simultaneously?

This dichotomy – between the competing demands for representing rich, *high-dimensional* representations and compressed, *low-dimensional* representations of

a cognitive task – lies at the heart of the experimental work that I will present in this thesis. In the subsequent sections, I will examine regions of the primate brain possessing unique features, positioning them as adept orchestrators of both specialist and generalist strategies. Furthermore, using a novel framework that utilises concepts from machine learning, I will operationalise these processes and relate them to behaviour. Concluding these considerations, I will examine two prominent theories—systems consolidation and feedback control—that elucidate how specialist and generalist processes operate harmoniously at varying timescales, ensuring their divergent objectives do not conflict at any given moment.

The prefrontal cortex and flexibility

The classic study by Shallice and Burgess (1991) discussed in the previous section demonstrates the complex relationship between the functionality of the prefrontal cortex (PFC) and an intuitive understanding of what defines intelligent behaviour. Modern research complements this notion by demonstrating that this brain area is engaged in cognitive processes supporting flexible behaviour such as planning, complex decision-making, working memory, and control of inappropriate actions – processes that fall within the purview of executive functions (Miller & Cohen, 2001). The interest in the PFC as a driving module for intelligent behaviour stems from its unique anatomical location and the heterogeneity of selectivity profiles observed in this area.

The prefrontal cortex occupies the anterior part of the cerebral cortex and has been proposed to be strategically positioned to serve an integrative function in cognition

(Miller & Cohen, 2001). Its uniqueness stems from the vast and dense connections it shares with both cortical and subcortical regions (Fuster, 2001). This includes bilateral cortico-cortical connections with visual cortices which allows the PFC to guide the processing of low-level sensory features in support of adaptive behaviour (Fuster, 2001). Furthermore, the PFC shares extensive and varied connectivity with the limbic structures such like the amygdala and hippocampus (Eichenbaum, 2017; Ghashghaei et al., 2007). This enables the integration of information stored in memory with the incoming sensory stimulation, as well as the contextualisation of current processing using emotional states such as novelty. Recent advances in empirical and theoretical neuroscience also demonstrate the importance of the PFC-cerebellum network for intelligent, flexible behaviour (Bostan & Strick, 2018; Pemberton et al., 2022; Yoon et al., 2017). More specifically, this network is believed to have parallel processing capabilities, whereby the PFC is involved in planning and executing, whereas the cerebellum optimises these processes for efficiency.

A traditionally employed approach in neuroscience, embodied by the representationalism doctrine, posits that the function of a specific brain area can be inferred from the selectivity profiles exhibited by its constituent neuronal cells. Specifically, if the firing rate of a cell tracks the feature X, its function is inferred to be related to the processing of feature X. This approach proved to be fruitful when researching the sensory cortices, as the stable selectivity profiles observed there led to the discovery of the hierarchical architecture of sensory information processing (Hubel & Wiesel, 1959). However, relating PFC cell selectivity (as well

as population level firing rates) to behaviour remains a challenging task, due to their special properties (Duncan, 2001). Specifically, research on PFC selectivity has demonstrated that regardless of the performed task, a significant proportion of cells in the area will be selective to at least some aspects of the experiment (Duncan, 2001). For example, empirical reports show that PFC cell are responsive to a variety of variables such like location, shape, colour, occurrence, motor response, reward prediction, or task rules (Duncan, 2001). Furthermore, these cells not only exhibit mixed selectivity profiles responding to multiple variables at the same time, but also often vary in terms of their temporal profile of firing (Stokes et al., 2013).

The dynamic heterogeneity and adaptability of PFC neurons challenge the traditional confines of representationalism. These cells demonstrate an unparalleled flexibility, adjusting their selectivity based on the nature of the information and the immediate environmental demands. To understand the PFC's contribution to intelligent behaviour, we need a paradigm that recognises intelligence as the emergent outcome of numerous interacting foundational processes.

A brave new world for brains and machines

A machine learning framework of brain function

New methodological advances are revolutionising neuroscience. We can now record from an unprecedented number of cells (Steinmetz et al., 2021), as well as precisely manipulate the neural dynamics of entire neural populations (Sadtler et al., 2014). This creates an opportunity to revisit and update the general framework that drives neuroscience discoveries. In the classical approach, only the activity of a small number of cells would be observed and a small, circuit-level model of the population's function would be devised based on the selectivity profiles of single cells (Richards et al., 2019).

Although this approach has shed light on some fundamental topics such as motion detection in the retina (Kim et al., 2014) or control of rhythmic movements (Marder & Bucher, 2001), it has profound limitations. More specifically, whilst identifying the representational properties of single cells or a population of cells certainly provides insights into the possible function that an area subserves, these coding properties should not be equated with function. This point was illustrated by Jonas and Kording using functioning microcontrollers as a compelling example (Jonas & Kording, 2017). The authors employed analysis methods typically used in empirically-driven neuroscience to dissect the data generated by a classical microprocessor. The aim was to use a traditional representational perspective on a "model organism" that performs computations that are both sufficiently complex and fully understood. The authors found that, although the analysis methods

offered an interesting description of the data, these techniques fell short of recovering the information processing hierarchy underlying the function of microprocessors (i.e., logical gates). These results illustrate that representations are a by-product of computations, and their description does not constitute a mature model of brain function. Instead, a framework that focuses on the mechanistic origin of these coding properties could provide a more fertile ground for formulating general models of brain function (Richards et al., 2019; Summerfield, 2022a).

The interactions between artificial intelligence research and neuroscience gave rise to a new perspective on brain function: a machine learning (ML) framework for neuroscience (Richards et al., 2019; Saxe et al., 2021; Summerfield, 2022a; A. Zador et al., 2023). It rests on the observation that artificial neural networks (ANNs) can reproduce complex computations that were long thought to be exclusively part of the human domain. Examples of such computations are speech recognition and generation, language processing, complex planning, and reasoning (Summerfield, 2022a). Crucially, ANNs are not only able to achieve human (or primate)-level performance when trained on the same tasks, but simultaneously, these systems seem to be able to reproduce the representational characteristics of the solutions employed by humans (or primates) (Flesch et al., 2022; Mante et al., 2013; Piwek et al., 2022; Stroud, Watanabe, et al., n.d.). Some of the examples are grid-cell-like selectivity profiles (Banino et al., 2018), visual illusions (E. Watanabe et al., 2018), and shape tunning curves (Pospisil et al., 2018). In some multi-layered ANNs representations supporting model-based decision-making were observed,

even though the system was explicitly designed to run model-free computations (J. X. Wang et al., 2018). Model-based behaviour observed as an emergent property of multilayer dynamics bears a striking resemblance with the computational motifs observed in the human prefrontal cortex (J. X. Wang et al., 2018).

The success of these algorithms can be attributed to a general principle guiding machine learning: the fact that the computations transforming inputs into the desired outputs are not hand-crafted but instead learned through experience (Richards et al., 2019). More specifically, to fulfil a specific learning goal, these systems fine-tune their internal multilayer dynamics (bounded by the architectural constraints) by gradually updating the network parameters according to a set of equations. Using machine learning terminology, these three components correspond to the *network architecture*, *learning rule*, and *objective function*. Importantly, the machine learning framework aims to identify what types of architectures, learning rules, and objective functions give rise to computations that utilise representations bearing close resemblance with the codes observed in humans and non-human primates (Richards et al., 2019; Summerfield, 2022a; A. Zador et al., 2023). The recent convergence of neuroscience and machine learning has ignited an exploration into the architectures, learning rules, and objective functions that, when combined, might offer a mechanistic understanding of the neural processes giving rise to intelligent behaviour.

Architectures describe what type of computations single units can perform and how these units are arranged relative to each other. Importantly, these features dictate

the information flow in an ANN, as well as put constraints on the computations that the network can run. In deep learning, networks consist of multiple layers, feedforward or recurrent, which can correspond to different brain regions (Richards et al., 2019). This multilayer structure often gives rise to emergent computations that are similar to the ones observed in humans (Summerfield, 2022a). It has been proposed that the PFC is best approximated by a network that has recurrently connected units (RNN), as it allows computations to be performed over time, which is essential in many cognitive tasks (Chaisangmongkon et al., 2017; Mante et al., 2013; H. F. Song et al., 2017). For example, through time-varying nonlinear integration, these models can reproduce working memory processes and context-dependent selection that has been observed in the non-human primate PFC (Mante et al., 2013).

Learning rules determine how the internal parameters of the network (i.e., synaptic weights and biases) are updated over training. In other words, they provide an estimate of the gradient described by the objective functions. ANNs with multiple layers are typically trained with the canonical error backpropagation algorithm. This method has been, however, disputed as a general biologically plausible learning rule, as it requires specific forward and backward flows of information as well as symmetric feedback weights (Richards et al., 2019). However, algorithms that not only approximate backpropagation-of-error in multilayer networks, but also are biologically plausible have been proposed recently (Greedy et al., 2022).

Objective functions describe the learning objective of the model and define how it performs when doing a specific task. More specifically, they are functions of the synaptic weights of an ANN and can be formulated in a task-agnostic manner. One of the commonly used objective functions in machine learning is cross-entropy loss (Goodfellow et al., 2016). Recently, objective functions that are more biologically plausible have been proposed, such as the objective length defined in the predictive coding framework (Whittington & Bogacz, 2017), the reward-scaled log-probability of a sequence of actions (a function maximizing reward in reinforcement learning; Roelfsema & Ooyen, 2005) or empowerment (measuring the degree of control an agent has over their environment; Klyubin et al., 2005). It is important to note that identifying which objective functions have shaped neural populations within the cortical areas is a notoriously difficult task, as these functions may not be directly observable in the brain (Richards et al., 2019).

Neural networks used as models of brain functions should be subject to constraints (Richards et al., 2019). That is, an algorithm that performs well on *all* possible tasks does not exist (Wolpert & Macready, 1997) and, therefore, a limited set of tasks should be chosen to make optimisation possible. This idea underlies the success of some of the most powerful algorithms in AI, as incorporating assumptions about a solution to a given task into the computations themselves (i.e., inductive biases) allowed for convergence (Richards et al., 2019). Similarly, in neuroscience, focusing on a limited set of tasks (i.e., the brain set) that are crucial for adaptation to a changing environment could foster the identification of features that are key for specific brain functions. Focusing on inductive biases

seems a particularly promising approach in this context, as it has been hypothesised that the remarkable computational properties that the primate brain possess originate from the local and global neural architecture shaped by evolutionary pressure (A. M. Zador, 2019). An approach that capitalises on the idea that complex computation can arise as an emergent property of constrained, multilayer systems seems an appropriate framework to analyse the relation between intelligent behaviour and the PFC.

Adaptive coding and task-optimised representations

It has been proposed that the heterogeneity of selectivity profiles of PFC cell results from their tendency to adapt to the current demands posed by the environment (Duncan, 2001). More specifically, in contrast to sensory neurons, which retain a life-long specialisation in representing a particular physical feature, PFC cells become selective to those aspects of the environment that are crucial for the current behavioural goal. On a computational level, this adaptive coding results in a correlated pattern of neural firing rates, confining the possible states the system can represent to a low-dimensional manifold (Dubreuil et al., 2022; Gao & Ganguli, 2015; Hirokawa et al., 2019). Such low-dimensional representations have important computational benefits like robustness to noise and the potential for cross-generalisation (Bernardi et al., 2020; Flesch et al., 2022). A wealth of empirical research confirms the PFC cells' ability to be adaptively tuned to task-relevant information (Bernardi et al., 2020; Cromer et al., 2010; Flesch et al., 2022; Hirokawa et al., 2019; Reinert et al., 2021). Yet, the mechanisms shaping this

adaptability and its precise role in generating intelligent behaviour remain areas of active research.

The learning of task-specific representations can be well understood within the framework of optimisation (Hennig et al., 2021). In the machine learning literature, optimisation refers to the process of finding or approximating the best possible solution from a large number of available choices, given a predefined objective (Goodfellow et al., 2016). That is, when training an artificial neural network, the weights of the connections between units are slowly and progressively adjusted, such that the model's predictions are as close as possible to the training dataset. This simple principle allowed some networks to perform remarkably well in tasks that were previously thought to be exclusively part of the human domain (Richards et al., 2019; Summerfield, 2022a). Taking inspiration from ML research may help us in finding the missing link between changes in behaviour and changes to synaptic plasticity (Hennig et al., 2021). It is important to emphasise, however, that gradual changes to local connectivity as a mechanism driving behaviour is not a new idea in neuroscience or psychology and has been previously used to explain perceptual learning and decision making (Engel et al., 2015), motor learning (Haith & Krakauer, 2013) and reinforcement learning (Niv, 2009). What the machine learning framework offers as a novel approach, is a method of identifying the objective that a specific brain area aims to achieve, along with the constraints under which this occurs.

As mentioned before, using a machine learning perspective to operationalise biological processes such as learning in neural networks, requires a particular focus on inductive biases. For example, biological brains face constraints that are alien to ANNs, such as the imperative of energy efficiency given finite metabolic resources. Metabolic cost is a constraint that seems to be of particular importance when exploring learning within the prefrontal cortex. In algorithmic terms, it might be analogously represented through adding a regularisation term, such as the L2 or L1 norm, to the objective function. Such formulations of metabolic constrains strongly influence the representational capacities of a system. Specifically, the network prioritises only those input dimensions with a direct relation to performance, filtering out irrelevant dimensions. It is important to emphasise that, even though certain neurons demonstrate a selectivity towards these irrelevant features purely by chance during the initial stages of learning, the pressure of metabolic constraints ensures an eventual reconfiguration. Over time, such neurons depart from their initial selectivity profiles, instead converging towards a configuration that adheres to metabolic efficiency. This process could be modelled by optimising weights for performance while simultaneously penalising large activation or weight values. Consequently, weights linked to irrelevant features would diminish to zero. An objective function containing an L2 regularisation term could thus offer a precise description of the adaptive coding property observed in the prefrontal cortex. Regrettably, due to the lack of a sufficiently resolved learning dynamic, no currently available data can be used to test this prediction.

Although the ML perspective on brain function consists of three components, recent advances mostly shed light on the influence of architectures and learning rules on learning dynamics (Richards et al., 2019). Objective functions have yet to receive substantial attention from the scientific community. This might be related to the fact that the objectives that shape neural systems are not directly observable. They can be, however, formulated using the stringent language of mathematics and thus indirectly inferred by comparing the learning dynamics of the constructed systems to the learning dynamics observed in real data (Richards et al., 2019). This process requires the data to meet a particular criterion: neural recording collection must span from the animal's initial exposure to the task through to its mastery. Regrettably, datasets fitting this description are rare. Chronic recordings from a specific brain area pose challenges, primarily because of the onset of necrosis at the recording site. Moreover, animal movement can displace electrodes, leading to recordings from varying neurons over time. While recent advancements in recording technologies exemplified by such devices as neuropixels 2.0 (Steinmetz et al., 2021) offer promising solutions to these challenges, the anticipation for a richer collection of learning-detailed datasets continues.

Optimisation can be positioned as one amongst several mechanisms driving adaptive behaviour. In conditions where the same task is being performed for an extended period, it is likely to offer a precise yet succinct description of learning dynamics – especially when biological constrains are taken into account. Although our environment consists of such periods of stability, we also often experience

intervals in which we are exposed to multiple interleaved tasks. In such scenarios, a traditional optimisation-based approach falls short, as evidenced by the phenomenon known as 'catastrophic forgetting' (French, 1999). This occurs when training a network sequentially on two tasks; the learning from the second task might overwrite the adjustments made for the first, deteriorating performance on the initial task. This underscores the need to consider alternative mechanisms to fully understand adaptive behaviour in volatile environments.

Random mixed selectivity

A changing environment requires the ability to switch between different task representations or utilising distinct features of the environment, both in rapid succession. Recent advances in computational neuroscience revealed that the ability of PFC neurons to mix different types of information (e.g., sensory, abstract, memory-based or motor) leads to important computational benefits in this context (Rigotti et al., 2013). It not only allows for a large number of diverse information types to be represented (high-dimensional representations) but can also capture the interactions among them (non-linear representations). The former characteristic is particularly important for flexible behaviour, as it can be used to contextualise current computations, whereby a different neural state is generated depending on a specific variable. In this section, the computational benefits of both high-dimensional and non-linear representations will be explored. Furthermore, I will discuss the type of algorithms that can give rise to these representations and detail how these representations can be utilised to generate adaptive behaviour without extensive learning.

To demonstrate how the PFC can generate these complex high-dimensional codes, an artificial toy network consisting of three putative neurons can be imagined (**Fig. 1.1a**; adapted from Fusi et al., 2016). These neurons exhibit various selectivity profiles to two types of stimulation: sound and light. Neuron number 1 responds selectivity to the light intensity, whilst the firing rate of neuron number 2 is exclusively modulated by the intensity of sound. Finally, neuron 3 is responsive to a *linear* mixture of both stimulation types. The network is required to discriminate yellow conditions from grey conditions representing different combinations of the light and sound stimulation. These conditions can be well described by the nonlinear exclusive OR (XOR) operator. Specifically, the yellow conditions reflect low or high intensities of both stimulations, e.g., simultaneous low-intensity flash and low-pitched sound or simultaneous high-intensity flash and high-pitched sound stimulation. In turn, the grey conditions reflect the remaining combinations of sound and light: high-pitched sound and low-intensity flash or low-pitched sound and high-intensity flash. As our model is constructed from two pure selectivity neurons and one linearly mixed selectivity neuron, all conditions are sitting on a two-dimensional plane. To allow for a linear readout of the yellow vs grey mapping, a plane that separates these conditions needs to be found. This is not possible given the selectivites of the three putative neurons, as all conditions lie on one plane (**Fig. 1.1a**). However, if one cell is allowed to exhibit nonlinear mixed selectivites to both stimulation types (i.e., an interaction between light and sound), the geometry of the representation changes dramatically (**Fig. 1.1b**). More precisely, the presence of this non-linear unit expands the dimensionality of neural

representations, which means that a 2-dimensional plane can now separate the grey from the yellow conditions. This geometry can be considered 'high-dimensional', because every possible type of information (light, sound and light x sound) can be read out. It is important to note here that a linear readout serves often as a 'yardstick' of the accessibility of the information encoded by a biological or artificial population of neurons (Fusi et al., 2016). The underlying assumption is that a linear readout requires only relatively simple threshold and weighted sum operations, and thus can be implemented by a single neuron.

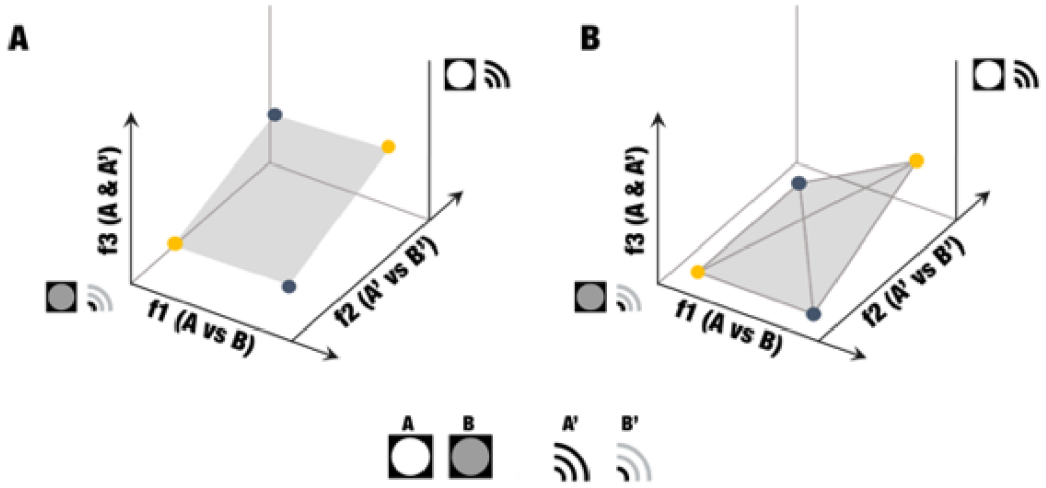


Figure 1.1. Non-linear mixed selectivity supports decoding of an XOR rule. Geometrical representations of the activity of 3 putative neurons with pure and linear mixed (a) or pure and nonlinear mixed (b) selectivity.

We can next expand our toy model to a larger neuronal population. Rather than individually handcrafting mixed selectivity profiles, we can draw connection strengths from a Gaussian distribution, creating a randomly connected population (Kaufman et al., 2022; Raposo et al., 2014). This guarantees each neuron's unique

representation of mixed variables, allowing the model to fully exploit high-dimensional representations. Now the population can act as a nonlinear kernel, projecting inputs into a high-dimensional space for enriched representation and improved separation. Theoretical work delineates the computational strengths of such neuronal reservoirs (Maass et al., 2002). Specifically, a readout neuron can integrate information from some of the reservoir neurons to approximate a task parameter that is crucial for the current behavioural goal. This approach bypasses the need to optimise local connectivity for each task, as multiple tasks can be addressed using the reservoir with distinct readout units for each task. Importantly, weight adjustments are primarily limited to the readout weights, which are few in number and thus lead to faster learning.

Empirical studies highlight the capacity of PFC cells to mix various features, facilitating the formation of high-dimensional representations (Rigotti et al., 2013). Furthermore, neural dimensionality—defined as the maximum number of possible readouts given a specific neural geometry—has shown to be predictive of task performance. Specifically, in a task that required a different response to a stimulus depending on the context, low dimensional neural representations led consequently to erroneous responses (Rigotti et al., 2013).

It is worth noting, however, that in their simplest form, reservoir computing networks struggle with tasks necessitating working memory retention (Enel et al., 2016). This is because the memory of a deactivated input is transient, as the firing rates revert to baseline post-stimulation, leading to information loss. Such

constraints undermine the motif's ability to account for human behavioural flexibility, especially given our frequent reliance on working memory. To circumvent this issue an alternative architecture is required (Enel et al., 2016).

Large theories and how to test them

What is feedback control and why do we need it?

One of the key principles underlying the machine learning framework of brain function is that complex cognitive functions generated by the brain can be explained and modelled by a multilayer system. More precisely, each of the layers performs simple computations, but their interactions give rise to emergent computations. Especially relevant in this context are concepts derived from control theory, such as observability and controllability (Kao & Hennequin, 2019). The former refers to the degree to which one can determine the internal state of the system based on its output over time. To determine the observability of a system is to examine the system's sensitivity to specific input patterns or initial conditions. This would ideally lead to the identification of the inputs that elicit the strongest responses of the readout units and thus, the most observable input directions. This idea is of particular importance when exploring the communication between regions (Kao & Hennequin, 2019). In that context, the inputs to one area that led to most observable readout responses of a second area constitute a communication subspace, whereas the input patterns that do not elicit a readout response define the first area's *private* modes (Semedo et al., 2019). Control

theory becomes, however, an especially interesting framework when explaining scenarios in which one area controls a second area. The controllability of a network can be determined by assessing which directions in the input space can be used to steer the current state of network along a specific direction in state-space (Kao & Hennequin, 2019) These ideas have recently been used to extend the architecture of the previously described reservoir computing network (Enel et al., 2016). Specifically, a reservoir network was endowed with working memory capabilities by having access to the output of a readout neuron as input to the subsequent timepoints. This allowed for an input driven attractor state to form and thus to maintain the information encoded by the readout unit, even after the original information stopped being supplied to the network.

The conceptual usefulness of control theory goes beyond improving the capabilities of reservoirs. It provides a theoretical framework for understanding how external controllers can guide activation dynamics of a separate network to solve a task, without the necessity of internal connectivity adjustments or destroying the information that is already stored in this network. Although using feedback loops or input-driven changes to the activation dynamics does not require connectivity updates, it can be regarded as learning. Compared to the previously described optimisation processes, learning through feedback control unfolds at a markedly accelerated pace.

Recent advancements have given rise to a sophisticated, systems-level modelling framework, wherein key functionalities of the prefrontal cortex (PFC), such as

working memory, are replicated through an interaction between a recurrent neural network (added to emulate PFC dynamics) and a cerebellar analogue functioning as a controller (Pemberton et al., 2022). Remarkably, this composite model demonstrated the capability to tackle a diverse range of tasks—from motor to cognitive—without necessitating the training of the RNN's recurrent connections. The study further elucidated the temporal dynamics of learning. Initial phases of learning appeared to transpire rapidly, being predominantly facilitated by the plasticity inherent in subcortical structures like the cerebellum. During these fast timescales, PFC connections remained largely unaltered, with the cerebellum capitalising only on its activation dynamics. However, as learning progressed over extended periods, the cerebellum transitioned into a different role – it served as a key component of an optimisation process that incrementally fine-tuned the recurrent weights within the PFC. Upon successful task representation transfer, the cerebellum disengaged, allowing the PFC to assume full control over the particular task. This conceptual framework, where one neural area can essentially 'upload' task-specific representations to another, offers a critical insight into the brain's capacity to coordinate processes with seemingly conflicting objectives.

System-level consolidation

Different timescales of learning in the brain can be explained by such large theories as the Complementary Learning Systems theory (CLS; McClelland et al., 1995). It posits that there are two different learning systems, each having a distinct anatomical substrate and each is responsible for generating different types of memory. The first one, implemented by the hippocampal structures, is responsible

for rapid (i.e., few shot) learning, with only a few exposures being sufficient to solve the task. The representations acquired by this fast system are highly sparse, to prevent interference. They are hypothesised to be specific - a relatively faithful copy of the experienced event. Importantly, it has been proposed that the memories maintained within the hippocampus decay relatively fast and thus that this area serves as a temporary storage module. The second learning system, implemented by the neocortex, is responsible for a slow learning process, akin to optimisation, that requires a vast number of exposures. The neural code used to store these memories is distributed, to allow for generalisation - that is, the extraction of common features among multiple different instances of experience. This format allows the neocortex to transform separate instances of experience into an abstract and rich model of the world that can be maintained over an extended period.

The theory calls these systems complementary because their interactions give rise to adaptive behaviour. When first encountering a problem, the hippocampal system can rapidly encode and solve the task. If this task becomes a stable part of the environment, the solution learnt by the hippocampus can be steadily transformed and uploaded to the cortex. The potential mechanism responsible for this transmission is based on hippocampal replay, whereby sequences of encoded experiences are rapidly reactivated during sleep or quiet wakefulness (Girardeau et al., 2009). The new and abstract memory can then be used to interpret new sensory information and thus enhance the rapid learning process.

Contemporary views on CLS expand the definition of a fast learning system, arguing that other subcortical areas like the cerebellum, basal ganglia and the thalamus can both, take part in consolidation and serve as a controller of cortical activity (Bostan & Strick, 2018; Pemberton et al., 2022; Sherman, 2016). Recent theoretical advances also elucidate why some memories stay hippocampal, while others are transferred to the cortex (Sun et al., 2023). Specifically, it has been proposed that neocortex consolidation takes places when the information encoded in the hippocampus can be applied to new but similar situations, i.e., aid generalisation. This account argues that transferring hippocampal memories that are not predictive of future behaviour can lead to overfitting, where the brain becomes too specialised in its responses to specific past experiences, reducing its ability to adapt to new or different conditions. Interestingly, this account seems in line with a perspective arguing that domain-general cognitive originate from a compositionality principle (Frankland & Greene, 2020).

Compositionality

It has been proposed that one way in which domain-general cognitive abilities can be implemented algorithmically is through capitalising on compositionality (Frankland & Greene, 2020). This rests on the assumption that most tasks can be solved by combining abstract pieces of already acquired knowledge to generate a novel solution.

It is vital to emphasize that solving a task using compositionality does not require lasting changes to the information that is stored in the agent's neural system. Instead, can be based on a short-term recombination of existing information for the

purpose of the currently performed task and followed by the relaxation of neural system to the task-agnostic state. Recent modelling efforts demonstrated that RNNs trained on multiple tasks acquire abstract dynamical motifs (Driscoll et al., 2022). The authors found that these networks acquired motifs such as attractors, discriminant boundaries and rotations through a slow learning process. These basic computational building blocks were then reused across multiple task types, e.g., in tasks that required working memory maintenance, the same attractor motif was employed. Importantly, when a new, previously unseen task was encountered, the networks were observed to dynamically transition between multiple existing motifs in an input driven fashion. A process akin to the previously described fast feedback-driven learning.

A dictionary of such abstract building blocks has been hypothesised to be either encoded in the genome or acquired through learning, depending on the complexity of the niche that an agent occupies. In situations where the complexity of the environment surpasses the memory capacity of the genetic code, a mixed strategy can be used, whereby a limited number of genetically hardwired connectivity motifs is integrated through experience into basic cognitive operations (A. M. Zador, 2019).

These considerations demonstrate that the cortex, specifically the PFC, can serve as a hub for a dictionary of abstract dynamical motifs that can be utilised by control signals stemming from separate brain areas (e.g., the cerebellum or hippocampus) and thus support learning on faster timescales. Even in situations where a new

task cannot be decomposed into basic computational operations, the PFC can be utilised by subcortical regions as a nonlinear kernel to generate rich, high-dimensional representations giving rise to fast but expensive solutions. However, over learning, the system consolidation mechanism will result in the PFC being updated by a new dynamical motif that supports an optimal, energy-efficient solution. Notably, there are other methods to produce task-specific, low-dimensional representations, including the activation of working memory processes.

Working memory is a computational resource

Working memory research has deep roots in both neuroscience and psychology. Over time, our understanding of this cognitive function has evolved from simple sensory encoding (Goldman-Rakic, 1995) to more complex transformative and preparatory roles that enhance the efficiency of upcoming behaviours (Spaak et al., 2017; Stokes et al., 2013). Historically, irrespective of the prevailing theory, working memory was often studied in isolation, with an emphasis on its capacity and the underlying neural storage mechanisms (Logie et al., 2021). However, contemporary machine learning perspectives suggest that memory should not be seen merely as passive data storage. Instead, it is viewed as playing an active role in computational processes, enhancing their efficiency (Dasgupta & Gershman, 2021).

In this context, working memory can be thought of as a process that capitalises on the space-time trade-off (Dasgupta & Gershman, 2021). More precisely, memory

can be used to store pre-computed information resulting in the reduction of computation time at the moment of decision. Conversely, by calculating information on-the-fly, one can save memory at the expense of using more computation time. Recent advances in computational neuroscience underscore this perspective, demonstrating that employing working memory changes the neural geometry in a way that supports the temporal decomposition of computations (Ehrlich & Murray, n.d.) Splitting complex computations into multiple simple classification problems that can be solved sequentially in time reduces the computational complexity at the time when a decision or action is required. This model was shown to capture human behaviour well and to reproduce neural features that are related to working memory processes such as context-dependent tuning and dynamic coding (Ehrlich & Murray, n.d.). Besides controlling neural geometry, working memory processes, can also be utilised to control neural dimensionality. This is possible by selecting features that enter working memory maintenance and are subsequently processed. For example, neurophysiological studies demonstrated that irrelevant sensory information that enters the PFC is only maintained during the stimuli presentation and fades away after stimulation is over (Jacob & Nieder, 2014).

Theoretical work suggested three types of dynamical structures that support working maintenance: fading memory, multiple point attractors, and continuous attractors. Fading memory can be observed in reservoir models as well as in models that exhibit a feedforward architecture (Goldman, 2009; Maass et al., 2002). As mentioned before, this form of memory is transient, as the firing rates will return to baseline after as specified amount of time. Thus these systems do not

constitute a good mechanism of adaptive working memory processes. In contrast, multiple attractor motifs maintain information indefinitely by capturing the neural state at one of the attractor points until this discrete information is needed for performance (Hopfield, 1984; X.-J. Wang, 2002). In turn, continuous attractors utilise a low-dimensional manifold that captures the neural state and thus can represent an arbitrary value depending on the positions within this manifold (Khona & Fiete, 2022; Seung, 1996). Both discreet and continuous attractor dynamics rely on the persistent firing of a large proportion of cells. However, recent analyses of classic experiment showing persistent activity demonstrate that this feature can be only observed in trial-averaged data (Miller et al., 2018). On a single trial level, only a few cells exhibit persistent firing during working memory delays with most cells demonstrating only sparse firing throughout this period.

This inconsistency resulted in the proposal that working memory maintenance might be partly based on short-term synaptic plasticity (Miller et al., 2018; Mongillo et al., 2008; Park et al., 2023). More precisely, connections between neurons that represent the same type of information are transiently strengthened, mimicking long-term Hebbian plasticity through the increase of calcium residuals at the presynaptic terminals (Mongillo et al., 2008) or short-term dendritic spine enlargement (Kasai et al., 2023). The maintenance of information represented in these slowly fading “impressions” does not require persistent firing rates after the encoding phase. Interestingly, feeding a non-specific input at the time of maintenance results in the transition of the content from activity silent format into persistent firing again (Mongillo et al., 2008; Stokes, 2015; Wolff et al., 2017a).

Crucially, maintenance processes resulting from short-term synaptic plasticity are energy-efficient, eliminating the need for metabolically expensive firing. Furthermore, theoretical studies have shown that short-term synaptic plasticity can foster the formation of transient attractor states without necessitating lasting changes to local connectivity (Fiebig & Lansner, 2017; Rourke & Butts, 2017). These dynamical motifs offer a means to momentarily modify the dynamical landscape of a network in response to immediate task requirements, making them a promising contender for rapid learning algorithms.

Learning shapes PFC representations

Combining CLS with control theory provides us with an intuition about how learning dynamics could evolve in the PFC. At the beginning of learning, when the agent encounters the problem for the first time, a fast generalist strategy should be employed. Depending on the similarity of the task structure to already encoded information, either a full reservoir computing algorithm or its combination with existing dynamical motifs could be utilised. In either scenario, the resulting representation will exhibit a high neural dimensionality. However, as learning advances, an optimisation-driven process gradually refines local connectivity, leading to a representation that is both task-specific and energy-efficient.

To probe this hypothesis, a particular dataset is essential. The selected task must be solvable by a range of representations, from high to low dimensionality. The XOR operation, as described earlier, is a good candidate since both reservoir and

task-optimized models can implement it (Ehrlich & Murray, n.d.). Specifically, a reservoir can project linearly non-separable input into a high-dimensional space and thus rapidly obtain the nonlinear interaction term (the XOR) necessary for task performance. The XOR can also be solved by decomposing the computations temporally, whereby one input dimensions is processed first, reducing the remaining computations at the moment of the decision to a binary choice. This strategy has been observed in task-optimised networks (Dubreuil et al., 2022; Ehrlich & Murray, n.d.). Furthermore, the experiment testing this hypothesis cannot be performed in humans, as the precise measurement of PFC representations requires an invasive recording method. However, such data can be collected using a non-human primate model such as macaques. Crucially, the neural data should be recorded with a sufficiently detailed learning resolution, starting from the primates' first task exposure up to their mastery. **Chapter 2** details a non-human primate experiment that aimed to elucidate how learning influences representations generated by the PFC.

Task-optimised representations, especially when metabolic efficiency was included in the objective used to generate them, should encode only the most informative features of the environment and thus low-dimensional. Such representations are not only efficient but also highly generalisable to novel experiences that share structural similarities with previously learned tasks (Bernardi et al., 2020; Cromer et al., 2010; Fusi et al., 2016). In **Chapter 3**, I will explore the types of representations acquired by the prefrontal cortex when introduced to a new task that shares structural elements with an already learned

task (as covered in **Chapter 2**). Specifically, I will investigate whether previously learnt information was used instantaneously to solve the new task instance. Additionally, I will explore whether the PFC represented the entire task structure in an abstract cross-generalisable format.

The geometry and dimensionality of neural representations can be also influenced by the engagement of working memory process. For example, through the selection of features that enter the maintenance phase, an agent can control the dimensionality of the representation that eventually is used to generate a decision. In **Chapter 4** I will thus explore whether the engagement of working memory can lead to representations that are optimised and energy-efficient, without slow optimisation-based learning. In contrast to **Chapters 2** and **3**, human electroencephalography (EEG) was used in the final experiment.

In the final section of this thesis, I will evaluate the empirical findings within the context of the utility of the machine learning framework for elucidating brain function. Specifically, I will scrutinise the constraints of optimisation-based frameworks for understanding neural learning mechanisms and explore alternative approaches rooted in genetically oriented perspectives. Finally, I will outline prospective directions for the collaborative advancement of neuroscience and machine learning.

Chapter 2: Learning shapes neural dimensionality in the prefrontal cortex.

Abstract

The relationship between the geometry of neural representations and the task being performed is a central question in neuroscience. The primate prefrontal cortex (PFC) is a primary focus of inquiry in this regard, as under different conditions, PFC can encode information with geometries that either rely on past experience or are experience agnostic. One hypothesis is that PFC representations should evolve with learning, from a format that supports exploration of all possible task rules to a format that minimises metabolic cost and supports generalisation. Here we test this idea by recording neural activity from PFC when learning a new rule ('XOR rule') from scratch. We show that PFC representations progress from being high dimensional and randomly mixed to low dimensional and rule selective, consistent with predictions from metabolically constrained optimised neural networks. These results show that previously conflicting accounts of PFC representations can be reconciled by considering the adaptation of these representations across learning in the service of metabolic efficiency and generalisation.

Introduction

Two seemingly discrepant accounts propose that PFC neural activity should track either low- (Cromer et al., 2010; Duncan, 2001; Ehrlich & Murray, n.d.; Flesch et al., 2022; Hirokawa et al., 2019; Reinert et al., 2021; Roy et al., 2010) or high-dimensional (Enel et al., 2016; Fusi et al., 2016; Maass et al., 2002; Rigotti et al., 2013) representations of the environment. Traditionally, it has been proposed that PFC cells are tuned adaptively to task-relevant information, leading to low-dimensional neural activity (Duncan, 2001). This results in the population displaying structured selectivity patterns, as commonly observed after training on a cognitive task (**Fig. 2.1a**, *low-dimensional*) (Duncan, 2001). A contrasting hypothesis suggests that the PFC may rely on high-dimensional, randomly mixed representations of task features to support complex cognition (**Fig. 2.1a**, *high-dimensional*) (Fusi et al., 2016; Rigotti et al., 2013). According to this notion, the PFC serves as a nonlinear kernel such that when a low-dimensional input is projected onto it, dimensionality expands, and a wide repertoire of responses can be generated (Enel et al., 2016; Maass et al., 2002).

Recently, it has been proposed that the PFC is capable of transitioning between high- and low-dimensional representations across learning, to accommodate the changing demands of the environment (Badre et al., 2021; Barak et al., 2013; Musslick et al., n.d.; Musslick & Cohen, 2021). For example, early in learning, high-dimensional representations may allow flexible exploration of all possible input–output mappings (“contingencies”) in order to discriminate which task rules are currently relevant (Enel et al., 2016; Fusi et al., 2016; Rigotti et al., 2013). This is

because a high-dimensional representation allows for a high number of linearly separable task features (**Fig. 2.1a**). Conversely, once an animal has learnt that only one set of contingencies is relevant, a low-dimensional representation may be used to minimise energy expenditure(Badre et al., 2021; Barak et al., 2013; Ehrlich & Murray, n.d.; Musslick & Cohen, 2021). In other words, different stages of learning impose different demands on the neural population. Learning could thus shape neural dimensionality and progressively push neural activity towards different solutions along the trade-off between discriminability and energy-efficiency, i.e., from a high-dimensional random regime towards a low-dimensional structured regime (Musslick et al., n.d.; Musslick & Cohen, 2021).

Recent connectionist approaches propose that learning in neural populations is the outcome of optimising an objective function, which is often subject to constraints(Hennig et al., 2021; Laughlin, 2001; Richards et al., 2019; A. Zador et al., 2023). In this context, the minimisation of energy consumption is a critical consideration that can influence neural dimensionality and its structure(Kao et al., 2021; Sussillo et al., 2015; Whittington et al., 2022). Specifically, under these constraints, neural systems are incentivised to find the most efficient representation of information by constructing a small number of highly informative dimensions. This results in a progressive elimination of neural dimensions that do not contribute to performance over the course of learning.

Here, we tested this idea in two macaque monkeys which learnt an exclusive-or (XOR) rule – a problem that can be solved by a range of representations, from low-

to high-dimensional (**Fig. 2.1a**) (Ehrlich & Murray, n.d.). Importantly, we tracked how the dimensionality and geometry of PFC representations changed across multiple training sessions of an XOR rule that was entirely new to the animals at the start of recording. We used a classical conditioning paradigm in which the nonlinear combination of the features of two objects presented in succession (XOR) predicted the outcome of the trial. Importantly, the animals were only required to fixate through both experiments. This classical conditioning design carries advantages relative to operant tasks in that changes in representation with learning are unlikely to be confounded with changes in behaviour(Asaad et al., 1998; Aston-Jones & Cohen, 2005; di Pellegrino & Wise, 1993; M. Watanabe, 1986). Later, we also show that our results hold in a previously collected delayed match-to-sample task (Constantinidis et al., 2016; Meyer et al., 2011; Qi et al., 2011).

We found that initially PFC selectivity was high-dimensional and randomly mixed with strongly linear representations. As learning progressed, PFC activity became low-dimensional, manifesting a non-linear but highly structured selectivity pattern across the neural population. We compared these results against predictions from both normatively defined generative models and optimised artificial neural networks. Collectively, these models provide a theoretical foundation for why representations should change with learning — namely that PFC representations optimise for task performance while also contending with metabolic costs.

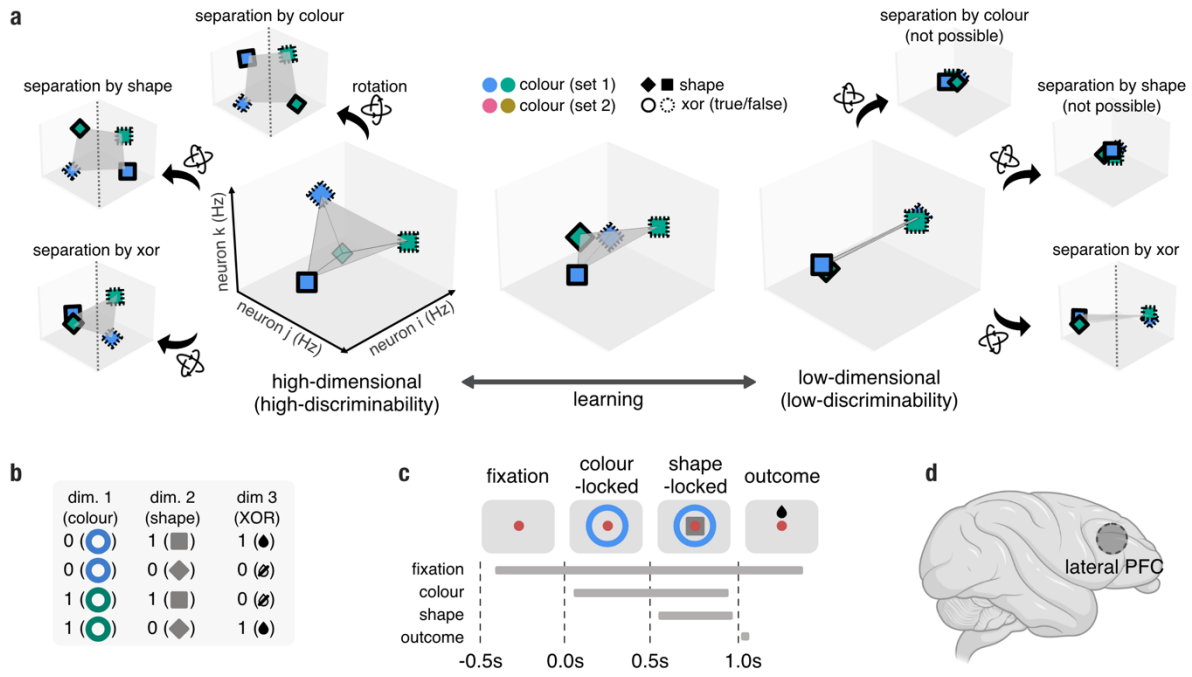


Figure 2.1. Potential effects of learning on neural geometry in the prefrontal cortex. Learning can reduce or expand neural dimensionality, changing how many linear decoding axes can be implemented on neural firing rates (discriminability). **a**, High-dimensional representations enable high discriminability. A high-dimensional regime allows the separation of all task features using three possible readout axes (left), whereas a low-dimensional representation only allows task-relevant features to be separated (right). **b**, Animals were incentivised to combine two passively viewed task features (colour and shape) in a non-linear fashion (XOR). For example, blue+square and green+diamond combinations were rewarded, whereas blue+diamond and green+square were not. **c**, Timeline of task events in a single trial. **d**, Neural data was collected from the lateral surface of the prefrontal cortex in two macaque monkeys (see *Methods* and **Supplementary Figure 5** for further details).

Methods

Animals and task

Two adult male rhesus macaques, monkey 1 and monkey 2, were trained in this study. The experiments were conducted in line with the Animals (Scientific Procedures) Act 1986 of the UK and licensed by a Home Office Project License obtained after review by Oxford University's Animal Care and Ethical Review committee. The procedures followed the standards set out in the European Community for the care and use of laboratory animals (EUVD, European Union directive 86/609/EEC). The animals were seated in a sound- and lighting-attenuated experimental booth. Their heads were restrained and faced a 19-inch screen. The centre of the screen was aligned with a neutral eye position. The animals performed a passive object-association task (**Fig. 2.1b,c**). Importantly, the animals were accustomed to an experimental setting but had no previous exposure to the task or stimuli introduced in this protocol. Neural recordings were collected from the first session as one of the main aims of the study was to capture learning dynamics. The animals were presented with a colour and a shape, a nonlinear combination of which predicted reward (**Fig. 2.1b,c**). The colours used in the coloured circles were designed in the CIELab colour space(Tkalčič & Tasič, n.d.). The L parameter (luminance) was kept constant for every colour which ensured that the stimuli where approximately isoluminant; parameters a and b varied with regard to valence but not value which resulted in a circular colour representation(Tkalčič & Tasič, n.d.). Colours were randomly assigned to conditions for each animal. Additionally, the second object had two features: one

relevant for reward prediction (shape) and one irrelevant (width) (for the duration and sequence in which stimuli were presented see **Fig. 2.1d**) The trial sequence was randomised. All trials with fixation errors were excluded. The dataset contained on average 237.9 ($SD = 23.9$) and 104.8 ($SD = 2.3$) trials for each of the 8 conditions, for monkey 1 and monkey 2, respectively.

Data acquisition and pre-processing

Before the start of the experimental protocol, a titanium head holder with two recording chambers was placed and fixed with stainless steel screws in each animal. The frontal recording chambers were implanted over the lateral prefrontal cortex (lPFC) of the right hemisphere in both animals. Data from a second chamber targeting inferotemporal cortex in the right hemisphere are not considered here. A craniotomy was made beneath each chamber to enable electrophysiological recording. Recording locations for each animal are shown in **Supplementary Figure 5**. Surgical procedures were carried out under general anaesthesia and were aseptic. A semi-chronic micro-drive system (SC-96, Gray Matter Research) with 1.5 mm interelectrode spacing, interfaced to a multichannel data acquisition system (Cerebus System, Blackrock Microsystems) was used for frontal recordings. Data were recorded over a total of 17 sessions in monkey 1 and 10 sessions for monkey 2. Notably, electrodes were manually advanced by a minimum of $62.5\ \mu m$ before every session to ensure that activity from new cells was recorded. Our setup did not allow for the same neurons to be sampled over days of training due the impact of cell necrosis and movement-related shift of electrode positions. Neural activity was amplified, filtered ($300\ Hz - 10\ kHz$), and

stored for offline pre-processing and analysis. Cluster separation was applied (valley seeking algorithm), and the binary spike train was smoothed using a Gaussian window ($\sigma = 50ms$). We collected spiking activity from 146 and 230 neurons for monkey 1 and monkey 2, respectively. Only cells sampled from the ventral and dorsal lateral frontal cortex were included in the data (**Supp. Fig. 6**). No neurons were excluded based on their selectivity profiles. Importantly, as the focus of this study was to track how learning influenced neural geometry and not the magnitude of firing rate (e.g., repetition suppression effects), we z-scored firing rates of each neuron across the whole session. The obtained firing rate data were then epoched from $200\,ms$ before to $1200\,ms$ after the colour onset. Next, all sessions were divided into four learning stages separately for each animal; the sessions within each stage were pooled across animals. All subsequent analyses were implemented in Python using custom-written code and run on combined data (monkey 1 and 2). Two types of analyses were used in this study: (1) timepoint-resolved, where a specific method was applied to every time point in the epoch to track how representations evolved in trial time, and (2) time-averaged, where a method was run on time-averaged data (i.e., $[\mathrm{t}_{400ms}, \mathrm{t}_{500ms}]$, shape-locked) in the time window preceding the trial outcome. There, we examined whether neural geometry reflected the colour and the shape and their combination (XOR) before the animals received feedback about the value of the trial.

Constantinidis et al. 2016 dataset

We also used an existing dataset of electrophysiological recordings (Constantinidis et al., 2016) which have been described in detail previously (Meyer et al., 2011; Qi

et al., 2011). In brief, neural activity was recorded from the ventral and dorsal lateral PFC (identical to the areas targeted in this study) in four rhesus monkeys who performed a feature match-to-sample task. More specifically, the animals were required to report after a delay period whether the shape of the first stimulus was the same as the shape of the second stimulus. Note that a match/no-match rule is equivalent to an XOR rule. Importantly, neural activity was recorded before the animals were exposed to the task rule (passive viewing) and after they had learned the rule. As both correct match and correct no-match trials were rewarded, the match/no-match signal was not confounded with a reward prediction signal. To test whether neural activity was pushed towards a minimal regime in such experimental conditions we employed the same decoding and selectivity measures as used in the analysis of our dataset (see **Fig 2.2**). We examined neural data averaged across the presentation of the second stimulus and the subsequent delay period ($[\mathrm{t}_{0ms}, \mathrm{t}_{2000ms}]$; stimulus 2-locked). Furthermore, neural activity was analysed for all stimulus pairs combined. For the 8 stimuli, we paired them into 4 sets of pairs and performed our analyses separately on each pair of stimuli (and averaged results over all 4 pairs) so that chance decoding was the same as in our dataset (i.e., 50%).

Multiple linear regression

We can model the firing rate $\boldsymbol{r}$ of a neuron (either from our generative models or our data) at a given time point as a linear combination of the three main task variables: colour, shape and the interaction between colour and shape (the XOR term):

$$\boldsymbol{r} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{r}$ is a vector of 1x K dimensionality containing the time-averaged firing rates for K trials; $\mathbf{X}$ is the design matrix of dimensionality K x D where rows correspond to the K trials and columns correspond to the value of the D task variables such as colour, shape and XOR ($\text{D} = 3$) in each trial. $\boldsymbol{\beta}$ is a D-by-1 vector populated with the coefficients for each of the task variable estimated for the nth neuron. Finally, $\boldsymbol{\epsilon}$ contains K residuals. The $\boldsymbol{\beta}$ vector specifies the coordinates of the nth neuron in the selectivity space spanned by D task variables (**Fig. 2.2a,b** and **Fig. 2.3b,c**). That is, every neuron can be represented as a point in a space where each axis corresponds to the cell's selectivity for a task variable.

Generative models

Neural selectivity can be defined by the matrix $\mathbf{S}_{data} = (\text{S}_{nd})_{1 \leq n \leq N, 1 \leq d \leq D}$, where the nth row corresponds to the nth neuron, and columns correspond to the regression coefficients for the three task variables colour, shape, and colour x shape (XOR) that form the axes of the considered space (Kaufman et al., 2022). This cloud of points is then centred by removing the mean ($\sum_d S_{nd} = 0$, for each of the D task variables). Here, we explored two types of selectivity distributions and their representational properties. Firstly, we examined a random mixed selectivity model in which selectivities are captured by a spherical multivariate Gaussian distribution $\mathbf{S}_{random} \sim \mathcal{N}_d(\mathbf{0}, \sigma^2 \mathbf{I}_d)$. In such a model, all variables can be decoded equally well from the population resulting in a high-dimensional representation and there is poor

cross-generalisation between variables. The second selectivity model we examine results from a system performing the XOR task while being constrained to exhibit low overall firing rates (i.e., a form of metabolic cost). We derived analytically that maximising XOR decodability while minimising such a metabolic cost results in neurons being selective only to the interaction term (colour x shape; XOR) and having no selectivity to the linear terms (colour or shape; **Supp. Materials Section 1**). A matrix describing the selectivity of such a population can be thus formulated as $\mathbf{S}_{minimal} \sim \mathcal{N}_d\left(\mathbf{0}, \sigma^2 \mathrm{diag}(0,0,1)\right)$. We call this the minimal XOR model. Importantly, to allow comparisons between the observed selectivity and model selectivity (minimal or random), the generative models were constructed using parameters derived from the data. Specifically, we used the mean value of diagonal entries of the covariance matrix $\widetilde{\mathbf{\Sigma}}_{data}$ estimated from $\mathbf{S}_{data}$ to set the value of the variance parameter σ^2 in both $\mathbf{S}_{random}$ and $\mathbf{S}_{minimal}$ (note that this ensures that both models have the same total variance). Furthermore, we showed that multiple linear regression was able to recover the underlying minimal and random models from artificially generated firing rates under various levels of noise (**Supp. Fig. 2**).

Optimised feedforward networks

We used $N = 400$ neurons in these networks and their firing rates were described by eq. 1 with σ = 2. The output $\boldsymbol{z}$ of these networks was given by a softmax readout

$$\boldsymbol{z} = \mathrm{Softmax}(\mathbf{W}_{\mathrm{out}}\boldsymbol{r} + \boldsymbol{b}),$$

(2)

where $\mathbf{W}_{\text{out}}$ are the two sets of readout weights and $\boldsymbol{b}$ is the readout bias. We optimized these networks with back-propagation using a canonical cross-entropy cost function

$$\mathcal{L} = \mathcal{H}(\boldsymbol{p}, \boldsymbol{z}) + \frac{\lambda}{2N} \|\boldsymbol{r}\|_2^2, \tag{3}$$

where the first part of eq. 3 denotes the cross-entropy loss $\mathcal{H}(\boldsymbol{p}, \boldsymbol{z})$ between the true probabilities of reward $\boldsymbol{p}$ (which were equal to 0 or 1, depending upon the stimuli for that trial) and the model's readout probabilities $\boldsymbol{z}$ and the second term corresponds to a metabolic cost on all firing rates. Before training, the values of the $\boldsymbol{\beta}$'s were drawn randomly from a Gaussian distribution with 0 mean and variance 0.05, and elements of $\mathbf{W}_{\text{out}}$ and $\boldsymbol{b}$ were set to 0. We trained two types of models: (1) with no regularisation ($\lambda = 0$) and (2) with high regularisation ($\lambda = 200$; **Supp. Fig. 1a** and **1b**, respectively). The training procedure was run using a learning rate equal to 0.005, 2000 traning steps and the Adam optimiser (Kingma & Ba, 2015). In line with our predictions, selectivity to task variables in models trained with high regularisation converged on the minimal selectivity regime whereas models trained with no regularisation produced randomly mixed selectivity (**Supp. Fig. 1**).

Decoding

To test what information was represented by the observed neural population as a function of learning, we employed linear SVM decoding (Bernardi et al., 2020; Rigotti et al., 2013). In contrast to the regression analysis, where the regression coefficients could be estimated for every neuron separately, a decoding analysis

required the construction of pseudo neural populations. To achieve this, sessions were divided into four learning stages. As each session had a different number of trials, the sessions were aligned to the last trial (to maximise learning effects) and the same number of trials per condition was used to construct the new pseudo-population for each learning stage (e.g. at least 99 trials per condition were selected from each of the sessions). This resulted in a matrix $\mathbf{X} = (\mathrm{X}_{knt})_{1\leq k\leq K, 1\leq n\leq N, 1\leq t\leq T}$ for each of the four learning stages, where the first dimension corresponds to K trials, the second dimension corresponds to N neurons (combined from two animals), and the last dimension corresponds to T time points. Then binary SVM classifiers were used to decode the task variables (colour, shape, width and XOR) at each time point for temporally resolved decoding (**Supp. Fig. 3a-d** and **Supp. Fig. 5a,b**) and from time-averaged firing rates ($[\mathrm{t}_{-100ms}, \mathrm{t}_{0ms}]$; outcome-locked, **Fig. 2.3f)**. An equivalent decoding procedure was used when analysing generative models (**Fig. 2.2f**). Decoding was performed in a cross-validated way where K trials were split randomly into set 1 and set 2, with each containing 50% of trials. The decoder was fitted using the set 1 and tested on set 2. The procedure was then repeated using set 2 as the training set and set 1 as the test set. Both decoding scores were then averaged. This procedure was repeated 10 times for different random splits of trials in sets 1 and 2, and these 10 resulting scores were then averaged. To estimate shattering dimensionality (Bernardi et al., 2020a; Rigotti et al., 2013; **Fig. 2.3h**), we used the same decoding approach except that we averaged decoding scores over all 35 possible dichotomies that could be theoretically represented given the task structure (i.e., three linear variables form a cube in state space that can be dissected into two

sets of 4 vertices in 35 possible ways). We estimated the decodability of each of these dichotomies and tracked their mean decoding accuracy as a function of learning. To assess learning-induced changes to the linear and nonlinear components of neural dimensionality separately, we split the dichotomies into linear ($\mathrm{n} = 7$) and nonlinear decoding axes ($\mathrm{n} = 28$) and tracked their mean decoding accuracy as a function of learning (**Fig. 2.3i**).

Cross-generalised decoding

To examine the neural geometry of the task representation we used cross-generalised SVM decoding (Bernardi et al., 2020). Trials were divided according to the to-be-decoded target variable labels as well as the splitting variable labels. More specifically, the splitting variable provides two instances of data which can be used to train and test the target variable classifier, respectively. In contrast to a typical cross-validation procedure, the testing happens not only on trials that were previously not seen but also on trials that correspond to different conditions. To achieve a high cross-generalisation score, it is therefore not sufficient to generalise across trial-wise noise but also to generalise across conditions (Bernardi et al., 2020). As the task examined here has three linear variables (colour, shape, width), 36 possible cross-generalisation axes can be identified (i.e., combinations of decoded variable and variable over which decoding was cross-generalised) per each of the 35 target variables. That is, the labels for 8 conditions ($2\ colours\ x\ 2\ shapes\ x\ 2\ widths$) are split into two sets of four labels (by colour, shape, or width). To identify training and testing exemplars within each of the sets, every possible binary combination of condition labels was determined ($4\ choose\ 2$).

This yielded 6 combinations for each of the sets. Thirty-six unique train-test splits (cross-generalisation axes) can be achieved by combining these two sets. Decoding accuracy scores estimated by running linear SVM decoders for each of these splits were then averaged to obtain a general cross-generalisation score of the analysed target variable.

Measuring similarity between selectivity distributions

To test the observed neural population for the presence of random mixed or minimal selectivity (**Fig. 2.2c,d,** and **Fig. 2.3d,e**), we firstly obtained regression coefficients for the three variables of interest (colour, shape and XOR; eq. 1) and constructed the selectivity space $\mathbf{S}_{data}$. To assess the similarity of $\mathbf{S}_{data}$ to $\mathbf{S}_{minimal}$ and $\mathbf{S}_{random}$, we computed the covariance matrix of $\mathbf{S}_{data}$ ($\widetilde{\mathbf{\Sigma}}_{data}$) as well as the covariance matrices of the expected random and minimal distributions given $\mathbf{S}_{data}$ ($\widetilde{\mathbf{\Sigma}}_{minimal}$ and $\widetilde{\mathbf{\Sigma}}_{random}$; see *Generative models)*. Finally, we calculated the normalised distance of the observed selectivity from model random selectivity

$$d_{i\,from\,random}(\widetilde{\mathbf{\Sigma}}) = \frac{\|\widetilde{\mathbf{\Sigma}}-\widetilde{\mathbf{\Sigma}}_{i_{random}}\|_2 - \mathbb{E}(\|\widetilde{\mathbf{\Sigma}}_{i_{random}}-\widetilde{\mathbf{\Sigma}}_{i_{random}}\|_2)}{\mathbb{E}(\|\widetilde{\mathbf{\Sigma}}_{i_{random}}-\widetilde{\mathbf{\Sigma}}_{i_{minimal}}\|_2) - \mathbb{E}(\|\widetilde{\mathbf{\Sigma}}_{i_{random}}-\widetilde{\mathbf{\Sigma}}_{i_{random}}\|_2)}, \tag{4}$$

and the normalised distance of the observed selectivity to minimal selectivity

$$d_{i\,from\,minimal}(\widetilde{\mathbf{\Sigma}}) = \frac{\|\widetilde{\mathbf{\Sigma}}-\widetilde{\mathbf{\Sigma}}_{i_{minimal}}\|_2 - \mathbb{E}(\|\widetilde{\mathbf{\Sigma}}_{i_{minimal}}-\widetilde{\mathbf{\Sigma}}_{i_{minimal}}\|_2)}{\mathbb{E}(\|\widetilde{\mathbf{\Sigma}}_{i_{random}}-\widetilde{\mathbf{\Sigma}}_{i_{minimal}}\|_2) - \mathbb{E}(\|\widetilde{\mathbf{\Sigma}}_{i_{minimal}}-\widetilde{\mathbf{\Sigma}}_{i_{minimal}}\|_2)}. \tag{5}$$

where the subscript i denotes a random draw and the expectations were computed over 1000 draws. From both the denominators and numerators, the distance within each of the models was subtracted to centre the measure around 0. More specifically, $\mathbb{E}(\left\|\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{i}_{random}} - \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{i}_{random}}\right\|_{\mathbf{2}})$(the expected difference between two different randomly drawn selectivity distributions from the random model) was, for example, subtracted from the denominator and numerator of $d_{from\,random}$ to account for within model distance. Additionally, both $d_{from\,random}$ and $d_{from\,minimal}$ were normalised by the distance between selectivities generated using both generative models ($\mathbb{E}(\left\|\widetilde{\boldsymbol{\Sigma}}_{i_{random}} - \widetilde{\boldsymbol{\Sigma}}_{i_{minimal}}\right\|_{\mathbf{2}})$) which resulted in the metrics being bounded between 0 and 1 (when $\left\|\widetilde{\boldsymbol{\Sigma}}_{data} - \widetilde{\boldsymbol{\Sigma}}_{minimal}\right\|_{\mathbf{2}}$ is equal or smaller than $\left\|\widetilde{\boldsymbol{\Sigma}}_{random} - \widetilde{\boldsymbol{\Sigma}}_{minimal}\right\|_{\mathbf{2}}$). This was done to allow for a comparison of similarity estimates across learning stages. The Euclidean distance metric was chosen as the main analysis tool in this study based on simulations in which we generated different proportions of random and minimal selectivity across a single population (from 0% minimal and 100% of random to 100% minimal and 0% random) and compared the precision with which multiple metrics recovered the true proportions. We compared the Euclidean distance metric to the PAIRs metric, which has been used previously in the literature(Hirokawa et al., 2019; Raposo et al., 2014), and to the symmetric Kullback–Leibler divergence estimate (KL divergence) which benefits from a strong theoretical basis and is assumption-agnostic. We found that, compared to the KL divergence and PAIRs metrics, the Euclidean distance measure yielded the highest precision of tracking learning-induced changes to neural selectivity (**Supp. Fig. 7a,b**). Specifically, our simulations showed that both the KL divergence and PAIRs can be used to precisely identify extreme selectivity

regimes (either strong random selectivity or strong minimal selectivity) but fail at identifying intermediate selectivity regimes showing a strong bias towards random mixed selectivity (**Supp. Fig. 7a,b**). As the focus of this study was to track learning dynamics, a metric that allows to identify a broad range of selectivity regimes was chosen for the final analysis. Nonetheless, the results from experiment 1 (**Fig. 2.3d,e**) were broadly replicated using the symmetric KL divergence estimate (**Supp. Fig. 7c,d**) and PAIRs (**Supp. Fig. 7e,f**).

Principal Component Analysis

PCA was used to visualise the population firing rate for different conditions in experiment 1 (**Fig. 2.3a**) and as a measure of neural dimensionality (**Supp. Fig. 3o**). Firstly, pseudo populations were constructed for each learning stage using the same procedure as described in the *Decoding* section. Then, firing rates were averaged in the time window preceding outcome presentation ($[\mathrm{t}_{400ms}, \mathrm{t}_{500ms}]$, shape-locked; **Fig. 2.3a**). Next, principal components were computed on condition averages and the population firing rates we projected onto the top three principal components. This was done separately for each learning stage. Note that width 1 and width 2 trials were pooled together. To compute how the variance explained (ratio) by the first PC changed as a function of learning (**Supp. Fig. 3o**), trials were randomly split into test and train 10 times; PCA was fitted then on train trials and the test trial firing rates were projected onto them to compute variance explained. The results from 10 random splits were then averaged. The null distribution for the permutation test was computed by randomly shuffling neurons between stage 1 and stage 4 and repeating the described PCA procedure ($n = 500$).

Statistical testing for decoding and cross. gen. decoding

Throughout the study, we employed non-parametric permutation tests to test statistical significance within each learning stage and between learning stages (learning-induced effects). Two types of null distributions were thus constructed: (1) for statistical testing in time-resolved and time-specific decoding analyses the labels describing the trial dimension (k) of the pseudo-population matrix $\mathbf{X} = (\mathrm{X}_{knt})_{1\leq k\leq K, 1\leq n\leq N, 1\leq t\leq T}$ were randomly permuted 500 times; (2) to test for learning-induced effects, the matrices $\mathbf{X}_{stage\ 1}$ and $\mathbf{X}_{stage\ 4}$ were concatenated along the neuron dimension (n) and then 500 new $\mathbf{X'}_{stage\ 1}$ and $\mathbf{X'}_{stage\ 4}$ matrices were generated by randomly assigning neurons to either $\mathbf{X'}_{stage\ 1}$ or $\mathbf{X'}_{stage\ 4}$. One-sided tests were used when testing the predictions of the minimal model and two-sided tests were used when no differences were expected. Additionally, to control for time-related family-wise errors a cluster-based permutation correction was added to the time-resolved decoding(Maris & Oostenveld, 2007).

Statistical testing for selectivity measures

To test whether observed selectivity was dissimilar to the random selectivity regime and similar to minimal selectivity regime 1000 random and minimal models were generated using data-derived parameters for each learning stage. Next, the $d_{rand.from\,rand.}$ and $d_{rand.from\,min.}$ distances were computed for 1000 randomly generated models according to eq. 4 and eq. 5 (with $\widetilde{\mathbf{\Sigma}}_{random}$ as input) to serve as null distributions for both comparisons. Note that the observed selectivity was

compared to random model selectivity when analysing the data's similarity to random (**Fig. 2.3d**) as well as minimal selectivity (**Fig. 2.3e**).

Results

The animals were trained to combine a colour stimulus (either blue or green) with a subsequently presented shape (either square or diamond) in a nonlinear fashion to predict the reward (the XOR between colour and shape) outcome of the trial (**Fig. 2.1c-d**). We sequentially recorded 376 neurons from the lateral PFC across both macaques (F**ig. 2.1c**; see *Methods, data acquisition and pre-processing*), i.e., new neuron were sampled each session. Importantly, to capture learning dynamics, we started recording from the first session in which the animals were exposed to the stimuli. The full set of sessions were divided up into four learning stages, e.g., the first 5 sessions in monkey 1 comprised the first learning stage for monkey 1. Data was pooled over animals within each of the learning stages. All selectivity and decoding analyses were run in time windows before the animals received feedback about the outcome of the trial (i.e., reward; for details see *Methods*, *data acquisition and pre-processing*).

Generative models of random and minimal selectivity

The aim of the study was to understand how the geometry of the neural representations changed over the course of learning. For this purpose, we firstly constructed and examined two types of neural geometries (Badre et al., 2021; Barak et al., 2013): (i) a high-dimensional random geometry (Enel et al., 2016; Fusi

et al., 2016; Maass et al., 2002; Raposo et al., 2014; Rigotti et al., 2013); and (ii) a low-dimensional, structured geometry (Chaudhuri et al., 2019; Cueva et al., 2020; Flesch et al., 2022; Ganguli et al., 2008; Gao & Ganguli, 2015; Hirokawa et al., 2019; Sadtler et al., 2014). The former is a well-established geometry (e.g., inherent in reservoir computing models (Enel et al., 2016; Maass et al., 2002)) whereas the latter was motivated by recent theoretical work which suggests that complex (nonlinear) tasks may require low-dimensional, highly structured population activity (Dubreuil et al., 2022; Flesch et al., 2022). These models make distinct predictions about the distribution of selectivity to the task-relevant variables in a 3-dimensional selectivity space (colour, shape, and their interaction, i.e., XOR).

In this selectivity space, each axis represents a neuron's response to one stimulus variable (e.g., 'shape selectivity' = higher firing rate for square than diamond). Hypothetically, one could imagine different distributions of variable encoding within this selectivity space. The properties of these distributions are determined by a covariance matrix: the diagonal elements describe the strength of coding of each variable (variance) whereas the off-diagonal entries determine the strength of the relationship between variables (covariance). In our generative models neural firing rates were simply constructed as a linear combination of task variables (colour, shape and XOR inputs represented using one-hot encodings; see *Methods, generative models* for details) with a specified covariance matrix of selectivities to the task variables. In line with previous studies, the high-dimensional random mixed model was constructed by allowing selectivity to be distributed randomly according to a spherical Gaussian distribution (**Fig. 2.2a**). In contrast, in the

minimally structured XOR selectivity model, only the interaction (i.e., the XOR) term carried variance (**Fig. 2.2b**). We derived this model mathematically and demonstrated that it minimises total firing rate activity (i.e., a metabolic cost) while maximising task performance (see **Supp. Materials, section 1**). Consistent with this, we found that feedforward networks trained using backpropagation to perform the task while also minimising a metabolic cost term converge to the minimal XOR selectivity model (see **Supp. Fig. 1a-f**; *Methods, optimised feedforward networks*). Additionally, in line with prior accounts (Barak et al., 2013; Bernardi et al., 2020), XOR decoding in the minimal model was more robust to noise (**Fig. 2.2e**) and required fewer neurons to implement a stable readout (see **Supp. Fig. 1g**).

We next established a metric to measure whether neural activity is better described by the random or minimal model. We first fitted a linear model to surrogate data generated by both our generative models, in which task variables (colour, shape, and colour x shape (XOR)) were used as predictors of each unit's firing rate (see *Methods, model section;* for similar analysis see Dubreuil et al., 2022; Mante et al., 2013). Then we measured the average within-model distance between two random models (**Fig. 2.2c**, red line; *Methods, eq. 4*) and the average between-model distance between the random and the minimal model (**Fig. 2.2c**, blue line, *Methods, eq. 5*). To test whether our measure captures learning dynamics, we constructed four artificial populations with varying proportions of random and minimal selectivity. As the proportion of the minimal model in this mixed population increased, it became more dissimilar to the average random model and more

similar to the average minimal model (**Fig. 2.2c**, black line). A reflected version of these results held true when the minimal model was used as reference (**Fig. 2.2d**).

Subsequently, to gain insight into the predictions of these models, we employed an established technique (Fusi et al., 2016; Rigotti et al., 2013) and trained linear decoders to decode all three task variables in both models. As previously suggested (Fusi et al., 2016; Rigotti et al., 2013), a randomly mixed selectivity model yielded a high-dimensional task representation, allowing for all variables, including the nonlinear XOR, to be decoded (**Fig. 2.2f**, red; cf. **Fig. 2.1a**, far left). In contrast, for the minimal model, only the XOR combination of shape and colour, and not shape or colour independently, could be decoded (**Fig. 2.2f**, blue; cf. **Fig. 2.1a**, far right).

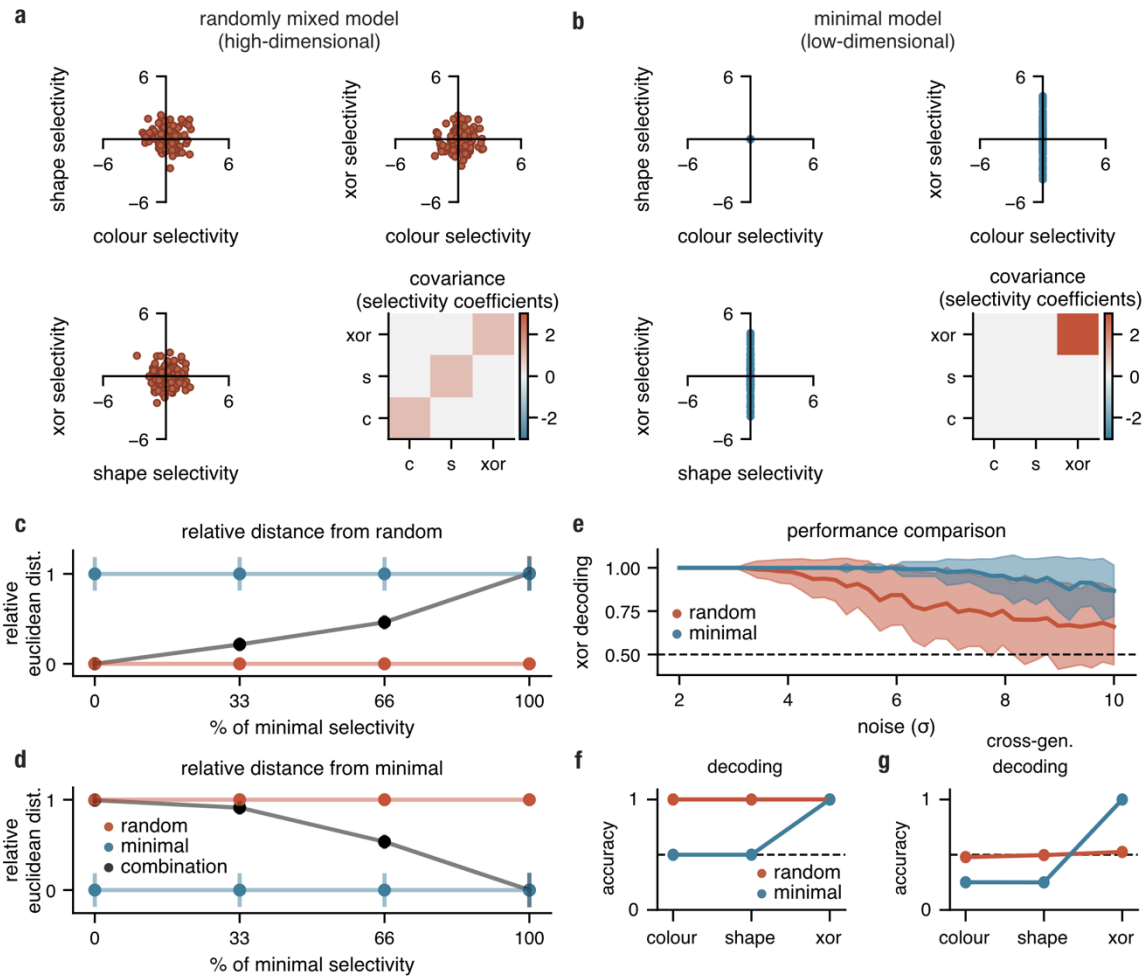


Figure 2.2. Predictions from random mixed and minimal selectivity models. **a**, Each neuron can be represented as a point in the 3-dimensional selectivity space spanned by colour, shape, and XOR (their interaction). In the random mixed model, selectivity is distributed according to a spherical Gaussian distribution in this space (*Methods, generative models*); the covariance matrix is computed between the selectivity coefficients. **b**, Analogous to **a** but for the minimal model; neurons are selective only for the XOR (interaction between colour and shape), as this is the only feature that is necessary to solve the task. **c**, Relative distance between the covariance matrices of either the random (red), minimal (blue) or the model with varying proportions of minimal selectivity (black) to the random selectivity model. **d**, Same as panel **c** but the distance is calculated relative to the

minimal selectivity model. Red and blue error bars show standard deviations ($\pm 1\ s.d.$ over 1000 randomly drawn models; see *Methods*, *measuring similarity between selectivity distributions*) of the relative Euclidean distance between the covariance matrix of the random, minimal model (with matched total variance to the data) and the covariance matrix expected from random selectivity; black error bars show the standard deviation of the relative distance between the surrogate covariance (with varying proportions of minimal selectivity) and random covariance ($\pm 1\ s.d.$ over 1000 random models). **e,** XOR decoding as a function of noise (σ) in the neural activities for the random and minimal selectivity models. Dashed grey line shows chance-level decoding. **f**, Mean (over 100 models) decoding of task variables for the random (red) and minimal (blue) models. Dashed grey line shows chance-level decoding. **g**, Mean (over 100 models) cross-generalised decoding of task variables for the random (red) and minimal (blue) models. Dashed grey line shows chance-level decoding.

While both models can perform the task, we expected their representation of the XOR variable to differ fundamentally. Specifically, the minimal model by design should represent the XOR in a format that generalises over all other task variables. This is not guaranteed in the random model. We verified this intuition using cross-generalised decoding, a method in which a linear decoder is trained to decode a given task variable (e.g., XOR = True vs. XOR = False) on a given set of task conditions (e.g., blue colour) and tested on a different set of task conditions (Bernardi et al., 2020) (e.g., green colour; see *Methods section, cross-generalised decoding*). We found that for the random model, cross-generalised decoding was

at chance-level for all task variables (**Fig. 2.2g**, red). This is because the random model, by design, exhibits no reliable structure in its representation of variables and therefore these dimensions are represented randomly in relation to each other. In contrast, the minimal model displayed maximal cross-generalised decoding for the XOR variable (**Fig. 2.2g**, blue), indicating that it can be decoded regardless of which set of task conditions the decoder is trained and tested on. This suggests that the minimal model is able to represent the XOR in a highly cross-generalisable format (**Fig. 2.1a**, far right). Consequently, the minimal model also exhibits below-chance cross-generalised decoding for colour and shape.

A low-dimensional task relevant geometry emerges over learning

We next applied these metrics to our neural data (**Fig. 2.3**). Inspecting the neural representation of task variables on the population level suggested that the geometry in the PFC changed over the course of learning (**Fig. 2.3a**) and resulted in a lower-dimensional, task-specific (XOR) representation by the end of learning (**Fig. 2.3a**, far right). To quantify this, we fitted a linear regression to our data, just as we did for our generative models (**Fig. 2.2c**), in which task variables (colour, shape, and colour x shape (XOR)) were used as predictors of each neuron's firing rate. We then examined how selectivity coefficients changed over learning (**Fig. 2.3b**, learning stage 1 and **Fig. 2.3c**, learning stage 4). We compared the covariance structure of these selectivity coefficients (**Fig. 2.3b,c,** bottom right) to the covariances obtained from the random selectivity model (**Fig. 2.3b**, red contours) and minimal selectivity model (**Fig. 2.3c**, blue line). At the beginning of learning (stage 1), PFC cells were randomly distributed in selectivity space

resembling the random model ($p = .305$, one-sided, **Fig. 2.3d** and $p = .827$, one-sided, **Fig. 2.3e**). However, in late learning (stage 4), selectivity diverged away from randomly mixed selectivity ($p = .000$, one-sided, **Fig. 2.3d**, compare black and red lines) and converged towards the minimal model ($p = .000$, one-sided, Fig. **2.3e**, compare black and blue lines).

We next applied our linear decoding analyses (**Fig. 2.2f, g**) to the neural recordings to establish whether these changes to selectivity distributions led to changes in the geometry of neural representations **(Fig. 2.3f,g**; **Supp. Fig. 3e-l)**. At the beginning of training (learning stage 1), the animals exhibited a high-dimensional geometry that allowed for every variable to be decoded, reminiscent of the random mixed model (compare **Fig. 2.2f**, red to **Fig. 2.3f,** stage 1; see **Supp. Fig. 3a-d** for temporally resolved decoding). In line with the random model, the task variables also exhibited poor cross-generalised decoding during early learning (compare **Fig. 2.2g**, red to **Fig. 2.3g,** stage 1). In late learning (stage 4), however, the representation became low-dimensional and task-specific. As predicted by our minimal model, the coding of the XOR strengthened ($p = .024$, one-sided) whereas the decoding of shape decreased ($p = .000$, one-sided; compare **Fig. 2.2f**, blue to **Fig. 2.3f,** stage 4). Moreover, the XOR variable became progressively represented in a format that generalised across the remaining task variables ($p = .000$, one-sided, compare **Fig. 2.2g**, blue to **Fig. 2.3g,** stage 4). Cross-generalised decoding for shape has not decreased over learning significantly ($p = .108$, one-sided). Unexpectedly, colour decoding remained high even in late learning (decoding: $p = .514$, one-sided; cross-decoding: $p = .846$, one-sided; **Fig. 2.2f**; **Supp. Fig. a, e**).

Finally, when we examined an additional task feature that was not predictive of reward (the width of the shape, see *Methods section, data and task*), we found a strong reduction in its representation in PFC over learning (decoding: $p = .026$, one-sided; cross. decoding: $p = .01$, one-sided; see **Supp. Fig. 3c, g, k**).

We next explored whether these learning-related changes were accompanied by a decrease in neural dimensionality (as measured with shattering dimensionality; see (Rigotti et al., 2013) for details). A neural representation described by three binary input dimensions (colour, shape and width) results in 35 dichotomies (division into two sets of four stimuli) that can be theoretically decoded. We found that the mean decoding accuracy of all dimensions (excluding colour, shape and XOR) decreased significantly over learning ($p = .044$, one-sided; **Fig. 2.3h**). Additionally, a principal component analysis computed on condition averages revealed that the proportion of variance explained by the first principal component increased as a function of learning ($p = .000$, one-sided; **Supp. Fig. 3o**; for details see *Methods, Principal component analysis*). Both the shattering dimensionality and the PCA analyses indicate that neural dimensionality decreased as a function of learning. Furthermore, as colour, shape, and width form a cube in the input space (when one-hot encoding is used) these dichotomies can be split into linear (a hyperplane can be used to divide all vertices into two sets) and nonlinear dimensions (a nonlinear function needs to be used to differentiate two sets of 4 vertices). We found that the decoding of linear dimensions decreased significantly over learning ($p = .012$, one-sided), while the decoding of non-linear dimensions did not ($p = .212$, two-sided; **Fig. 2.3i**). The progressive change over learning, not

only in the neural dimensionality but also in the type of computations that are represented (linear vs non-linear) conflicts with a static model of the PFC as a high-dimensional nonlinear kernel (e.g., such as that utilised in reservoir computing Enel et al., 2016; Maass et al., 2002).

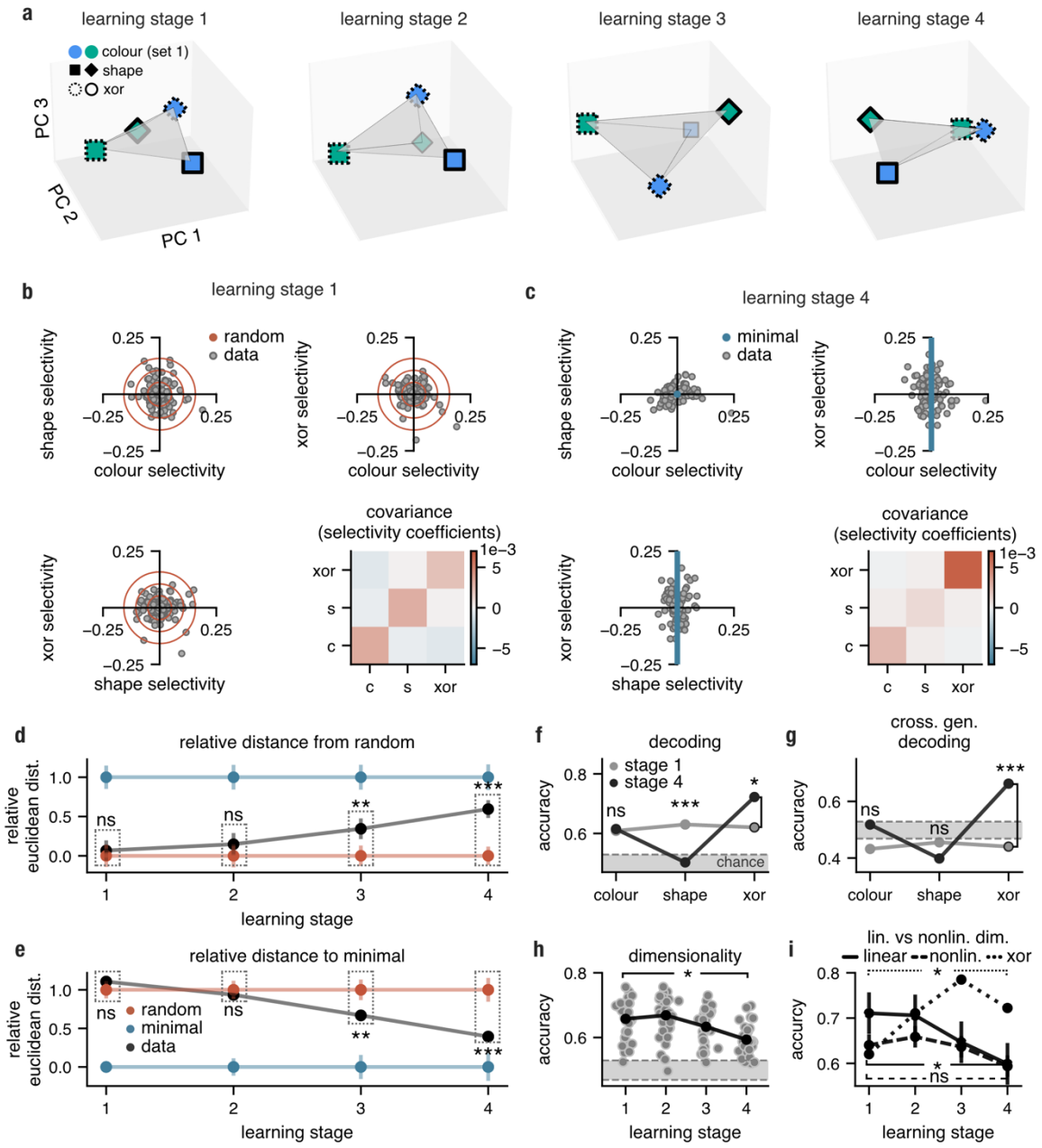


Figure 2.3. Analysis of neural representations in macaque PFC over learning.

a, Trial-averaged population firing rate during the late shape period for each task

condition projected onto the top three principal components as a function of learning (*Methods, principal component analysis*); the same scale was used for all plots. **b**,**c**, Neural selectivities in learning stages 1 and 4, respectively. Each point represents the selectivity of one neuron for each of the task variables (colour, shape and XOR). We show all 3 possible pairs of the 3 axes. In panel **b**, the 3 red rings show spherical Gaussian distributions corresponding to 1, 2, and 3 standard deviations of the data. In panel **c**, the 3 blue rings show ellipses from the minimal model that correspond to 1, 2, and 3 standard deviations of the data (note that they all lie on top of one another along the XOR selectivity axis). The bottom right panel shows the covariance matrix of the selectivities computed from the data for stage 1 of learning (**b**) and stage 4 of learning (**c**). **d**, Relative Euclidean distance between the covariance matrix of selectivity coefficients from the data and the covariance matrix expected from random selectivity (with matched total variance) as a function of learning (*Methods*, *measuring similarity between selectivity distributions*). Red and blue, respectively, error bars show mean ($\pm 1\ s.d.$ over 1000 randomly drawn models) of the relative Euclidean distance between the covariance matrix of the random, respectively minimal, model (with matched total variance to the data) and the covariance matrix expected from random selectivity; black error bars show standard deviation of relative distance between the observed covariance and random covariance ($\pm 1\ s.d.$ over 1000 random models). **e**, Same as panel d but we show the relative distance from the covariance matrix expected from minimal selectivity (with matched total variance). **f**, Linear decoding of task variables for learning stages 1 (grey) and 4 (black). **g**, Cross-generalised linear decoding of task variables for learning stages 1 (grey) and 4 (black). **h**, The learning dynamic of the

decoding of all possible task dichotomies excluding colour, shape and XOR (shattering dimensionality); solid black line indicates the mean over these dichotomies. **i**, Mean $\pm 1\ s.d.$ linear decoding of all linear dichotomies (dashed line), non-linear dichotomies (solid line), and the XOR dichotomy (dotted black) as a function of learning. All p-values were calculated from permutation tests (***, $p < 0.01$; **, $p < 0.01$; *, $p < 0.05$; †, $p < 0.1$; $n.s.$, not significant). Grey shading in panels **f**–**g** shows the mean $\pm 1\ s.d.$ of chance-level decoding obtained by shuffling trial labels (for details see *Methods, statistical testing*).

To replicate our findings, we also performed a re-analysis of an existing dataset (Constantinidis et al., 2016; Meyer et al., 2011; Qi et al., 2011) in which recordings were taken from primate ventral and dorsolateral PFC before and after learning a delayed match-to-sample task that was similar in structure to ours (for details see *Methods, existing lPFC dataset*). This task required active behaviour and the reward signal was disentangled from the XOR (i.e., match/no match) signal. We found that, as in our data, learning pushed neural activity in the PFC towards a minimal XOR regime (**Supp. Fig. 4**).

Our findings indicate that neural activity in the PFC shifts between two distinct selectivity regimes throughout the learning process. Initially, the PFC maximally expanded the representational space by encoding all available variables. Subsequently, after a combination of task variables that predicted the trial's outcome was identified, the PFC reduced its dimensionality and maintained a minimal model of the task structure.

Discussion

The prefrontal cortex has the capacity to generate both low-dimensional and high-dimensional representations. However, the conditions under which each regime is employed currently remain unclear. Our study investigated how the dimensionality and geometry of neural activity changed over learning. We observed that, as learning progressed, neural activity in the PFC transitioned from being high-dimensional with high-discriminability to being low-dimensional and abstract. This transition in the representational strategy was accompanied by a change in population structure, from random and mixed, to minimal and structured.

We found that the transition between high-dimensional and low-dimensional representations can be reproduced by a model that solves an XOR problem under metabolic cost constraints. One recent theoretical proposal (Dubreuil et al., 2021) used alternative constraints to argue that the emergence of structured selectivity is largely contingent on the complexity of the task. This was particularly evident in tasks that necessitate flexible input–output mappings, such as XOR rules. While our model explicitly defines the XOR nonlinearity as an input, this alternative model instead leveraged the temporal dimension to express the nonlinearity through recurrent dynamics (Dubreuil et al., 2021). However, the two models are not mutually exclusive, as both demonstrate that shortly before a decision is made, the system only retains relevant information (i.e., the XOR). This is accomplished by employing different forms of regularisation, such as an L2 cost on neural activities

in our model, and a rank constraint on the recurrent connectivity matrix in the alternative model (Dubreuil et al., 2021). Another prominent theoretical proposal (Bernardi et al., 2020) describes neural geometries similar to those found in our optimised networks: output variables are represented in an abstract, cross-generalisable fashion. However, one difference with this proposal is that the metabolic cost constraints in our model cause task-irrelevant variables to no longer be represented after learning (**Supp. Fig. 1h-i**); and this is consistent with the findings in our data **(Fig. 2.3h)**.

The acquisition of a low-dimensional representation following the learning of a single XOR rule raises the question of which regime the PFC might adopt when confronted with more complex tasks. Although XOR operations necessitate nonlinear integration and abstraction from sensory input, they can be reduced to simple stimulus-response pairings once the rule has been learnt or when a memory-based strategy is utilised. However, some tasks are more difficult to decompose, as they require switching between multiple orthogonal or conflicting subtasks. It has been suggested (Musslick et al., n.d.; Musslick & Cohen, 2021) that in conditions requiring the performing of multiple tasks in series, a high-dimensional representation could be employed by the PFC in order to maximise flexibility and prevent interference.

We found that the PFC acquired a low-dimensional representation to solve a complex cognitive task. This corroborates the findings of Hirokawa et al. (Hirokawa et al., 2019) who found that neural activity covaried with behaviourally relevant

variables on a single cell level, thus occupying a low-dimensional manifold. On the other hand, some studies have suggested that an increase in neural dimensionality is predictive of performance (Rigotti et al., 2013). It is possible that the structure of the task and training provides an explanation for these contrasting findings (Musslick & Cohen, 2021). In our study, recordings were initiated from the onset of task training, spanning a period of five weeks, with animals experiencing the entire task structure from the first session. In contrast, many other investigations into the primate PFC's involvement in complex cognitive tasks train animals in a fashion that decomposes the task into multiple subcomponents and either builds up task knowledge across training (Rigotti et al., 2013) or presents them in a serial (block-wise) manner (Bartolo et al., 2020; Bernardi et al., 2020). Additionally, whereas in our study, one source of information (width) was always irrelevant, in other studies, information becomes periodically relevant and irrelevant across multiple blocks, which may promote encoding of currently irrelevant information(Mante et al., 2013). These training differences may promote information encoding in a high-dimensional manner that favours discriminability over metabolic efficiency.

Although we operationalised the metabolic constraint using an L2 cost on neural firing rates, several other related formulations are possible. The energy expenditure of a population can be reduced not only through a global (population-level) reduction of firing rates but also through the reduction of the number of actively firing neurons. This would naturally correspond to an L1 (rather than L2) cost on neural firing rates, which encourages sparseness of representations. Indeed, different levels of sparseness of neural firing rates could affect whether the

representation is low- or high-dimensional and would also indirectly reduce an L2 metabolic cost on neural firing rates. However, our results indicate that the PFC likely does not rely on sparseness as a mechanism for metabolic cost reduction, as we did not observe any learning-induced changes in this measure (**Supp. Fig. 3p**). Additionally, imposing costs on the strength of synaptic connections rather than on neural firing rates would also be an interesting avenue for future work, especially when considering more realistic models with recurrent connectivities.

It is perhaps surprising that, given the key role of the PFC in the development and acquisition of structured knowledge, only a few studies have investigated how the structure of PFC representations changes during several training days of an entirely novel task (Corrigan et al., 2022; Reinert et al., 2021). By tracking changes in neural activity across learning, it is possible to identify the biological principles that are required to produce representations supporting higher cognitive functions (Richards et al., 2019). In this study, we have demonstrated that the constraint of metabolic cost reduction parsimoniously describes the changes in neural selectivity and geometry in non-human primates learning a complex cognitive rule. This provides a foundation for explaining how PFC representations should change with learning. Future experiments should extend this paradigm, to track changes in learning even more complex and naturalistic tasks (Hunt et al., 2021); those that have a compositional structure (Driscoll et al., 2022; Yang et al., 2019) the influence of different learning curricula(Dekker et al., 2022); and how these representations change within the same individual neurons as opposed to pseudo populations (Steinmetz et al., 2021; Tolias et al., 2007; Zhao et al., 2023).

Chapter 3: Low-dimensional representations support cross-generalisation

Abstract

Information stored within the PFC is believed to serve a distinct role in adaptation, contrasting with other memory types. In essence, cortical memories are posited to capture commonalities across multiple specific experiences, thereby reflecting the environment's abstract structure. We tested this hypothesis using a non-human primate model. Two macaques used the representations acquired in a prior experiment (Experiment 1; **Chapter 2**) to facilitate the learning of an XOR rule within a new stimulus set. Our findings indicate that, during early learning stages, the animals used the existing XOR readout to categorize both novel and familiar stimulus combinations. However, as learning advanced, alignment extended to other task features. Consequently, the XOR problem was solved through a dynamic neural landscape, capitalising on an initial contextual signal to process the subsequent task features in a state-dependent manner. The distinct learning rates of generalisation between the XOR and the remaining task features implies that such a dynamic landscape emerged uniquely in Experiment 2.

Introduction

The ability to leverage stored information to enhance the learning of new tasks is fundamental for adaptability. Within the primate brain, two primary types of memories are recognized: those housed within the hippocampus and those in the neocortex (McClelland et al., 1995; Samborska et al., 2022; Sun et al., 2023). Hippocampal memories, believed to be precise reflections of past events, act as a transient storage mechanism due to their hypothesized rapid decay rate. Conversely, neocortical memories emerge from a gradual learning process that resembles optimisation, necessitating numerous exposures. The neural encoding of these memories is distributed, facilitating generalisation.

It has been proposed that neocortical memories are constructed by relaying only specific types of information from the hippocampus to the neocortex via a systems consolidation mechanism (McClelland et al., 1995). Only memories that are predictive of future events, and thus support generalisation are transferred to the lasting storage of the neocortex (Sun et al., 2023). The consequences of this process for PFC representations are well described by a discriminability-generalisability trade-off (Badre et al., 2021; Barak et al., 2013). At the beginning of learning, the PFC connections do not reflect any prior knowledge of the new task and thus generate high-dimensional representations that support flexibility (Rigotti et al., 2013; left, Fig. 1a). However, the specificity of these representations poses challenges for cross-generalisation to novel stimuli (*high discriminability*, *low generalisability*; left, **Fig. 3.1a**). As learning progresses, the representations within the PFC become optimised to perform the task and thus low-dimensional (**Chapter**

2; Wójcik et al., n.d.). This condensed representation is then poised for efficient adaptation to previously unseen problems (right: *low discriminability*, *high generalisability*; **Fig. 3.1a,b;** (Bernardi et al., 2020; Cromer et al., 2010).

Our previous research in non-human primates (**Chapter 2**; Wójcik et al., n.d.) demonstrated that PFC representations evolve from high-dimensional to low-dimensional across learning in the service of metabolic efficiency. This learning trajectory resonates with established theories of neocortical memory formation (McClelland et al., 1995; Sun et al., 2023). In line with the discriminability-generalisability trade-off (Badre et al., 2021; Barak et al., 2013) as well as the abstract nature of neocortical memories (McClelland et al., 1995; Sun et al., 2023), these low-dimensional representations should exhibit a capacity for cross-generalisation to new sensory instances that share the same abstract structure.

We examined this hypothesis using two macaque monkeys trained to cross-generalise a previously learned exclusive-or (XOR) rule to new stimuli. Notably, we monitored shifts in the dimensionality and geometry of PFC representations over numerous training sessions for both the new and old XOR problems. Our methodology incorporated a classical conditioning paradigm, where a non-linear combination of successive object features (colour and shape) determined the trial's outcome (XOR). The monkeys were solely tasked with fixation throughout both experiments. Employing this classical conditioning approach, as opposed to operant tasks, ensures that representational changes during learning were not

conflated with behaviourally driven neural activity (Asaad et al., 1998; Aston-Jones & Cohen, 2005; di Pellegrino & Wise, 1993; M. Watanabe, 1986).

We found that the PFC aligned the new XOR representation to the old XOR readout early in training, allowing for the same readout to be used to solve both instances of the task. Furthermore, we observed that the animals used the feature presented first (colour) to generate an abstract context signal before the presentation of remaining task features. This suggests that the PFC occupied a preparatory state that permits the immediate, context-appropriate transformation of the remaining task features to generate an XOR signal. In essence, the PFC employed the same abstract dynamical landscape for both XOR rules, echoing optimal models previously proposed for context-dependent computations (Dubreuil et al., 2022).

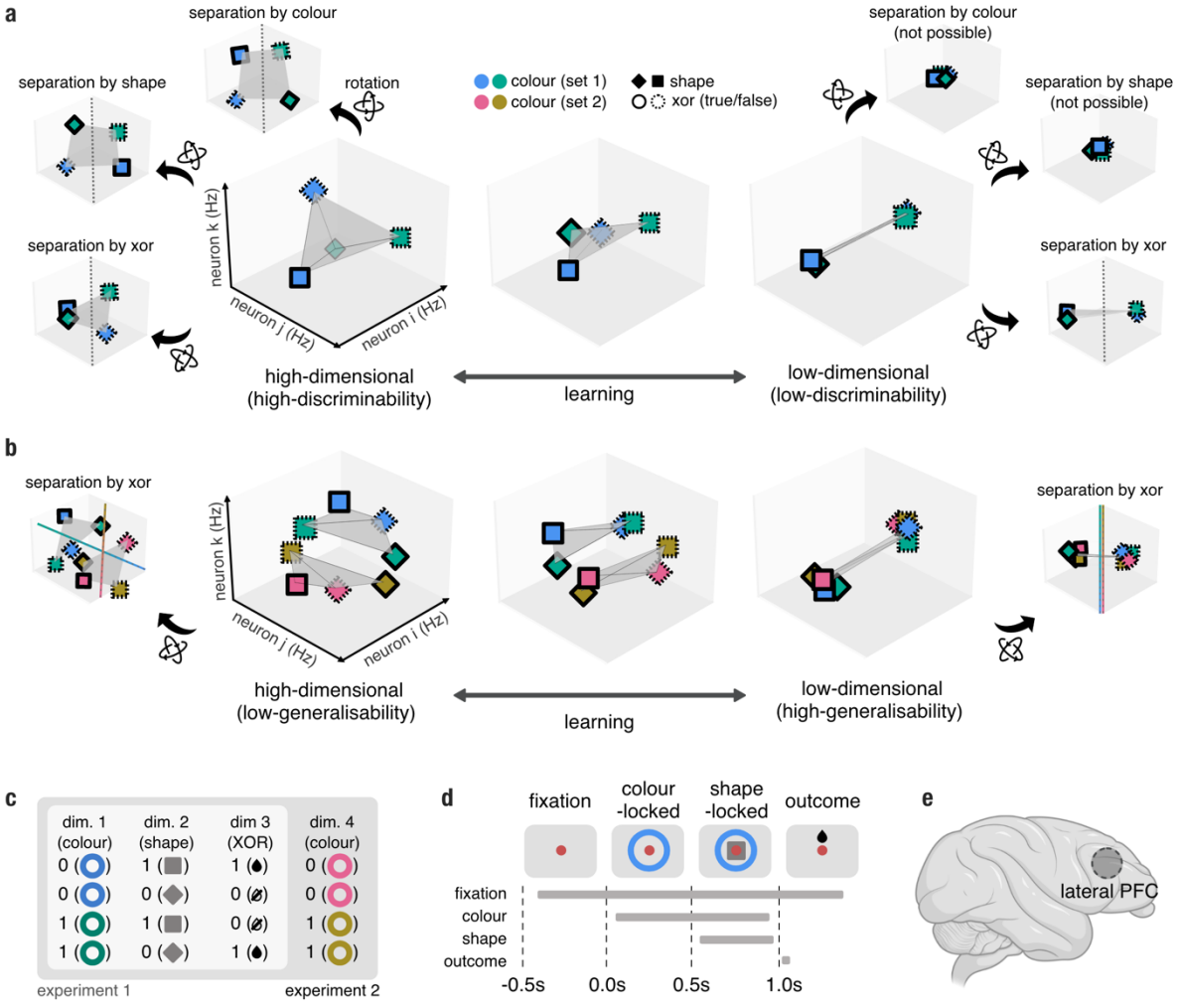


Figure 3.1. Potential effects of learning on neural geometry in the prefrontal cortex. Learning can reduce or expand neural dimensionality, changing how many linear decoding axes can be implemented on neural firing rates (discriminability). **a**, High-dimensional representations enable high discriminability. A high-dimensional regime allows the separation of all task features using three possible readout axes (left), whereas a low-dimensional representation only allows task-relevant features to be separated (right). **b**, Low-dimensional representations enable high generalisability. When the task representation is high-dimensional (left), XOR discriminant from one task (blue and green shapes) fails to correctly

differentiate the XOR feature in the second task (pink and khaki); i.e., a rotation that aligns both XOR discriminant axes is not possible. When the neural code is low-dimensional, both tasks can be immediately aligned to a common axis enabling a shared XOR discriminant that generalises across tasks. **c**, In Experiment 1, animals were incentivised to combine two passively viewed task features (colour and shape) in a non-linear fashion (XOR). For example, blue+square and green+diamond combinations were rewarded, whereas blue+diamond and green+square were not. In experiment 2, an additional set of colour–shape contingencies was introduced. Monkeys could generalise the learned rule from experiment 1 to the new set of colours in experiment 2. **d**, Timeline of task events in a single trial. **e**, Neural data was collected from the lateral surface of the prefrontal cortex in two macaque monkeys (see *Methods* and **Supplementary Figure 5** for further details).

Methods

Animals and task

Two adult male rhesus macaques, monkey 1 and monkey 2, were trained in this study. The experiments were conducted in line with the Animals (Scientific Procedures) Act 1986 of the UK and licensed by a Home Office Project License obtained after review by Oxford University's Animal Care and Ethical Review committee. The procedures followed the standards set out in the European Community for the care and use of laboratory animals (EUVD, European Union

directive 86/609/EEC). The animals were seated in a sound- and lighting-attenuated experimental booth. Their heads were restrained and faced a 19-inch screen. The centre of the screen was aligned with a neutral eye position. The animals performed a passive object association task (**Fig. 3.1c,d**). Importantly, neural recordings were collected from the very first session, as one of the main aims of the study was to capture learning dynamics. In an experiment performed previously (experiment 1; **Chapter 2**), the animals learnt that a nonlinear combination of colour and shape information was predictive of reward (blue and green colours; **Fig. 3.1c,d**). In the current experiment (experiment 2) a second set of colours was additionally introduced to test whether the rule learnt in the first experiment cross-generalised to the new sensory domain (**Fig. 3.1c**). The colours used in the coloured circles were designed in the CIELab colour space (Tkalčič & Tasič, n.d.). The L parameter (luminance) was kept constant for every colour which ensured that the stimuli where approximately isoluminant; parameters a and b varied with regard to valence but not value which resulted in a circular colour representation (Tkalčič & Tasič, n.d.). As colours were randomly assigned to conditions for each animal, this circular representation ensured that regardless of which colour pair was assigned to which XOR mapping, the initial colour similarity/dissimilarity within colour pair was kept constant. Additionally, similarly as in experiment 1 (**Chapter 2**), the second object had two features: one relevant for reward prediction (shape) and one irrelevant (width) (for the duration and sequence in which stimuli were presented see **Fig. 3.1d**) The trial sequence was randomised. All trials with fixation errors were excluded. Each session in the dataset contained on average 101.0 ($SD = 18.6$) and 54 ($SD = 1.1$) trials per each of the 16 conditions

for monkey 1 and monkey 2, respectively. Please note that the difference in the standard deviation between animals was due to monkey 1 being often able to participate in the experiment for longer periods of time than monkey 2.

Data acquisition and pre-processing

The surgical procedures adopted in this study mirrored those in experiment 1 as we used the same animals as subjects (detailed in **Chapter 2**, *Data Acquisition and pre-processing*). We transitioned to experiment 2 from experiment 1 once the animals consistently displayed a reward prediction signal. Monkey 1's data spanned 8 sessions, while Monkey 2's spanned 15. Similarly as in experiment 1, we manually adjusted the electrodes a minimum of 62.5 µm before each session to capture activity from new cells. The neural activity recorded was then amplified, filtered (ranging between 300 Hz-10 kHz), and saved for later analysis. Using the valley-seeking algorithm, we separated clusters and smoothed the binary spike train with a Gaussian window ($\sigma = 50ms$). We collected spiking activity from 205 neurons for Monkey 1 and 151 for Monkey 2, limiting our data to cells from the ventral and dorsal lateral frontal cortex (see **Supp. Fig. 6**). Notably, no neurons were excluded based on their selectivity. As the understanding of the influence of learning on neural geometry was central for this study, we standardized (z-scored) the firing rates of each neuron throughout the session to exclude repetition suppression effects. The processed firing rate data were then segmented, capturing activity from $200\,ms$ before to $1200\,ms$ after colour onset. We split all sessions into four distinct learning stages for each monkey and pooled sessions from both animals for analysis. All data analysis was conducted in Python and run

on combined data (monkey 1 and 2). Two types of analyses were used in this study: (1) timepoint-resolved, where a specific method was applied to every time point in the epoch to track how representations evolved in trial time, and (2) time-averaged, where a method was run on time-averaged data (e.g., $[t_{400ms}, t_{500ms}]$, colour-locked or shape-locked) in the time window preceding the shape display or trial outcome. In the former time window, we examined the neural geometry when only the colour information is known (i.e., context representation), whereas just before outcome onset (latter time window), we examined whether neural geometry reflected the colour and the shape and their combination (XOR) before the animals received feedback about the value of the trial.

Multiple regression

Similarly as in experiment 1, neural selectivity was obtained using multiple regression with context, shape, width and XOR variables used as predictors of average firing rates (see **Chapter 2**, *Multiple regression* for more details). Importantly, to examine whether the same PFC cells represented task variables irrespective of whether the old stimulus sets (experiment 1, **Chapter 2**) or the new stimulus set (experiment 2, current study) was shown, we employed a Pearson correlation metric. Specifically, we compared how similar the colour, shape and XOR coefficients in stimulus set 1 are to coefficients for the same variables in stimulus set 2, which yielded three correlation scores for each of the main task variables (**Supp. Fig. 5c-h**). This was done for each of the four learning stages to explore whether selectivity for stimulus set 1 aligns with selectivity for stimulus set 2 as a function of learning, consistent with a shared abstract representation.

Decoding

We used the same SVM-based decoding methods as previously described for experiment 1 (see **Chapter 2**, *Decoding* for details). Briefly, to investigate how neural representation changed with learning, we employed linear SVM decoding, which necessitated the construction of pseudo neural populations by dividing sessions into four learning stages and ensuring uniform trial numbers across these newly constructed pseudo-sessions. Using a matrix of the $trials\ x\ neuron\ x\ time\ points$ dimensionality, we applied binary SVM classifiers for temporally resolved decoding (separately for each timepoint) and time-averaged firing rates (averaged over $[\mathrm{t}_{-100ms}, \mathrm{t}_{0ms}]$ shape-locked and outcome-locked). Decoding was cross-validated, employing alternating training and test sets across 10 iterations, with the resulting scores averaged.

Uniform Manifold Approximation and Projection

UMAP was used to visualise the neural trajectories formed to dynamically solve and cross-generalise an XOR problem. It is a dimensionality reduction technique that utilises the principles of Riemannian geometry and algebraic topology to balance the representation of local vs global patterns embedded in the data (McInnes et al., 2018). The algorithm constructs a graph from the data and then optimises a low-dimensional projection of the original data so that it best approximates the structure of the previously constructed graph (McInnes et al., 2018). UMAP is especially useful to visualise data that include non-linear relationships like the XOR operation. This method was run on condition-averaged

data with width 1 and width 2 trials pooled together. Next, we concatenated trajectories for each of the 8 conditions (4 colours x 2 shapes) along the time dimension to obtain a common solution for each trajectory. Only the first two components were used to visualise to obtained dynamical landscape (**Fig. 3.2c,d**)

Principle Component Analysis

The same PCA procedure was used here as in experiment 1 (see **Chapter 2**, *Methods, Principle Component Analysis* for details). Similarly as for the UMAP analysis, width 1 and width 2 trials were pooled together and then the neural data were condition-averaged.

Cross-stimulus set generalised decoding

Cross-generalised decoding performed for experiment 2 data (both temporally resolved and run on time-averaged firing rates; **Fig. 3.3b,f**, **Fig. 3.4b,f,j** and **Supp. Fig. 3a-d**, **Supp. Fig. 5a,b**, respectively) differed in one aspect from the algorithm described in **Chapter 2**, *Cross-generalised decoding*. As the aim of the analysis described here was to identify the neural format of the main task variables used across stimulus sets, only one splitting variable was used (i.e., stimulus set) to obtain cross-generalisation scores for the task-relevant variables (context, shape, and XOR). This reduced the possible cross-generalisation decoding axes to four possible binary decoding problems (e.g., when performing cross-stimuli set generalised decoding for the context variable we can: (1) train on differentiating colour 1 from colour 2 in stimulus set 1 and test on differentiating colour 3 from colour 4 in stimulus set 2, (2) train on differentiating colour 3 from colour 4 in

stimulus set 2 and test on differentiating colour 1 from colour 2 in stimulus set 1, (3) train on differentiating colour 1 from colour 3 and test on differentiating colour 2 from colour 4, and (4) train on differentiating colour 2 from colour 4 and test on differentiating colour 1 from colour 3; these four decoding scores were then averaged). Using this procedure, we explored the cross-stimulus set generalisation potential of the context, shape, width and XOR variables.

Statistical testing for selectivity measures

As we tested whether selectivity for task variables was similar in stimulus set 1 to variables in stimulus set 2 and whether this selectivity alignment changed over learning, two null distributions were thus constructed: (1) statistically significant selectivity alignment was assessed by comparing the observed correlation to a distribution ($n = 500$) of correlations obtained after shuffling one of the selectivity vectors; (2) learning-induced effects in selectivity alignment were assessed by comparing the observed difference in alignment between stage 1 and stage 4 to a distribution of differences computed after randomly shuffling neurons between stage 1 and 4. These null distribution were then compared to the observed correlation or difference in correlations (learning effect) to obtain p values.

Statistical testing for decoding and cross-set generalised decoding

The same procedure for estimating statistically significant decoding effects was used here as in experiment 1 (see **Chapter 2**, *Statistical testing for decoding and cross. gen. decoding*). Briefly, non-parametric permutation tests were used to determine statistical significance within and between learning stages. We

generated two null distributions for time-resolved decoding and learning-induced effects, permuting trial labels and concatenating matrices across specific stages. To address potential time-related family-wise errors in time-resolved decoding, we incorporated a cluster-based permutation correction (Maris & Oostenveld, 2007).

Results

Similarly as in experiment 1 (**Chapter 2**), the animals were trained to combine the old as well as the new colour stimuli (blue, green, pink or khaki) with a subsequently presented shape (either square or diamond) in a nonlinear fashion to predict the reward (the XOR between colour and shape) outcome of the trial (**Fig. 3.1c-d**). We sequentially recorded 356 neurons from the lateral PFC across both macaques (**Fig. 3.1c**), i.e., new neurons were sampled each session. Importantly, to capture learning dynamics, we started recording from the first session in which the animals were exposed to the new stimuli. The full set of sessions were divided up into four learning stages, e.g., the first 2 sessions in monkey 1 comprised the first learning stage for monkey 1. Data was pooled over animals within each of the learning stages.

A dynamical solution for the XOR operation

Our previous empirical and theoretical work (Stroud, Wójcik, et al., n.d.; Wójcik et al., n.d.) demonstrated that an optimal XOR representation relies on maintaining only the XOR information at the decision time and discarding the input features after the XOR computation has been performed. This research does not directly

conceptualise the mechanism for optimal XOR construction. Importantly, such mechanistic model has been proposed in the past and relies on utilising time as an auxiliary computational dimension (Buonomano & Maass, 2009; Ehrlich & Murray, n.d.; Mante et al., 2013). Specifically, if the information about one input feature is accessible before the remaining input feature, the first feature can be used to contextualise the processing of the second feature through recurrent connections (Buonomano & Maass, 2009). This mechanism was reproduced in RNNs trained on a task that required the integration of different types of information depending on context (Mante et al., 2013). Is it important to note, however, that vanilla RNNs are not able to reproduce the low-dimensional dynamics of solving an XOR observed in the brain. More precisely, the PFC was found to discard irrelevant input features even during presentation (**Chapter 2**; (Wójcik et al., n.d.). This extent of dimensionality reduction is difficult to reproduce in RNNs in a setting where active inputs generate ongoing activity. Recent advances in computational neuroscience demonstrate, however, that an appropriate level of reduction can be achieved by introducing a form of regularisation to the model (Dubreuil et al., 2022; Stroud, Wójcik, et al., n.d.). These considerations led to the conceptualisation of an XOR representation as a dynamical, low-dimensional landscape that supports context-dependent processing (**Fig. 3.2a**).

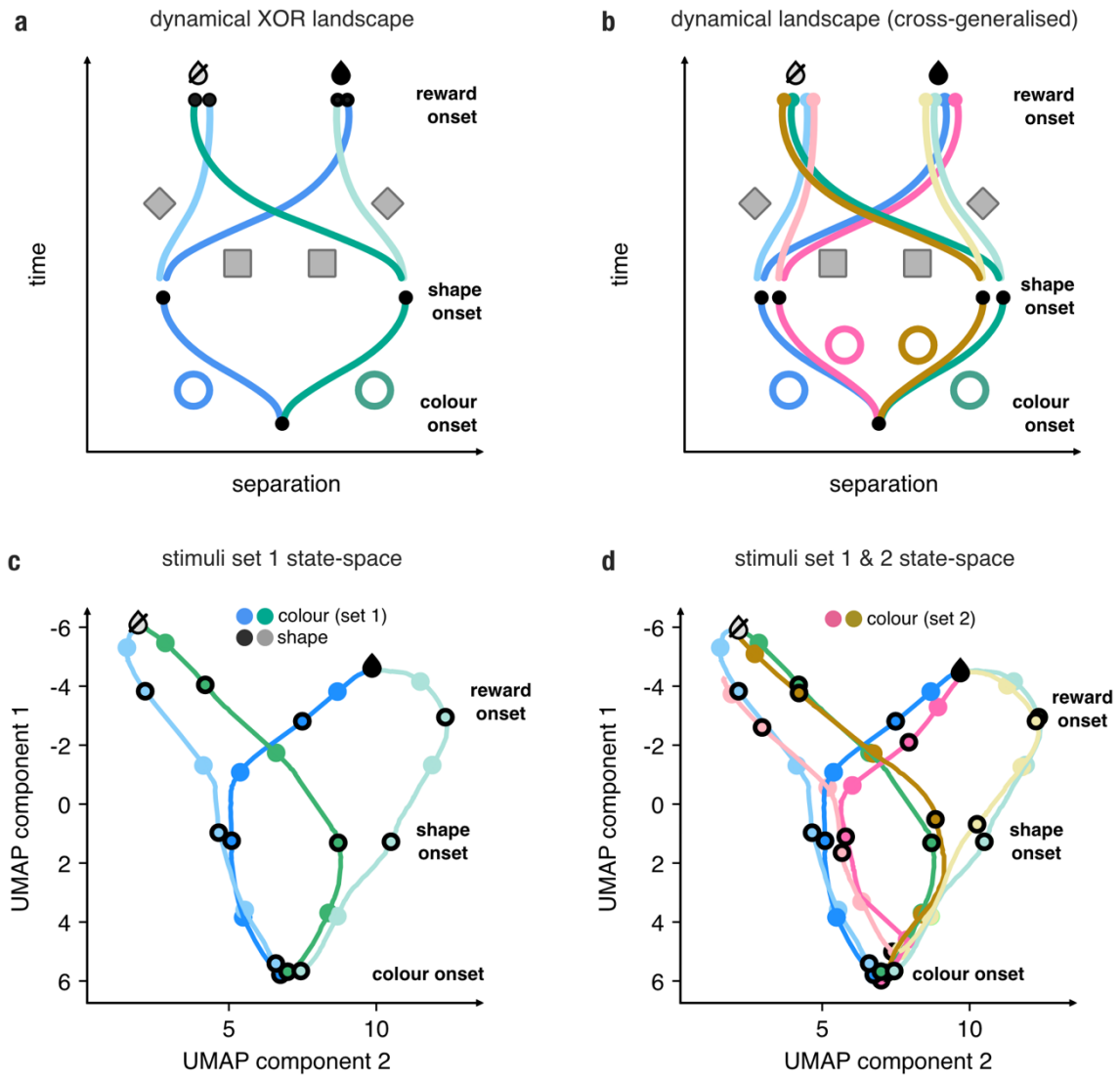


Fig. 3.2. Solving an XOR problem by utilising a dynamical landscape. **a**, schematic illustration of a putative state-space landscape of a system solving an XOR problem using time as an additional computational dimension. The system firstly processes the colour information utilising different areas of the state-space. The processing of the shape feature is then contextualised as this information enters the state-space from different locations (depending on the preceding colour input). Next, activity is guided to the respective *reward* or *no reward* attractors. **b**, the new stimuli set (**Fig. 3.1c**) can be processed using the already learnt dynamical landscape. **c,d**, state-space representations obtained by reducing the neural

dimensionality of condition-averaged trajectories using UMAP (see *Methods*, *Uniform Manifold Approximation and Projection* for details) for stimulus set 1 and stimuli set 1 and 2 combined, respectively; note that the distance between two scatter points along a trajectory reflects $250ms$.

The described mechanism of solving an XOR leads to important predictions in the context of cross-generalisation. If the animals learnt to solve the XOR task using an abstract dynamical landscape in experiment 1 (**Fig 3.2a**), the effect of generalising this representation to a novel setting should be observable at two timepoints. More precisely, if the first feature (colour) was used to contextualise future processing, the neural activity within the PFC should reflect a common coding pattern between the new and old task instance (stimuli set 2), i.e., an abstract context signal (colour onset; **Fig 3.2b**). Consequently, this should lead to the PFC constructing one XOR readout before decision time, irrespective of whether the new or old task instances is being solved (shape onset; **Fig 3.2b**).

To test these predictions, we introduced a new colour pair (stimulus set 2) that followed the same shape–outcome associations as the previous colour pair (stimulus set 1, **Chapter 2**; **Fig. 3.1c**). Randomly interleaving stimulus set 1 trials and stimulus set 2 trials allowed us to test whether a shared neural representation would be used for both stimulus sets (**Fig. 3.1b**). As in experiment 1 (**Chapter 2**), we recorded neural activity from the very first session in which the animals were exposed to the new stimulus set and divided the experimental sessions into four distinct learning stages. We subsequently examined neural state-space

trajectories following the learning process (stage 4) to determine whether the animals employed a dynamic landscape for constructing and cross-generalizing an XOR representation. To reduce the dimensionality the neural data, we employed UMAP (see Methods, *Uniform Manifold Approximation and Projection*). Given its ability to visualize non-linear relationships and preserve the global structure in the resultant projection, UMAP proves ideal for illustrating a dynamic landscape, particularly one akin to the one used for a XOR computation (see **Fig 3.2a**). Our findings highlighted that, by the end of the learning, neural activity closely mirrored the predicted dynamic landscape (**Fig. 3.2c**), with the new stimulus set being processed utilising the same structure (**Fig. 3.2d**).

It is important to note that UMAP primarily served as a visualisation tool in this context, given the inherent challenges in interpreting its component values. Following this, we employed decoding and cross-generalised decoding to quantitatively evaluate how neural representation changed as a function of learning. We next compared these metrics to the changes in the structure of neural selectivites. Specifically, for each neuron, we computed its selectivity profiles for each variable (colour, shape, XOR), separately for each stimulus set, and measured their correlation between different sets. Notably, to test for cross-generalisation at each processing stage, these analyses were conducted for each learning stage in both the colour-locked and shape-locked time windows (**Fig 3.3** and **Fig. 3.4**, respectively).

The PFC constructs an abstract representation of context over learning

Hypothetically, the PFC could represent the new colours separately from the already learned pair (high-dimensional representation), or it could generate an abstract representation (shared across sets) of the colours' contextual meaning (low-dimensional representation). Through an exploration of the neural activity in the colour-locked period (colour-locked; **Fig. 3.1d**), we observed that the PFC collapsed the stimulus set dimension and progressively aligned corresponding colours from both sets according to their meaning (**Fig. 3.3a**). Although context was equally well decodable in early as well as in late learning ($p = .84$, two-sided, **Fig. 3.3b**, black, decoding; see **Supp. Fig. 5a** for temporally resolved decoding), its format changed from an item-specific representation into an abstract code. We tested this explicitly by training a linear SVM classifier to decode the context in stimulus set 1 and tested this model on differentiating context in stimulus set 2, and vice-versa (see *Methods, cross-stimulus set generalisation*). We found that the representations of new and old colours became aligned as a function of learning, to allow for a single context readout ($p = .01$, one-sided, **Fig. 3.3b**, purple, 'cross. decoding'; see **Supp. Fig. 5b** for temporally resolved cross. decoding). This change in geometry was reflected in a change in neural selectivity, as with the progression of learning, PFC cells exhibited an increasingly positive correlation between selectivity for colour pair 1 and colour pair 2 ($p = .000$, one-sided; **Supp. Fig. 5d**; **Fig. 3.3c-d**). Additionally, no learning-induced changes were detected when colour decoding was analysed separately for set 1 and set 2 ($p = .99$, two-sided and $p = .32$, two-sided, respectively; **Fig. 3.3e**). We also investigated the

representation of the stimulus set (the old set 1 vs the new set 2) and found that the animals were able to clearly distinguish between the old and new colours during the early stages of learning (**Fig. 3.3f**, black, 'decoding'). As learning progressed, however, we observed a reduction in stimulus-set decoding ($p = .000$, one-sided) and corresponding cross-context decoding of stimulus set ($p = .02$, one-sided, **Fig. 3.3f**, purple, 'cross. decoding'), as well a reduction in the shared selectivity ($p = .026$, one-sided, **Supp. Fig. 5c**; **Fig. 3.3g, h**). This indicates that the animals collapsed the set 1 vs set 2 dimension. The emergence of an abstract representation of context before the appearance of the shape suggests that the network occupies a preparatory activity state that permits the immediate, context-appropriate transformation of shape information to reward prediction (XOR representation) (Buonomano & Maass, 2009; Dubreuil et al., 2022; Mante et al., 2013; Stokes et al., 2013).

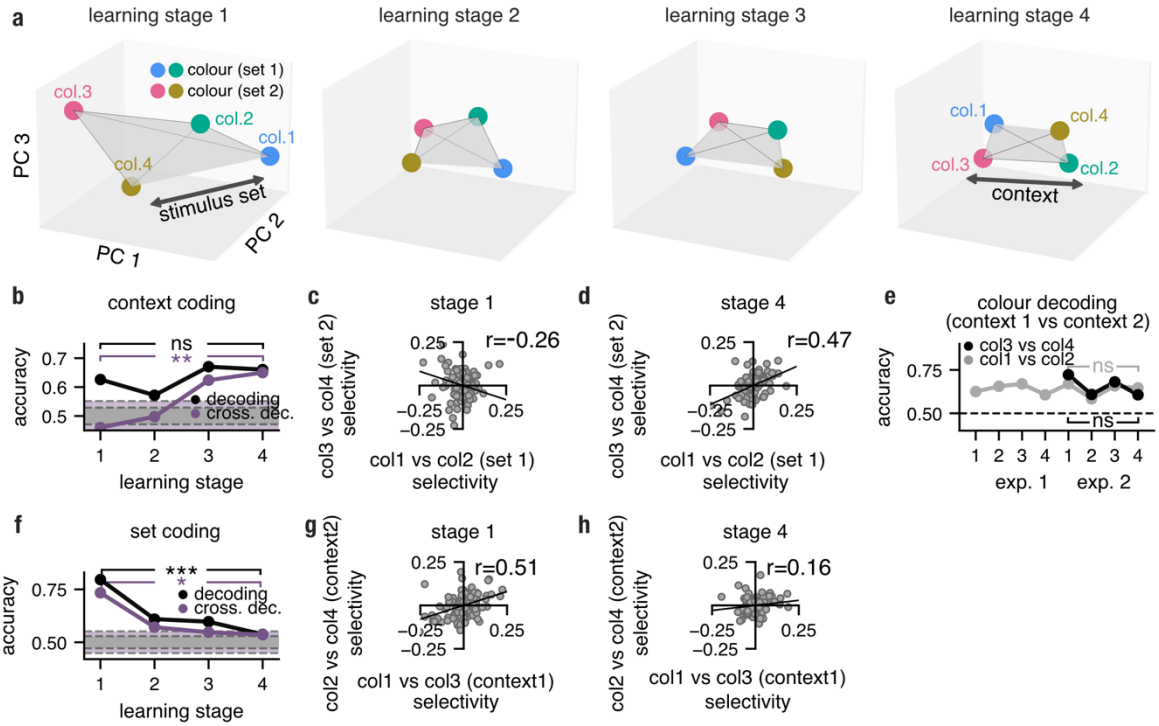


Figure 3.3. **The PFC aligns new and old colours as a function of learning to allow a single readout of context. a**, Trial-averaged population firing rate during the late colour-locked period for each colour projected onto the top three principal components as a function of learning (*Methods, principal component analysis*); the same scale was used for all plots. **b**, Decoding (black) and cross-generalised decoding (purple) of context as a function of learning. The purple and grey shaded areas indicate chance-level cross. gen. decoding and chance-level decoding, respectively. **c, d**, Selectivity of PFC neurons in stage 1 and stage 4 for context in stimulus set 1 and stimulus set 2 in the colour-locked period. The line of best fit is shown in black. **e**, Decoding of colour separately for stimulus set 1 (grey, 'blue vs green') and stimulus set 2 (black, 'pink vs khaki'). **f-h**, Analogous to **b-d** but for set coding (f) and set selectivity (g, h). All p-values were calculated from permutation tests (***, $p < 0.01$; **, $p < 0.01$; *, $p < 0.05$; n.s., not significant). Shading in panels b and f shows the mean ±1 s.d. of chance-level decoding.

The PFC uses one XOR readout to classify both new and old stimuli early in learning

We next focused on the neural activity in the shape-locked period, which reflected the state of the PFC neurons when all necessary information for outcome prediction was available (colour and shape). Inspecting the neural representation of task variables on the population level suggested that set 1 and set 2 geometries aligned as a function of learning to allow a single XOR readout (**Fig. 3.4a**). We found that in both early (stage 1) and late learning (stage 4), XOR decoding was high and no learning-induced changes to the XOR signal were detected ($p = .708$,

two-sided, **Fig. 3.4b**, black, decoding). This suggests that the animals were able to rapidly cross-generalise the XOR representation between stimulus sets as early as in stage 1. We tested this explicitly by training a linear SVM classifier to decode the XOR in stimulus set 1 and tested this classifier on differentiating the XOR in stimulus set 2, and vice-versa. We observed high scores of cross-generalised XOR decoding in both stage 1 and stage 4, as well as a trend-level increase in this measure as a function of learning ($p = .052$, one-sided, **Fig. 3.4b**, purple, 'cross. decoding'). We also observed a learning-induced increase in the shared selectivity for XOR between set 1 and set 2 ($p = .012$, one-sided, **Supp. Fig. 5h; Fig. 3.4c-d**;). These results suggest that, although the PFC constructed an abstract (shared across sets) representation of the XOR early in learning, it continued to align the set-specific XOR boundaries until this abstract format fully dominated the XOR representation. No differences between shape decoding were detected across the learning stages ($p = .104$, two-sided, **Fig. 3.4f**, black, decoding). However, both cross-generalised decoding ($p = .006$, one-sided, **Fig. 3.4f**, purple, 'cross. decoding') and selectivity analyses ($p = .008$, one-sided, **Supp. Fig. 5g**; **Fig. 3.4g,h**;) showed that the PFC progressively aligned the shape representation between stimulus sets. It is important to note that the same shapes were used in both stimulus set 1 and stimulus set 2. These results suggest that, at the start of the learning process, the PFC encodes shape differently (using a high-dimensional representation) depending on the stimulus set, and that only during late learning (stage 4), the representations converge to a single shared axis (i.e., become low-dimensional). We next explored the representation of context (which we already introduced in **Fig. 3.3**) during the shape-locked period. We found no learning-

induced changes to the geometry (decoding: $p = .956$, two-sided; cross. decoding: $p = .398$, one-sided) or selectivity ($p = .382$, one-sided, **Supp. Fig. 5f**; **Fig. 3.4j-l**). We speculate that the PFC already transformed the context signal in the colour-locked period, so that after this preparatory state was used to guide early shape processing, context information did not serve any relevant function in the late shape-locked period. Finally, we found that a similar level of XOR and shape decoding was maintained as a function of learning in both set 1 and set 2 ($p = .156$, one-sided, and $p = .266$, one-sided, respectively; **Fig. 3.4e, i**). The decoding of colour in set 1 approached, however, the levels of colour decoding in set 2 with learning ($p = .014$; one-sided, **Fig. 3.4m**). These changes to neural geometry were accompanied by slight changes in the dimensionality of representations (**Supp. Fig. 5i-j**). Specifically, as compared to the beginning of experiment 1, the mean shattering dimensionality of set 2 decreased slightly (trend-level) over learning in experiment 2 (stage 1: $p = .11$, one-sided; stage 4: $p = .086$, one-sided), whereas the dimensionality of set 1 in experiment 2 remained consistently lower (stage 1: $p = .038$, one-sided; stage 4: $p = .044$, one-sided).

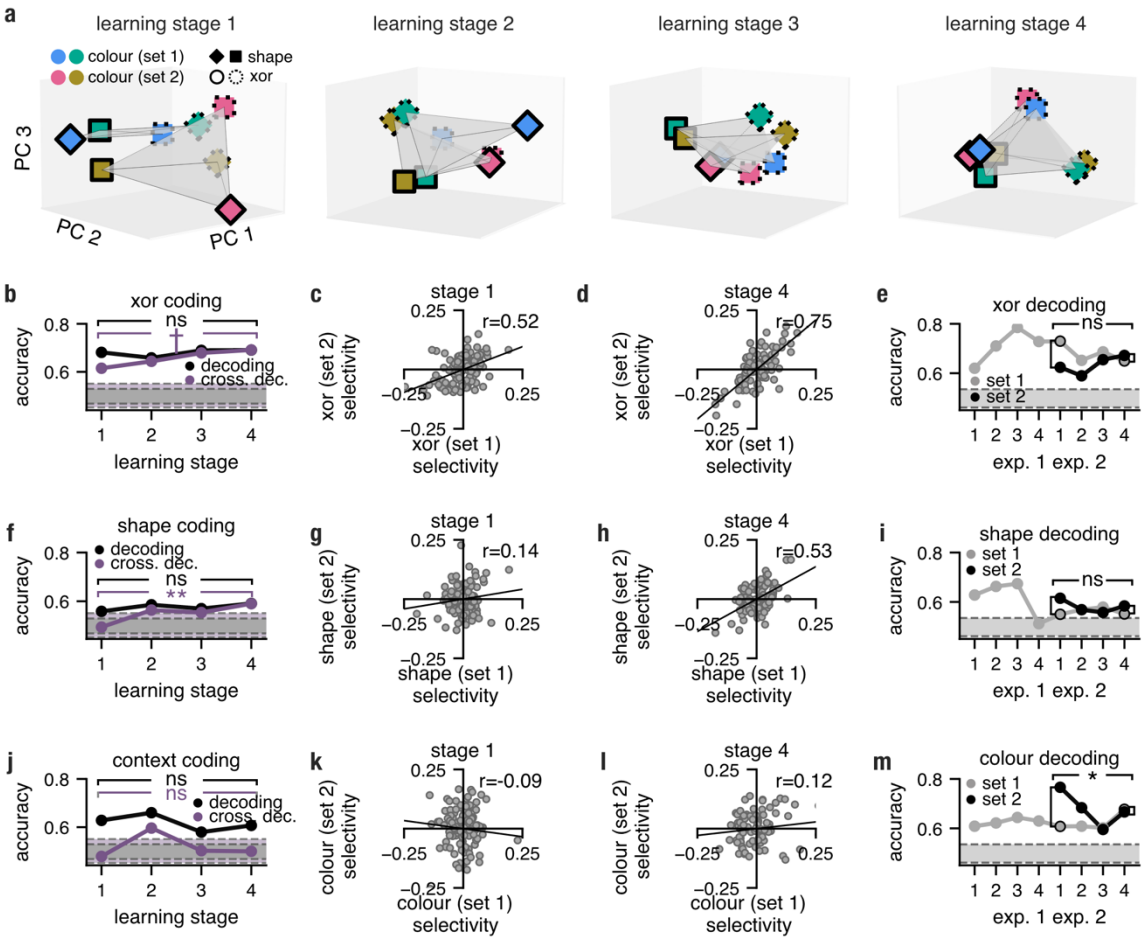


Figure 3.4. The PFC utilised abstract stimulus representations when previously unseen stimulus combinations were introduced. a, Trial-averaged population firing rate during the late shape-locked period for each condition projected onto the top three principal components as a function of learning (*Methods, principal component analysis*); the same scale was used for all plots. **b**, Decoding (black) and cross-generalised decoding (purple) of XOR as a function of learning. The purple and dark grey shaded areas indicate chance-level cross. gen. decoding and chance-level decoding, respectively. **c,d**, Selectivity of PFC neurons for XOR in stimulus set 1 and stimulus set 2 in the shape-locked period in stage 1 and stage 4, respectively. The line of best fit is shown in black. **e**, Decoding of XOR shown separately for stimulus set 1 (grey) and stimulus set 2 (black). The grey

shaded area indicates chance-level decoding. **f-m**, Analogous to b-e but for shape (f-i) and context (j-m). All p-values were calculated from permutation tests (***, $p < 0.01$; **, $p < 0.01$; *, $p < 0.05$; †, $p < 0.01$; $n.s.$, not significant). Shading in panels b, f, and j shows the mean ±1 s.d. of chance-level decoding and cross-set decoding.

Discussion

The complementary learning systems theory posits that neocortical memories should represent abstract elements of the environment that are useful for predicting future events and therefore support adaptability (McClelland et al., 1995; Sun et al., 2023). We tested this hypothesis by exploring whether the PFC is able to utilise stored representations to facilitate learning of previously unseen stimuli combinations. We found that after learning an XOR representation in a previous experiment, animals started to use the same XOR readout early in learning when a new stimulus set was introduced. Crucially, the animals also later generalised the colour-locked preparatory state across task instances, constructing an abstract contextual signal. Together, these findings support the notion that cortical memories, specifically those housed by the PFC, capture the abstract structure of the environment in service of generalisation and adaptation.

Many experimental efforts are guided by the implicit assumption that the animals have no prior knowledge nor have received training related to the task being studied, thus constituting a *tabula rasa* system. Nevertheless, it is unlikely that the

entirety of the animal's experience is unrelated to the experimental task, and that new representations do not interact with information already stored in the neural system. Our design enables a detailed comparison between the learning dynamics of constructing an XOR representation from the ground up (as seen in experiment 1, detailed in **Chapter 2**) and the dynamics of leveraging this pre-existing knowledge to facilitating the learning of a new instance of the same task (as seen in experiment 2). In the latter experiment, we observed a swift generalisation of the XOR readout dimension to the new set of stimuli early in learning. Despite this rapid adaptation, ongoing changes in neural geometry persisted throughout learning. For instance, the context and shape dimensions of both the old and new stimuli sets began aligning along common axes across four learning stages. Contrary to the speed with which XOR was aligned, the pace of alignment for shape and colour was notably slower. Such observations hint at distinct underlying mechanisms for the cross-generalisation of input versus readout features. The XOR representation, for instance, appears to have been cross-generalised swiftly, potentially via utilising the activation dynamics in the PFC (as detailed in **Chapter 1**, *What is feedback control and why do we need it?*). Meanwhile, the more gradual alignment of input features, especially colour, suggests the potential influence of a slow optimisation process. The varied rates of cross-generalisation challenge the idea that the entire task representation was initially generalised as an abstract dynamical landscape. Instead, it seems that such solution was only realised in Experiment 2, where the introduction of two additional colours likely prompted the PFC to formulate a preparatory state for context-dependent processing.

An alternative strategy for generalising previously learnt knowledge is exemplified by neural networks that develop compositional representations (Driscoll et al., 2022; Yang et al., 2019). After being trained on multiple cognitive tasks, these models have the ability to decompose a variety of complex operations into simple computational motifs. Furthermore, compositional models are able to recombine these basic operations to rapidly solve tasks that were not previously seen (Driscoll et al., 2022; Yang et al., 2019). Initial neural dimensionality could thus depend on the similarity of the new problem to the already learnt operations and the number of computational motifs required to effectively solve it. In the current study, the dimensionality of the stimuli set 2 at the beginning of the learning was similar to the dimensionality of stimuli set 1 when the animals were first exposed to the task in experiment 1 (**Chapter 2**). Although the XOR readout was rapidly cross-generalised, this alignment did not translate into an immediate reduction of neural dimensionality. This suggest an alternative interpretation to the one described in the previous paragraph, whereby an abstract dynamical landscape was only learnt in experiment 2. The PFC could have acquired the two-stage solution for the XOR operation in experiment 1 but cross-generalised only the later, shape-locked motif early in experiment 2. The slow, learning-resolved alignment of the stimuli sets in the colour-locked period argues, however, against such a scenario. Specifically, utilising stored dynamical motifs through the control of activation dynamics should results in instantaneous alignment.

It is important to address the constraints inherent in our experimental design. While our results revealed that animals were quick to develop an XOR dimension in a

novel task instance, it remains uncertain whether this neural facilitation resulted in any behavioural advantages. To be more specific, given the reliance on a classical conditioning paradigm, no explicit behavioural readout was required. The only behaviour that was expected from the animals was to consistently maintain fixation during the stimuli display. However, in our past experiments, incorporating a response requirement posed challenges. In a previous experiment involving a different animal cohort, the acquisition of an XOR representation was hindered by heuristic-driven strategies. Specifically, animals resorted to oversimplified representations or stereotypical response patterns to obtain either 75% or 50% of the achievable reward (Unpublished). Once established, these strategies proved resistant to change. A potential remedy might be to decompose the XOR task into elementary computations or to block same-colour trials mirroring a reversal learning paradigm. As touched upon in **Chapter 2**, such design choices could potentially influence, however, the geometry and dimensionality of the acquired representations. Interestingly, empirical studies observing animals in more naturalistic settings have demonstrated an improved ability to learn complex tasks (Leggio et al., 2005; Nithianantharajah & Hannan, 2006). As such, environmental enrichment emerges as a promising avenue for introducing an XOR problem in its full format in animal models.

Future studies could build on the design explored in the described experiment to further examine the role of information encoded in the PFC for adaptation. Our design predominantly facilitated the examination of neural markers for *near transfer*. This pertains to the cross-generalisation of memories to facilitate the

learning of a problem that shared both the abstract structure and some of the sensory characteristics of the previously learnt task. For instance, the ability to cross-generalise the XOR representation might stem from the use of identical shape features in both the old and new stimulus sets. An alternative design, where every input feature is entirely novel, could allow the investigation of neural markers of *far transfer*. This denotes the capacity to cross-generalise the abstract structure of a task, facilitating the learning of a problem that has no sensory parallels to the already learnt task. While this design holds promise in probing the efficacy of purely abstract representations in facilitating learning, it poses some challenges for analysis. Specifically, given the lack of arbitrary relation between the new and old task instances, deducing the trajectory of representation alignment becomes challenging. For illustration, if in the current design new shapes were also introduced to stimuli set 2, predicting whether animals associated the colour blue with pink or khaki would be difficult. Consequently, a *far transfer* design would necessitate an initial use of part of the data to identify the direction of alignment, followed by an evaluation of alignment robustness. Future studies exploring *far transfer* should therefore account for the need of additional data in statistical power estimation.

Finally, an optimal, energy efficient XOR representation at decision time can be constructed through the engagement of processes that are alterative to metabolically constrained optimisation explored in **Chapter 2** and **Chapter 3**. More precisely, agents can use working memory to capitalise on the temporal structure of the task and thus reduce the complexity of computations at decision time

(Dasgupta & Gershman, 2021; Ehrlich & Murray, n.d.). This concept is explored in the next chapter, **Chapter 4**.

Chapter 4: Working memory shapes neural geometry in human EEG

Abstract

Working memory has been traditionally studied as a passive storage for information. Recent advances in cognitive neuroscience suggest, however, that the content entering working memory undergoes transformations that support future behaviour. A machine learning perspective underscores this notion by conceptualising memory processes as a computational resource. Here, we explore this perspective by examining whether maintenance process can shape neural geometry and dimensionality and thus lead to minimal representations. Specifically, we recorded EEG in human participants who learnt to solve a XOR task. We hypothesised that separating task features by a working memory delay would result in participants temporally decomposing the XOR computation by processing one of the task features early in trial time. In line with our predictions, participants transformed the first feature from a sensory to an abstract format and subsequently maintained this pre-processed information throughout the delay. This process was related to the construction of a minimal representation at the moment of decision – an representation that has been found in recent non-human primate research to be energy-efficient and optimal for performance. These results demonstrate that slow optimisation-based learning might not be the only mechanism responsible for generating minimal representations, especially in a setting involving maintenance processes.

Introduction

Working memory, the capacity to briefly retain and manipulate information, is posited as a foundational element of adaptability. Although traditionally perceived as sensory-based (Goldman-Rakic, 1995), this understanding has been increasingly contested by empirical (Spaak et al., 2017; Stokes, 2015; Stokes et al., 2013; Wolff et al., 2017b) and theoretical studies (Stroud, Watanabe, et al., n.d.). Pioneering electrophysiological research in both humans and primates has unveiled that information in working memory undergoes transformations, yielding abstract representations that guide future behaviour, rather than merely mimicking sensory codes (Spaak et al., 2017; Stokes, 2015; Stokes et al., 2013t). Theoretically, the shift from sensory to abstract coding is recognised as an inherent computational characteristic of working memory, optimising performance (Stroud, Watanabe, et al., n.d.).

Insights from artificial intelligence further highlight this notion, demonstrating that memory processes can be viewed as a resource for computations rather than a passive mechanism for storage (Dasgupta & Gershman, 2021). In this light, working memory adapts computations to the current task demands by modulating the space-time trade-off (Dasgupta & Gershman, 2021). Specifically, pre-computed information can be stored in working memory, and thus reduce the computation time at the moment of the decision. Alternatively, processing information in real-time at decision time can conserve memory but demands more computation time and is energy expensive. Progress in computational neuroscience supports this perspective, indicating that utilising working memory

can change neural geometry in a way that supports the temporal decomposition of computations (Ehrlich & Murray, n.d.). Furthermore, decomposing complex operations into several simple problems sequentially in time reduces the computational complexity at the time when an action is required. This suggest that working memory processes can be used to shape the neural geometry of a problem.

Furthermore, working memory can be utilised to filter out information that is irrelevant to task performance and thus to control neural dimensionality. It has been shown, for example, that irrelevant features that enter the PFC are represented only briefly, during the presentation time, but then are rapidly discarded once the stimulation is over (Jacob & Nieder, 2014). Similarly, human EEG studies suggest that only object features that are relevant to task performance are maintained in working memory, whereas irrelevant features are discarded after encoding (Bocincova & Johnson, 2019). This process of selection was parsimoniously described using a dynamical system perspective applied at a population level (Mante et al., 2013). It has been demonstrated that input dimensions that are aligned to the selection vector of a dynamical system are amplified by recurrent dynamics whereas input dimensions that are orthogonal will be suppressed by relaxation dynamics. These considerations indicate that working memory could constitute a process that controls neural dimensionality and consequently leads to different geometries of task representations.

Extensive empirical and theoretical investigations demonstrate that a particular type of representations – low-dimensional codes – are characterised by enhanced efficiency on various metrics. (Barak et al., 2013; Bernardi et al., 2020; Cromer et al., 2010; Cueva et al., 2020; Flesch et al., 2022). For example, these representations can reduce the impact of noise on ongoing decision processes and provide robust and stable readouts (Fusi et al., 2016; Wójcik et al., n.d.). The compressed format allows also for them to be easily generalisable to new situations and thus facilitate adaptability (Reinert et al., 2021; Wójcik et al., n.d.). Previous non-human primate research demonstrated that low-dimensional representations are acquired over learning spanning multiple days and that metabolically constrained optimisation provides a parsimonious description of this process (Chapter 2; Stroud, Wójcik, et al., n.d.; Wójcik et al., n.d.). It is important to note that these observations were made under conditions where all necessary inputs were still available from the environment at the moment of the decision. In an alternative setting, involving delay periods, neural dimensionality could be, however, reduced through the selection of features that enter working memory.

Here, we explored whether engaging memory processes shapes neural dimensionality and geometry in a way that leads to the construction of low-dimensional representations. To test these predictions, we asked human participants to learn an exclusive-or (XOR) mapping – a problem that can be solved by a range of representations, from low- to high-dimensional (Ehrlich & Murray, n.d.). Importantly, by continuously recording the EEG signal, we tracked how the dimensionality and geometry of neural representations changed across learning.

Based on theoretical work that suggest that working memory can be used to temporally decompose computations(Dasgupta & Gershman, 2021; Ehrlich & Murray, n.d.), we hypothesised that separating task features with a memory delay would result in the preparation of an abstract, low-dimensional context signal (representing the first task feature) before all necessary information for solving the task were presented. This decomposition should lead to a minimal, low-dimensional representation at the moment of the decision and be accompanied by reaction time and performance biases (i.e., switch costs). Specifically, we hypothesised that the task feature presented first would be considered a first-order policy and trial-to-trial changes to this policy would impact behavioural performance more than changes to subsequently presented features (Mayr & Kliegl, 2003; Vandierendonck et al., 2010).

In line with these predictions, we found that participants progressively constructed a contextual signal from the sensory information provided by the first feature and subsequently learnt to maintain it throughout the delay. Furthermore, the irrelevant dimensions of the contextual cue were filtered across trial time (but not over learning), resulting in the maintenance of only a low-dimensional representation of the first feature. The preparation of this information was associated with reduced neural dimensionality at decision time. Importantly, these low-dimensional representations at decision time bore a remarkable similarity to the energy-efficient representations observed in non-human primates acquired after extensive training spanning several weeks (Wójcik et al., n.d.). However, in contrast to the predictions of slow, optimisation-based learning, neural dimensionality did not decrease as a

function of learning. Instead, it remained low throughout learning and decreased in trials where a failure of XOR coding was present. These results demonstrate that low-dimensional representations, instrumental in facilitating complex decision-making, can be constructed utilising working memory selection processes.

Methods

Participants

Twenty-five participants (14 women and 11 men) were recruited for the study. The mean age of the sample was 23.3 years. Before taking part in the experiment, all participants declared their written consent according to the guidelines approved by the MS IDREC of the University of Oxford (approval no R62730/RE001 and R51332). No previous history of neuropsychiatric or neurological disorders was reported by the participants who had either normal or corrected-to-normal vision. All participants received an appropriate monetary reward for their time and effort.

Procedure

Participants were seated in a sound-proof and dimly lit room. Their heads were supported by a chinrest which also ensured a constant viewing distance of $70\ cm$. Visual stimuli were then projected on the screen at a spatial resolution of $1024\ x\ 786$ pixels with a $60\ Hz$ refresh rate. The Psychtoolbox run on MATLAB was used to design and run the task.

Task

Participants were encouraged to learn a delayed match-to-sample task (**Fig. 4.1a, b**). First, a coloured circle was presented (colour). After a delay, a shape was displayed, and participants were instructed to choose one of two response buttons (left or right). A nonlinear combination of the colour and shape predicted whether the left or the right response button was correct (the XOR rule). Importantly, the task consisted of two instances of the XOR rule (**Fig. 4.1c**). For example, in colour pair 1 (blue and green) the combination "blue-square" and combination "green-diamond" predicted the left response button to be correct. "Green-square" and "blue-diamond" predicted the right response button to be the correct one. In stimulus colour pair 2, the colours pink and khaki shared the same shape-response mapping as the colours blue and green in colour pair 1, respectively (e.g., "pink-square" and "khaki-diamond" were mapped onto the left response button). Importantly, the participants were not informed about the implicit structure of the task and needed to infer it through trial and error. As the task consisted of eight different stimuli-response combinations which formulated two instances of the same XOR rule, two different solutions could be employed by the participants: (1) they could either memorise all combinations (a high-dimensional strategy) or (2) capitalise on the first feature (colour) being presented first and generate a low-dimensional representation at decision time, i.e., transform the colour information into an abstract context signal (**Fig. 4.1c**) before the shape onset to reduce the computational complexity at the moment of the decision. Each of the conditions ($n = 8$) was repeated 100 times across 20 blocks of 40 trials. The sequence of

trials was randomised. Importantly, 5 task variables could have been represented in the task: colour, context, shape, XOR and motor coding (left vs right response).

Stimuli

The colours of the coloured circles stimulus were chosen using the CIELab colour space (Tkalčič & Tasič, n.d.). To ensure that the colours were approximately isoluminant, the same L parameter was used for every colour. Furthermore, to ensure a circular (equidistant) colour representation, parameters a and b were fixed with respect to value but not valence. Note that colours were randomly mapped to the XOR rule for every participant before the experiment. This was done to ensure that the initial dissimilarity between colours on a physical level did not influence the learning of the XOR rule. The shapes were custom-made in Adobe Illustrator.

EEG acquisition and pre-processing

EEG signal was collected using 61 Ag/AGcl electrodes (EasyCap, Herrshing, Germany) mounted in the cap in line with the 10-20 system (Acharya et al., 2016). The data was amplified and collected at a $1000\,Hz$ sampling rate using the NeuroScan SynAmps RT amplifier and the Scan 4.5 software (Compumedics NeuroScan, Charlotte, NC). The electrode system was grounded with an electrode placed at the surface of the right elbow. To allow efficient ocular artefact detection and elimination, horizontal and vertical electrooculography (EOG) was collected. Specifically, two vertical channels were attached to the suborbit and supraorbit of the left eye and two horizontal channels were attached at the exterior canthi of both

eyes. The impedance of all channels was checked throughout the duration of the experiment and kept below the level of $10\ k\Omega$. After collection, the EEG data were pre-processed using the MNE toolbox that was run on Python 3.5 (Gramfort et al., 2013). Firstly, the data was re-referenced using the right mastoid channel. Then we downsampled the signal to $250Hz$ and used a bandpass filter from 0.01 to $40Hz$. The signal was subsequently epoched into segments starting $-200ms$ and ending $1500ms$ colour-locked and shape-locked. For a decision time analysis data was epoched $-200 - 0ms$ locked to the button press. Horizontal and vertical eye movement, blinks, ECG and movement artefacts were identified using Independent Component Analysis. Then, we estimated artefact rejection thresholds separately for each participant using the Autoreject algorithm and discarded or repaired corrupt epochs (Jas et al., 2017). We also employed the RANSAC algorithm to identify and repair contaminated sensors (Bigdely-Shamlo et al., 2015). Note that automated pre-processing methods have been shown to be less prone to error than methods based on visual inspection and support transparency, reliability, and scalability of EEG analysis (Jas et al., 2017). Our pre-processing pipeline resulted in a mean value of 728 ($SD = 85$) colour-locked, 728 ($SD = 83$) shape-locked, and 740 ($SD = 97$) response-locked epochs of that were then used in the subsequent analysis phase. For analyses presented in **Figure 4.1d, h**, **Figure 4.2b-d, f-h, j, l, n**, **Figure 4.3b-d, f, h, j**, and **Figure** 4.**5** these data was divided equally into four learning stages.

Behavioural analysis

Performance was calculated as the proportion of correct responses in each of the four learning stages. Similarly, reaction times were operationalised as median (over trials) time measured from the onset of the shape stimulus and response button press in each of the four learning stages. Next, we computed three types of switch costs for both performance and reaction times using all trials (**Fig. 4.1i, f, g** and **j, k, l**, respectively). We first contrasted performance scores and reaction times obtained from context switch trials where the context (where context denotes the predictive value of the colour, shared between two colours, e.g., the square should be mapped to the left button in the context of both the blue and the pink colour; **Fig. 4.1c**) on the current trial changed as compared to the previous trial against trials in which the context remained the same (**Fig. 4.1i, j)**. To test whether context switches impacted performance and reaction times more than shape switches we compared context switch trials to shape switch trials (**Fig. 4.1f, k)**. Finally, context switch trials were compared to colour switch trials, where the colour on the current trial changed as compared to the previous trials but the context remained the same (**Fig. 4.1g, l**). This was done to measure a context-independent impact of a sensory switch on performance and reaction times. Importantly, as frequentist statistics were used in the behavioural analysis, reaction times distribution were transformed using a log transform to approximate a normal distribution.

Decoding

To determine the amount of information represented for each task variable by the brain linear discriminant analysis (LDA) decoders were employed. The inherent

presence of noise and spatial correlations present in the EEG signal can make decoding particularly difficult. Specifically, because electrophysiological potential can propagate freely through brain tissue, the same source signal can be measured by multiple EEG sensors leading to a spatial covariance structure that is prevalent in the observed signal. Additionally, the process of referencing introduces artefacts present at the reference site to every sensor in the dataset. To account for these characteristics a LDA decoder with shrinkage was used. The shrinkage factor was chosen based on the Ledoit-Wolf lemma (Ledoit & Wolf, 2004). The decoding analysis was run within each participant for each time point separately. First, trials were split randomly into equal test and train sets. Next, an LDA classifier was trained on the train data and subsequently tested on the test data (and vice-versa). The scores obtained from these two procedures were then averaged. This was repeated for 10 random $50\%/50\%$ splits and the scores were again averaged. Importantly, to obtain an estimate of colour decoding that was not biased by context information (i.e., an estimate of pure irrelevant information) we run two separate decoders that differentiated colours within the same context: (1) colour 1 vs colour 3 (context 1), and (2) colour 2 from colour 4 (context 1). The scores were then averaged. Decoding analyses were also run on time average data. For this purpose, learning resolved data (4 learning stages) were time-averaged (trial time) in the colour-locked period ($[\mathrm{t}_{100ms}, \mathrm{t}_{500ms}]$, colour-locked), delay-locked period ($[\mathrm{t}_{500ms}, \mathrm{t}_{1500ms}]$, delay-locked), and shape-locked period ($[\mathrm{t}_{100ms}, \mathrm{t}_{500ms}]$, shape-locked). When analysing neural activity at the moment of decision the EEG signal was averaged in the response-locked period ($[\mathrm{t}_{-200ms}, \mathrm{t}_{0ms}]$).

Neural dimensionality

Neural dimensionality was operationalised in this study as shattering dimensionality, a metric that has been applied in several previous experiments (Bernardi et al., 2020; Rigotti et al., 2013; Wójcik et al., n.d.). That is, we employed the same decoding procedure described in the *Decoding* section to estimate the decoding of 35 theoretically possible variables of a task that has three linear input variables (colour, context and shape). These scores were then averaged to obtain one measure of neural dimensionality.

Cross-colour generalised decoding

To explore the neural geometry of the task information represented by the participants, a cross-generalised decoding approach using LDA was utilised. Specifically, we tested whether learning the task resulted in representations that were shared across different levels of the irrelevant feature (colour). This resulted in four possible binary decoding problems (e.g., when performing cross-generalised decoding for the context variable we can: (1) train on differentiating colour 1 (context 1) from colour 2 (context 2) in and test on differentiating colour 3 (context 1) from colour 4 (context 2), (2) train on differentiating colour 3 (context 1) from colour 4 (context 2) and test on differentiating colour 1 (context 1) from colour 2 (context 2), (3) train on differentiating colour 1 (context 1) from colour 4 (context 2) and test on differentiating colour 2 (context 2) from colour 3 (context 1), and (4) train on differentiating colour 2 (context 2) from colour 3 (context 1) and test on differentiating colour 1 (context 1) from colour 4 (context 2); these four

decoding scores were then averaged). Using this procedure, we explored the cross-colour generalisation potential of the context, shape, XOR and motor coding variables.

Full cross-generalised decoding

To explore the neural geometry of the neural representations at the moment of the decision we used cross-generalised LDA decoding (Bernardi et al., 2020; Wójcik et al., n.d.). Similarly as in **Chapter 2**, trials were split according to the to-be-decoded target variable labels as well as the splitting variable labels. That is, the splitting variable provided the training and test instances of the data for the decoding of the target variable. The task examined in this chapter had three linear input variables (colour, context and shape) which resulted in 36 possible cross. gen axes per each of the task variables (for more details see *Chapter 1, Cross-generalised decoding*; Wójcik et al., n.d.). The decoding scores of the 36 cross. gen. axes were then averaged to obtain a single score for each of the task variables still present at the moment of the decision (i.e., context, shape and XOR).

Significance testing

Time-resolved one-sample t-tests (one-tailed) were used to test for statistical significance of the observed decoding accuracy scores. To correct for family-wise errors a cluster-based correction was employed (Maris & Oostenveld, 2007). To test for learning-induced changes in accuracy, reaction times, decoding and selectivity measures we compared data from learning stage 1 to data from learning stage 4 using a t-test for dependent samples. Furthermore, shaded areas in all

figures represent 95% confidence intervals obtained through a resampling with replacement procedure ($n = 1000$). In analyses where decoding scores were compared to the chance level, one-sample t tests were used. Test statistics were supplemented with Cohen's d effect size estimation. Two participants were excluded from the response-locked analyses due to trigger failure and insufficient trials for shattering dimensionality estimation.

Results

Human behaviour exhibiting evidence of employing a hierarchical strategy

Participants learned to solve the task rapidly with a performance above chance level in the first learning stage, $M = .75,\ t(24) = 9.16, p = .0$ (**Fig. 4.1d)**. Task accuracy also significantly increased with learning from stage 1 ($M = .75$) to stage 4 ($M = .96$), where participants exhibited a ceiling-level performance, $t(24) = -7.22,\ p = .000,\ d = 1.51$. We next explored whether participants utilised a hierarchical representation to solve the task, whereby the colour information was used as a first-order policy and the shape information as a second-order policy. We found that performance in context switch trials ($M = 89.2\%$; where context in the current trial changed in comparison to context in the previous trials) was lower compared to context stay-trials ($M = 91.3\%$; where the same context was presented in the current and previous trial), $t(24) = -3.85, p = 0.0, d = 0.3$ (**Fig. 4.1e**). Importantly, a switch in context had a greater negative impact on

performance than a switch in shape, $M = 89.2\%\ vs\ M = 90.1\%, t(24) = -2.96, p = 0.0, d = 0.13$ (**Fig. 4.1f**). We found no differences in performance when context-switch trials and colour-switch trials were compared (when the previous colour was different from the current colour, but the context stayed the same) $M = 89.2\%\ vs\ M = 89.8\%, t(24) = -1.18, p = 0.13, d = 0.08$ (**Fig. 4.1g**). The learning-induced increase in performance was not accompanied by a decrease in reaction times, $M = 554\ ms\ vs\ M = 485\ ms,\ t(24) = 1.15, p = 0.13, d = 0.19$ (**Fig. 4.1h**). We also found that the switch-costs in reaction times mirrored the pattern of results found in performance switch costs. Context switch trials were characterised by slower median reaction times compared to context stay trials, $M = 478\ ms\ vs\ M = 464\ ms,\ t(24) = 4.34, p = 0.0, d = 0.22$ (**Fig. 4.1j**) and shape switch trials, $M = 479\ ms\ vs\ M = 475\ ms,\ t(24) = 4.42, p = 0.0, d = 0.15$ (**Fig. 4.1k**). Interestingly, we also found that participants exhibited slower reaction times in context switch trials as compared to colour switch trials, $M = 480\ ms\ vs\ M = 481\ ms,\ t(24) = 2.07, p = 0.02, d = 0.07$ (**Fig. 4.1l**). These results indicate that participants used a highly structured representation of the task, in which the colour constituted a first-order policy compared to the shape feature. This led to worse and slower adaptation on trials where a switch in this higher-order policy took place. Next, we explored whether these behavioural effects translated into hypothesised changes in neural codes.

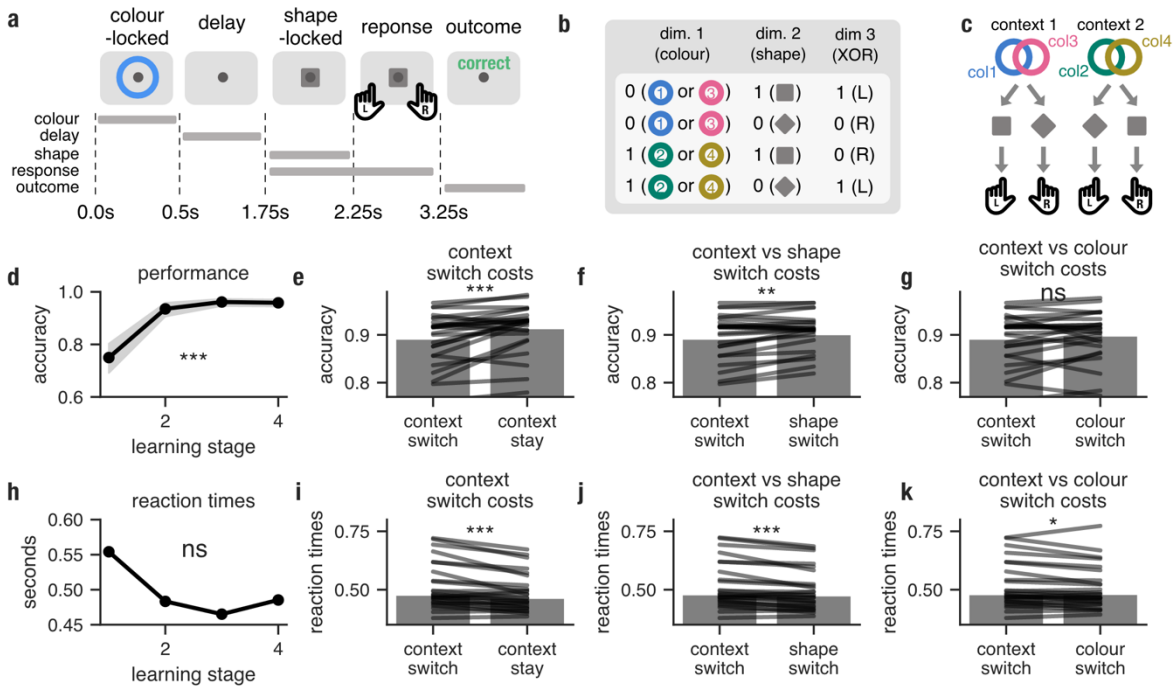


Figure 4.1. **Experimental design and behavioural results**. **a**, Timeline of task events in a single trial. **b**, Participants learnt by trial-and-error to combine two task features (colour and shape) in a non-linear fashion (XOR) and respond either by pressing the left or the right response button. For example, blue+square and green+diamond combined with left button press (XOR == True (1)) were followed by positive feedback; blue+diamond and green+square combined with right button press (XOR == True (1)) were also followed by positive feedback. Any other combination was followed by negative feedback information. **b**, The task consisted of three stimuli features: colour ($n = 4$), shape ($n = 2$) and XOR conditions ($n = 2$; aligned fully with left and right button responses). **c**, Participants could generalise the meaning of colour 1 (e.g. blue) to colour 3 (e.g. pink), and vice versa, as both shared the same shape-XOR mapping. This could result in colours being grouped (clustered) by the context information they provided for future (in trial time) shape-XOR (response) classification. **d**, Mean performance accuracy plotted as a function of learning stages; shaded area indicated 95% CIs obtained through random resampling with replacement ($n = 1000$); accuracy in stage 1 was compared to accuracy in stage 4. **e**, Comparison of accuracy between context switch trials (the context of the current trial

changed compared to the previous trial) and context stay trials (the context of the current trial remained the same compared to the previous trial); horizontal lines illustrate differences in accuracy for every participant individually. **f**, Analogues to e but comparing context switch trials to shape switch trials (the shape of the current trial changed compared to the previous trial). **g**, Analogues to e but comparing context switch trials to colour switch trials (the colour of the current trial changed compared to the previous trial; as the context remained the same this measure captures changes caused by sensory differences only). **h-l**, Analogues to panels d-g but plotted for reaction time medians (before transformations). All p-values were calculated using t-tests for dependent samples (***, $p < 0.01$; **, $p < 0.01$; *, $p < 0.05$; †, $p < 0.1$; n.s., not significant); reaction times were transformed using the log transformation prior running parametric tests.

Humans learn to maintain the context information and construct the XOR feature over learning

To solve the task participants were required to construct an XOR dimension by mixing nonlinearly the context and shape features. As the colour stimulus is displayed only for 500ms at the beginning of the trial, participants needed to maintain the colour information in working memory till the moment when the shape was displayed. The cue stimuli consisted of two dimensions, the irrelevant colour dimension and the relevant context dimension. The former represents sensory differences between colours (blue vs pink or green vs khaki; **Fig. 4.1c**) and the latter dictates the context within which the next feature (shape) should be processed (blue vs green or pink vs khaki; **Fig. 4.1c**). We hypothesised that participants would exhibit an increase in context coding but not colour coding in the

delay period and that this would lead to an increase in XOR coding in the shape-locked period – both processes were predicted to be learning-resolved.

We first employed LDA decoding on each time-point separately, on trials pooled across learning stages. This time-resolved analysis showed that the context information peaked early, in the colour-locked period, was maintained throughout the delay to then increase slightly after shape presentation, $cluster\ 1\colon 0.082 - 1.188s, p = 0.001$ and $cluster\ 2\colon 1.208 - 3.2s, p = 0.001$ (**Fig. 4.2a**). Next, the EEG signal was averaged in the colour-locked, delay-locked and shape-locked periods to explore learning-induced changes in decoding scores (see *Methods, decoding* for more details). More specifically, the mean activity in the time window was taken for each electrode, to create a single feature map (over electrodes) for each time window, which was then subjected to classifier analysis. We found that context coding in stage 4 increased as compared to stage 1 in: (1) the colour-locked period: $M = 0.515\ vs\ M = 0.50,\ t(24) = -1.43, p = 0.08, d = 0.32$; (2) the delay-locked period: $M = 0.519\ vs\ M = 0.507,\ t(24) = -1.8, p = 0.04, d = 0.35$; and (3) the shape-locked period: $M = 0.515\ vs\ M = 0.504,\ t(24) = -1.34, p = 0.1, d = 0.25$ (**Fig. 4.2b-d).**

We hypothesised that working memory processes can control the dimensionality of neural representations by selecting features for maintenance. We tested this prediction by exploring the learning dynamics of the colour representation. Importantly, to obtain an isolated measure of colour coding (unrelated to context) we focused on decoding colour that provided the same context information (blue

vs pink or green vs khaki). Decoding analyses demonstrated that colour information peaked in the early colour-locked period of the trial and then rapidly declined over time to reach chance levels before the delay-locked period, $cluster\ 1\colon 0.082 - 0.484\ ms, p = 0.006$ (**Fig. 4.2e**). No learning-induced changes to colour decoding were detected between stage 1 and stage 4 in the colour-locked, $M = 0.494\ vs\ M = 0.499;\ t(24) = -0.92, p = 0.18, d = 0.15,$ and the delay-locked, $M = 0.498\ vs\ M = 0.498,\ t(24) = 0.03, p = 0.51, d = 0.0$ periods (**Fig. 4.2f-g**). A learning induced increase of colour decoding between stage 1 and stage 4 in the shape-locked period was observed, $M = 0.491\ vs\ M = 0.506,\ t(24) = 2.01, p = 0.03, d = 0.38$ (**Fig. 4.2h**). It is likely that this effect represents a global change in the geometry of the task representation rather than increase in colour coding per se, as colour decoding in stage 4 was not different from chance levels, $M = 0.506,$ $t(24) = 1.0, p = 0.33.$ These results suggest that the neural system likely discarded the colour information as a function of trial time capitalising on firing rate decay in the absence of input stimulation. Only information relevant for performance (context) was entered working memory and was maintained.

Time-resolved decoding also revealed that the shape information rapidly peaked, persisted through the presentation period and then got gradually discarded as a function of trial time after stimulation was over, $cluster\ 1\colon 1.772 - 2.556s, p = 0.001$ (**Fig. 4.2i**). Furthermore, no learning-induced (stage 1 vs stage 4) changes to shape coding were detected, $M = 0.516\ vs\ M = 0.508,\ t(24) = 0.79, p = 0.782, d = 0.15$ (**Fig. 4.2j**). XOR and motor decoding followed the same temporal evolution with a relatively (compared to shape) late peak and a subsequently slowly

decaying dynamic; XOR: $cluster\ 1: 1.892 - 3.2\ ms, p = 0.001$; motor: $cluster\ 1: 1.45 - 1.772\ ms, p = 0.039, cluster\ 2: 1.792 - 3.2\ ms, p = 0.001$ (**Fig. 4.2k** and **Fig. 4.2m**). In line with our predictions, the XOR signal exhibited a learning-induced increase from stage 1 to stage 4, $M = 0.496\ vs\ M = 0.516$, $t(24) = -1.85, p = 0.04, d = 0.4$ (**Fig. 4.2l**).

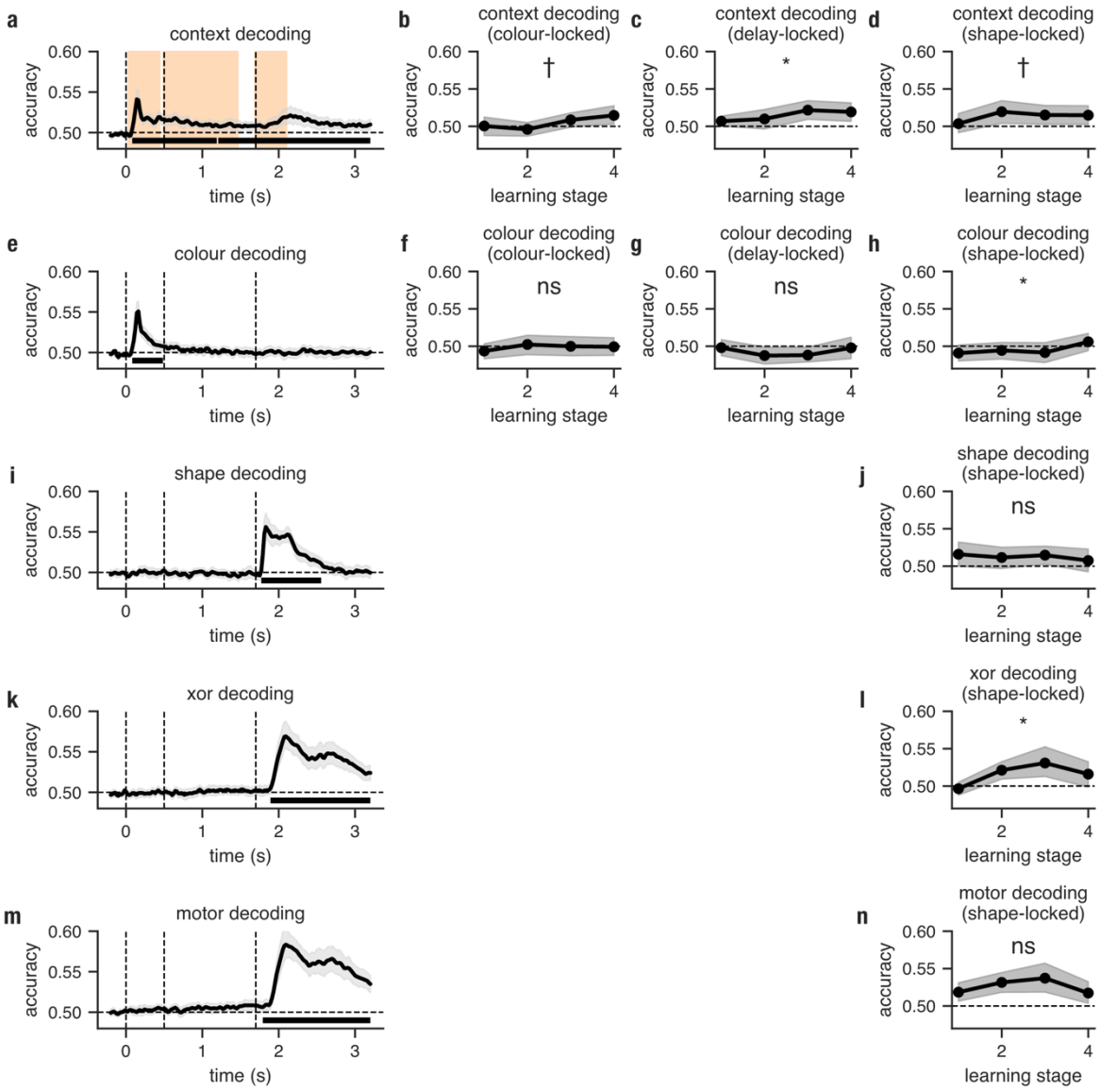


Figure 4.2. Decoding of task variables plotted as a function of time and learning. a, Time-resolved decoding of context; horizontal bars indicate statistical significance; the pale orange areas indicate the time windows for which subsequent time-averaged

decoding analyses were run (panels b-d). Vertical three dashed lines show the onset of the colour, delay, and shape, respectively; **b-d**, The learning dynamics of the context decoding computed from the time-averaged signal in the colour-, delay-, and shape-locked periods, respectively. **e-h**, Analogues to a-d but for within-context colour decoding (captures only the physical properties of the colour). **i,j**, **k,l**, and **m,l** Analogues to a-d but for shape, XOR and motor decoding, respectively. The shaded area indicated 95% CIs obtained through random resampling with replacement ($n = 1000$); All p-values for stage 1 vs stage 4 comparisons were calculated using t-tests for dependent samples (***, $p < 0.01$; **, $p < 0.01$; *, $p < 0.05$; †, $p < 0.1$; n.s., not significant)

The neural format of the context signal evolves over learning

A defining characteristic of low-dimensional task representations is that they can be easily cross-generalised to sensory different instances of the same task. Guided by this consideration, we next tested whether task features like context, shape, XOR and the motor signal were coded in a similar format across different levels of the irrelevant feature (colour). For example, to explore the cross-colour generalisation of context we trained a classifier on differentiating blue from green and tested it on the discrimination of pink vs khaki, and vice versa (see *Methods, cross-colour generalised decoding* for details). Decoding scores above chance obtained from such an analysis would suggest that the same abstract format of context was used in both sets of colours. It is important to note, however, that the lack of colour coding in the delay- and shape-locked periods indicates that the alignment of discriminant boundaries for each of the task variables (context, shape, XOR and motor coding) across different levels of colour would be trivial in these trial periods.

We found that cross-colour generalised context decoding mirrors the temporal dynamics of simple context decoding only in the shape-locked period, $cluster\ 1\colon 1.892 - 2.315\ ms, p = 0.001,$ $cluster\ 2\colon 2.335 - 2.516\ ms, p = 0.011$ (**Fig. 4.3a**). Therefore, context transforms into an abstract format only later in the period. Furthermore, the learning-resolved data reveal that contextual information transforms into an abstract cross-generalisable format over learning in both the colour-locked (stage 1 vs stage 4: $M = 0.492\ vs\ M = 0.5, t(24) = 1.6, p = 0.06, d = 0.31$) and delay-locked (stage 1 vs stage 4: $M = 0.502\ vs\ M = 0.523,\ t(24) = 1.71, p = 0.05, d = 0.3$) periods (**Fig. 4.3a-d**). Time-resolved cross-colour generalised decoding of shape ($cluster\ 1\colon 1.772 - 2.375\ ms, p = 0.002$), XOR ($cluster\ 1\colon 1.892 - 3.2\ ms, p = 0.001$) and motor coding ($cluster\ 1\colon 1.289 - 3.2\ ms, p = 0.001$; **Fig. 4.3e,g,i**) and their learning dynamics (**Fig. 4.3f,h,j**) follow the same pattern of results as observed for simple decoding (**Fig. 4.2i-n**) with only the XOR exhibiting a learning induced effect, $M = 0.484\ vs\ M = 0.513$ (stage 1 vs stage 4), $t(24) = 2.19, p = 0.02, d = 0.45$. As discussed above, only cross-colour generalisation of context in the sample-locked period is non-trivial, however, as strong colour coding was only observable in the colour-locked period. Taken together, these findings imply that participants constructed abstract representations of task features but that the mechanism responsible for this transformation relied heavily on discarding colour information early in trial time. That is, the dimensionality of the task representation was reduced as a function of time and not learning, through the selection of which information enters working memory. It is worth noting the U-shaped learning dynamic observed across each

task variable when learning dynamics were examined. This might be because task-specific geometrical transformations to the neural activity might reflect preparation processes and shift with reaction times. Averaging a temporally shifting neural signal using the same time window might thus result in a U-shaped dynamic. We next analysed response-locked activity to examine neural representations that were not biased by preparation-relates shifts.

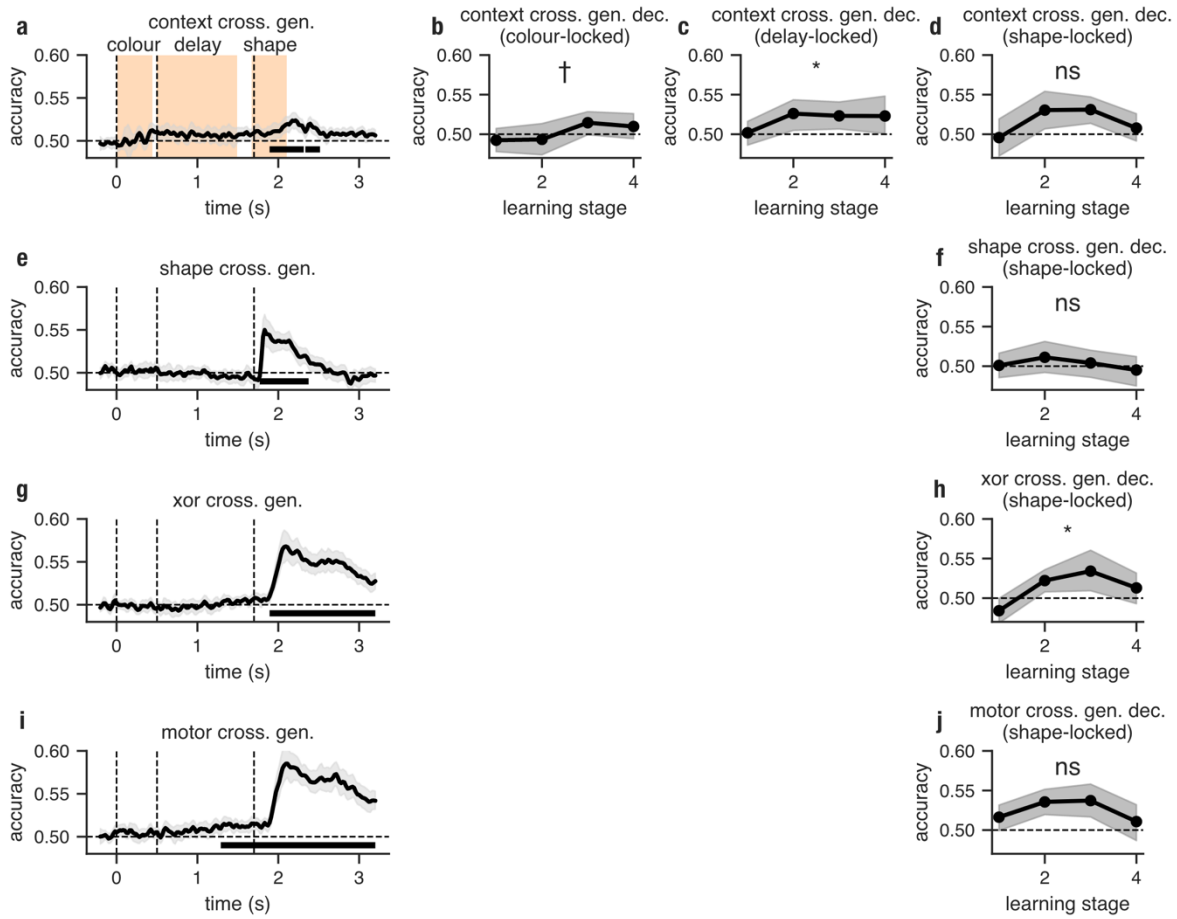


Figure 4.3. Decoding of task variables generalised across colour and plotted as a function of time and learning. a, Time-resolved cross-colour decoding of context; horizontal bars indicate statistical significance; the pale orange areas indicate the time windows for which subsequent decoding analyses were run (panels b-d). Vertical three

dashed lines show the onset of the colour, delay, and shape, respectively; **b-d**, The learning dynamics of the cross-colour context decoding computed from the time-averaged signal in the colour-, delay-, and shape-locked periods, respectively. **e,f** and **g,h**, and **I,j** Analogues to a-d but for shape, XOR and motor cross-colour decoding, respectively. The shaded area indicated 95% CIs obtained through random resampling with replacement ($n = 1000$); All p-values for stage 1 vs stage 4 comparisons were calculated using t-tests for dependent samples (***, $p < 0.01$; **, $p < 0.01$; *, $p < 0.05$; †, $p < 0.1$; n.s., not significant).

Analysis of neural geometry reveals evidence of using minimal representations at the time of the decision

We hypothesised here that the engagement of working memory processes could result in the preparation of a part of the necessary computations and thus lead to utilising a low-dimensional task representation at the decision timepoint. To test these predictions, we analysed the geometry and dimensionality of averaged neural responses, shortly before the motor response was triggered ($[\mathrm{t}_{-200ms}, \mathrm{t}_{0ms}]$).

By employing the same decoding and cross-generalised decoding analyses as in our non-human primates study (**Chapter 2**; Wójcik et al., n.d.), we found that neural representations found at the level of the entire human cortex after learning closely resemble the representations found in trained non-human primates and task-optimised neural models. More specifically, we observed that decoding of the XOR feature increases with learning (stage 1 vs stage 4), $M = 0.516\ vs\ M = 0.572$, $t(23) = 4.05, p < .001, d = 0.77$ (**Fig. 4.4a**). Moreover, the XOR feature became

progressively represented in an abstract format over learning as evidenced by the increase in cross-generalised decoding from stage 1 to stage 4, $M = 0.512\ vs\ M = 0.56,\ t(23) = 3.52, p = 0.001, d = 0.66$ (**Fig. 4.4b**). Additionally, in line with an metabolically constrained, minimal model of solving an XOR, the cross-generalised decoding of context and shape decreased over learning (stage 1 vs stage 4); context: $M = 0.514\ vs\ M = 0.494,\ t(23) = 1.54, p = 0.068, d = 0.32$; shape: $M = 0.513\ vs\ M = 0.493,\ t(23) = 1.86, p = 0.038, d = 0.41$ (**Fig. 4.4b**). Note that the XOR decoding is colinear with the motor decoding at later stages of learning (i.e., XOR True == Left and XOR False == Right; **Fig. 4.1c**). However, in early learning (stage 1), when participants have not yet constructed a prominent XOR representation, motor coding is not related to XOR coding. This is corroborated by the above chance-level decoding of motor response in both stage 1 ($M = 0.592, t(23) = 6.65, p < .000$) and stage 2 ($M = 0.582,\ t(23) = 6.69, p < .000$) and no significant change in motor coding over learning (stage 1 vs stage 4), $M = 0.592\ vs\ M = 0.582, t(23) = 1.3, p = 0.205, d = 0.12$ (**Fig. 4.4c**). We next explored whether the process of learning influenced the neural dimensionality of the representations used to solve the task. An analysis run on all trials (correct and incorrect combined) indicated that, although shattering dimensionality (see *Method, neural dimensionality* for details) was low throughout the entire experiment, an increase over learning (stage 1 vs stage 4) was observed, $M = 0.507\ vs\ M = 0.515,\ t(22) = 2.73, p = 0.012, d = 0.5$ (gold line, **Fig. 4.4d**). These results deviate from predictions of the minimal model (**Chapter 2, Fig. 2.2**) and from the observed learning dynamics of neural dimensionality in non-human primates (**Chapter 2, Fig. 2.3**) which both demonstrate a dimensionality reduction

over learning. We hypothesised that low neural dimensionality in learning stage 1 might be related to a high proportion of error trials. That is, a failure in maintaining context information or constructing the XOR in error trials might have resulted in a low-dimensional representation at the decision timepoint and thus decreased neural dimensionality when all trials we analysed. As the number of error trials was not large enough for a shattering dimensionality analysis, we compared the learning dynamic of shattering dimensionality computed on all trials to the dimensionality dynamic obtained from only correct trials (gold vs green; **Fig. 4.4d**). This data split revealed that dimensionality in correct trials did not change over learning from stage 1 to stage 4, $M = 0.518\ vs\ M = 0.517,\ t(22) = 0.51, p = 0.615, d = 0.1$. Moreover, a statistically significant difference between shattering dimensionality computed on all trials and computed on correct trials was only detected in learning stage 1, $M = 0.507\ vs\ M = 0.517, t(22) = 3.71, p = 0.001, d = 0.54$, i.e., the learning stage containing the most error trials. These findings suggest that the observed increase in dimensionality reflected the differences between the neural geometry found in correct and incorrect trials rather than being indicative of dimensionality expansion as a learning strategy. To address this explicitly, we next explored whether correct trials in stage 1 are in fact characterised by a more prominent representation of context or the XOR. In line with these predictions, we found that correct trials compared to all trials in stage 1 were characterised by an increase in both XOR coding, $M = 0.579\ vs\ M = 0.513, t(22) = -4.66, p = 0.0, d = 0.81$, and XOR cross-generalised decoding, $M = 0.569\ vs\ M = 0.51,\ t(22) = -4.01, p < .000, d = 0.71$ (**Fig. 4.4e, f**). Cross-generalised decoding of shape was also lower in correct trials as compared to all trials, $M = 0.489\ vs\ M = 0.507,$

$t(22) = 3.0, p = 0.003, d = 0.34$ (**Fig. 4.4f**). Correct trials in stage 1 exhibited also slightly higher context coding compared to all trials, $M = 0.533\ vs\ M = 0.521$, $t(22) = -1.82, p = 0.083, d = 0.18$ (**Fig. 4.4e**). These results indicate that participants, similarly to non-human primates in experiment 1 (**Chapter 2**) acquired a low-dimensional task representation at the end of learning. The learning process responsible for these changes seems, however, be different than metabolically-constrained optimisation as the dimensionality of correct trials did not change over the progression of the experiment.

To further support this claim we explored how the main task variables (colour, shape and XOR coding) change as a function of experiment progression. We compared neural representations used in correct trials in stage 1 to representations obtained from correct trials in stage 4. We observed no differences in decoding ($M = 0.51\ vs\ M = 0.508,\ t(22) = 0.15, p = 0.443, d = 0.03$) and cross-generalised decoding ($M = 0.489\ vs\ M = 0.491,\ t(22) = -0.16, p = 0.561, d = 0.04$) of shape and no differences in decoding ($M = 0.579\ vs\ M = 0.573$, $t(22) = 0.57, p = 0.714$, d=0.07) and cross-generalised decoding ($M = 0.569\ vs\ M = 0.563,\ t(22) = 0.42, p = 0.662, d = 0.06$) of XOR (**Fig. 4.4g, h**). Context decoding ($M = 0.533\ vs\ M = 0.508$, $t(22) = 2.37, p = 0.013, d = 0.44$) and cross-generalised decoding ($M = 0.52\ vs\ M = 0.489,\ t(22) = 2.54, p = 0.009, d = 0.5$) decreased from stage 1 to stage 4 when only correct trials were analysed (**Fig. 4.4g, h**). It is important to note that the changes to context coding might be related to an overt engagement of working memory processes in the context of an experimental design that includes working memory delays. The observed pattern of results, especially the lack of

learning-induced changes to the shape and XOR coding, thus provides further evidence that the minimal XOR representation found at the moment of the decision is not a product of a metabolically constrained optimisation. An alternative learning strategy could rely on using working memory to control neural dimensionality and geometry.

To test whether working memory engagement resulted in the reduction of neural dimensionality at the moment of the decision (i.e., the space-time trade-off hypothesis) we related the strength of mean cross-generalised context decoding in the delay-locked period to the mean shattering dimensionality (all trials) measured at the moment of the decision at a group level. In line with the space-time trade-off hypothesis, we found that the strength of abstract context decoding during maintenance was negatively related to neural dimensionality at the decision time in learning stage 1, $r(23) = -0.37, p = 0.04$ (**Fig. 4.4i**). However, this effect was not present in learning stage 4, $r(23) = 0.08, p = 0.63$ (**Fig. 4.4j**). This suggests that participants that maintained a stronger representation of abstract contextual information also exhibited lower dimensional representations at decision time.

Our previous non-human primate research showed that learning-induced changes to task representations are accompanied by changes to the distribution of selectivity across the PFC neural population (**Chapter 2**; Wójcik et al., n.d.). Motivated by this result, we explored next whether the acquisition of a minimal XOR representation at the moment of the decision in humans was mirrored by similar

changes in the distribution of selectivity exhibited by EEG sensors. It should be emphasised, however, that selectivity analyses performed in neuron space and in EEG sensor space lack conceptual parity for a host of reasons. Firstly, in contrast to single-cell firing data, EEG signal carries an intrinsic correlation structure reflecting spatial relations between sensors. Secondly, the two differ in terms of the level of analysis, as the method used in **Chapter 2** was performed on data sampled from one region (PFC) and the results reported here reflect the functioning of the entire cortex. Consequently, the interpretation of the measure is fundamentally different. In the case of neuron space, the characteristic of the selectivity distribution and how it changes provides information about the plasticity and specialisation of the sampled region, whereas selectivity in EEG sensor space reflects rather how computations are distributed across different regions of the cortex.

With these considerations foregrounded, we tested whether neural activity is better described by the random or minimal model (see **Chapter 2**, *Methods, generative models)*. Similarly as in the NHP experiment (**Chapter 2**; **Fig. 2.3**), we fitted a linear regression to our EEG data, in which task variables (colour, shape, and XOR) were used as predictors of the electrophysiological signal recorded at each sensor. We next computed the covariance between these selectivity coefficients for each learning stage (stage 1 and stage 4 shown in **Fig 4.4l, m**). Using a Euclidean distance measure (**Chapter 2**, *Methods, measuring similarity between selectivity distributions.*), we compared these covariance matrices to the covariances obtained from the random selectivity model (red dotted line; **Fig. 4.4n, o**) and minimal selectivity model (blue dotted line; **Fig. 4.4n, o**). We found that the

selectivity observed in all stages of learning is different from random mixed selectivity and that no learning-induced effects were detected, $M = 0.535\ vs\ M = 0.432,\ t(23) = 1.18, p = 0.876, d = 0.23$ (**Fig. 4.4n**; black lines). However, upon examining distance from minimal selectivity, we found that neural activity is progressing towards the minimal model over learning, $M = 1.184\ vs\ M = 0.94$, $t(23) = 3.76, p = 0.001, d = 0.73$ (**Fig. 4.4o**; black lines). The observed selectivity was not less dissimilar to minimal selectivity than the random model used in this analysis as a null hypothesis (**Fig. 4.4o**). It can be thus concluded that neural activity at the moment of the decision was strongly non-random in sensor space. As previously motioned, this is a natural consequence of the multicollinearity of EEG signals. This structured signal, seemed to progress slightly towards a covariance specified by the minimal model over learning which suggests a uniform engagement of cortical areas in the XOR computation.

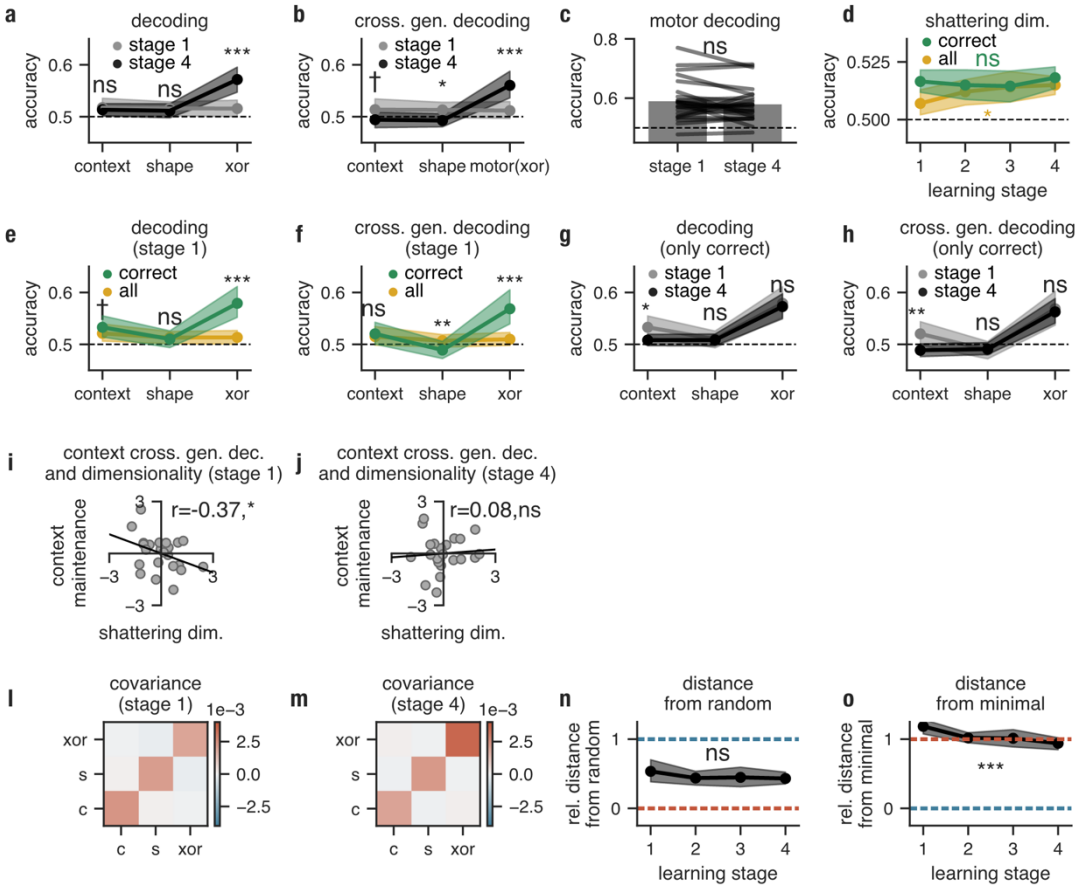


Figure 4.4. Neural geometry but not dimensionality changes over learning before the decision. **a**, Linear decoding of task variables for learning stages 1 (grey) and 4 (black). **b**, Cross-generalised linear decoding of task variables for learning stages 1 (grey) and 4 (black). **c**, Comparison of motor decoding (response) in stage 1 and stage 4; horizontal black lines represent value pairs for each participant. **d**, The learning dynamic of mean decoding of all possible task dichotomies excluding context, shape and XOR (shattering dimensionality) run on all trials (gold) and only correct trials (green); solid lines indicate the mean and shaded areas 95% confidence intervals over participants, respectively. **e**, Linear decoding of task variables in learning stage 1 run on all trials (gold) and only correct trials (green). **f**, Cross-generalised linear decoding of task variables in learning stage 1 run on all trials (gold) and only correct trials (green). **g**, **h**, Decoding analyses analogues to **a,b** but run only on correct trials. **i**,**j**, Correlation between normalised mean cross-generalised context decoding scores in the delay-locked period

(context maintenance) and the shattering dimensionality computed at the moment of the decision on all trials in stage 1 and stage 4, respectively. **l,m**, Covariance matrix of the selectivities computed from the data for stage 1 of learning and stage 4 of learning, respectively. **n**, Relative Euclidean distance between the covariance matrix of selectivity coefficients from the data and the covariance matrix expected from random selectivity (with matched total variance) as a function of learning (**Chapter 2**, *Methods, generative models*). Red and blue, respectively, error bars show the mean of the relative Euclidean distance between the covariance matrix of the random, respectively minimal, model (with matched total variance to the data) and the covariance matrix expected from random selectivity; the black line and shaded areas show mean and 95% confidence intervals over participants. **o**, Same as panel l but we show the relative distance from the covariance matrix expected from minimal selectivity (with matched total variance). All p-values were calculated using a t-test for dependent samples (a-h; l-m) and Pearson's correlation coefficient (***, $p < 0.01$; **, $p < 0.01$; *, $p < 0.05$; †, $p < 0.1$; $n.s.$, not significant).

Discussion

The recent convergence of neuroscience and machine learning has led to a novel conceptualisation of the role of working memory. It has been proposed that this process actively participates in improving the efficiency of computation rather than just being a passive data storage system (Dasgupta & Gershman, 2021). Our research echoes this perspective, suggesting that working memory selection shapes neural dimensionality, resulting in minimal and abstract task representation at decision time. Notably, these representations closely mirror the energy-efficient

neural codes observed in non-human primates following intensive training (**Chapter 2**; Wójcik et al., n.d.).

The feature selection mechanism discussed here is based on the assumption that irrelevant features are filtered out after they enter working memory. However, feature selection can also take place at the early perceptual stages through the engagement of selective attention. This is supported by empirical reports demonstrating that the PFC, specifically the inferior temporal junction, can exert top-down control over posterior cortical areas involved in visual processing (Zanto et al., 2010, 2011). Specifically, these frontal areas can enhance or diminish the processing of low-level physical features in visual cortices (from V1 to V5;Ruff et al., 2008) during the early stages of perceptual processing (~100ms) and thus control what information enters later stages of information processing such as working memory maintenance. The observed consequence of attention selection is the diminished representation of the irrelevant feature early in trial time. The early perceptual selection account seems, however, at odds with the described here data. In particular, the strength of colour (irrelevant) and context (relevant) decoding was similar during the presentation period of the coloured cue. Crucially, the irrelevant feature was only discarded during the delay after it entered working memory. Therefore, we believe that the selection mechanism observed here reflects working memory selection rather than selective attention.

In the present study, we also showed that human participants rapidly acquired an abstract representation of context over learning. It should be noted, however, that

learning here might not be an effect of slow updates to synaptic weights that accompany building a task representation from scratch – a process akin to metabolically constrained optimisation found in non-human primates (**Chapter 2**). Specifically, it is possible that human participants were already in the possession of a hierarchical XOR representation that could have been used to solve the task and the observed rapid learning speed reflected the alignment of the task representation stored in long-term memory to the new sensory instance. In line with this, our previous research in non-human primates demonstrated that after learning optimal, energy-efficient XOR representations, animals were able to rapidly alight the already acquired readout dimension to the representation of previously unseen stimuli combinations (**Chapter 3**; Wójcik et al., n.d.). An alternative explanation relates to the capacity of human and non-human primates to perform compositional generalisation (Frankland & Greene, 2019; Lake et al., 2017; Reverberi et al., 2012). That is, new, previously unseen problems can be solved by combining basic computational motifs that were previously learnt. Although compositionality has been proposed as a cohesive force that orchestrates higher cognition, experimental efforts examining multitask representations are scarce (Driscoll et al., 2022). This is since including multiple tasks in one experimental design reduces the number of trials that humans or animals can perform in each task and thus dramatically decreases the statistical power necessary to extract the representations of each task. A renewed enthusiasm for compositionality was kindled by the rise of deep neural network models which began to shed light on computational mechanisms underlining compositional generalisation (Yang et al., 2019). Recently, a dynamical systems perfective was

applied to explore representations acquired by RNNs trained on multiple tasks (Driscoll et al., 2022). The models decomposed the tasks into a subset of abstract dynamical motifs such as attractor states, input amplifications, rotations, and discriminant boundaries. Each task was solved through smooth input-driven transitions between these dynamical motifs. Furthermore, tasks that were previously not encountered, could be solved by rapidly repurposing these already learnt motifs. Future empirical research could focus on testing these theoretically principled predictions provided by RNN models in settings utilising chronic, lifelong recordings in animals(Steinmetz et al., 2021) or convolution-based designs in human EEG/MEG (Ruesseler et al., n.d.).

Our findings, emphasizing the role of working memory in model-based decision-making, align with recent research exploring different factors influencing the arbitration between model-free and model-based strategies (Doll et al., 2015; Otto et al., 2013; Smittenaar et al., 2013). Several studies have suggested that employing a model-based learning strategy depends strongly on the currently available computational resources. For example, when the working memory of human participants was overloaded with content that was not related to the task at hand, they were more likely to use a model-free approach (Otto et al., 2013). Furthermore, it has been shown that the PFC, an area consistently shown to be involved in working memory processes, is deterministically associated with controlling the balance between model-free and model-based learning (Smittenaar et al., 2013). Recent advances in psychology extend this notion, demonstrating that the learning strategy that human participants employ, is strongly dependent

on prior instructions (Castro-Rodrigues et al., 2022) and that humans can flexibly switch between strategies to reduce resource expenditure (Feher da Silva et al., 2023). Together these findings highlight the role of meta-reasoning in model-based vs model-free strategy arbitration. This process has been defined as the monitoring, evaluation and possible modification of the agent's own cognitive process involved in solving a particular problem (Griffiths et al., 2019). This can be achieved by maximising the expected value of an action while minimising the time and resources needed to perform these computations which can be formalised as the value of computation (Griffiths et al., 2019). Future studies could explore whether involving working memory, through the introduction of delays in tasks, modulates the value of computation in a manner that favours model-based solutions (Braver, 2012).

Finally, dimensionality reduction can be implemented in several ways in neural systems. Alignment to the selection vector (Mante et al., 2013; Stroud, Watanabe, et al., n.d.) as well as the introduction of metabolic cost as a regulariser (Stroud, Wójcik, et al., n.d.; Wójcik et al., n.d.) both represent optimisation-based approaches that require incremental updates to synaptic weights and mirror a slow learning process in biological systems. These mechanisms seem at odds with the learning speed with which the optimal representation was acquired in the data presented here, as participants were able to generate an energy-efficient solution in the first learning stage of the experiment. An alternative explanation for these discrepancies might be provided by a family of models that can reproduce realistic maintenance dynamics without the necessity of optimising recurrent weights

(reservoir computing networks; Enel et al., 2016). Specifically, it has been shown that combining a randomly connected pool of recurrent units with feedback control is sufficient to generate attractor states necessary for solving tasks with delays (Enel et al., 2016). This was achieved by allowing a context readout to be fed back as input to the model. Although reservoir computing is often characterised by an increased learning speed compared to fully optimised models, it is unclear whether these models can reproduce such fundamental computational properties of working memory processes like the sensory-to-abstract coding transformation. In principle, the feedback-driven attractor states could be dominated early in trial. It is important to note here, however, that contextual feedback control relies on the availability of a context oracle to explicitly inform the model which variable should be maintained in an attractor state. Future theoretical studies should thus not only explore whether mixed reservoir regimes can reproduce sensory-to-abstract coding transformations, but also investigate possible unsupervised techniques of context estimation (Collins & Frank, 2016; Rios & Itti, 2020).

While the data presented here shed light on how working memory processes could be utilised to control neural dimensionality, there are potential limitations that should be considered. Firstly, to establish a causal role of working memory selection in reducing the dimensionality of neural activity at the decision timepoint, a control condition could be introduced, whereby the colour information persisted throughout the entire trial. The neural representations and their learning dynamics observed in the current study could be then compared to representation acquired in the second, continuous colour input setting. Secondly, the low difficulty of the

task, and consequently a small number of error trials, prevented us from performing a dimensionality analysis on representations that led subsequently to erroneous responses. The difficulty of the task could be increased by reducing the memory delay period. Finally, the motor response signal and the XOR signal were collinear in the study after learning. These two features could be disentangled by introducing a third-level policy and thus a second interaction term. In the current design, the XOR feature represents an interaction between the shape and the colour features. Displaying two abstract stimuli (e.g., “#” and “&” signs) representing the XOR (True vs False) and randomising the sides on which they would be presented (left vs right) would provide a marker of the XOR that is not confounded by motor code.

Building on the results described here, future studies could explore the relationship between working memory processes and low-dimensional task representations using perturbation approaches. Specifically, techniques such as TMS (Rose et al., n.d.) or pinging (Wolff et al., 2017b) have been successfully used in the past to reactivate information stored in working memory in an activity-silent format (Stokes, 2015). Using these methods to non-specifically perturb the system right after the stimulus presentation could reveal, for example, whether discarding irrelevant features is related to the decay of neural activity (i.e., in the early delay) (Souza & Oberauer, 2015).

Chapter 5: General Discussion

The ability of primates, particularly humans, to rapidly adapt to their environment is a key line of research in both neuroscience and psychology. In this thesis, I explored the neural mechanisms that facilitate such adaptation across different time scales, and ultimately lead to behaviours we instinctively recognise as intelligent. Such neural processes often emerge from the complex interactions between various brain regions. In this context, the prefrontal cortex stands out as a crucial area consistently linked to higher cognitive functions. In the research described in this thesis, I utilised a machine learning perspective to design and analyse empirical studies which used both humans and non-human primate models to explore how the PFC supports intelligent behaviour. In the following section, I will discuss the insights gained from each chapter as well as their collective contributions to our understanding of PFC function. I will then reflect on the limitations of the machine learning perspective that were evident in the empirical efforts described here. Finally, having these considerations foregrounded, I will reflect on the future of the cycle of mutual inspiration observed between machine learning and neuroscience.

Chapter summaries and contributions

In **Chapter 1,** I reviewed research on human evolution to identify the ecological niche humans occupied and, by extension, the cognitive challenges their brains evolved to overcome. This exploration revealed that humans evolved in a

generalist-specialist ecological landscape which required the development of both a general-purpose cognitive toolkit as well as a capacity to specialise in a specific environment. I theorised that these seemingly contradictory processes can both be implemented by the same neural system, by orchestrating generalist and specialist processes at different time scales. Generalist processes were well approximated by utilising the activation dynamics of a reservoir network to generate high-dimensional representations. In turn, specialist processes were operationalised by slow, metabolically constrained optimisation. Placing reservoir computing and optimisation within the broader vistas of theories such as complementary learning system theory and feed-back control led to the formulation of the core hypothesis of this thesis. More specifically, I postulated that at the beginning of learning, when an agent is confronted with a novel problem, a reservoir-based, high-dimensional strategy will be employed by the PFC. As learning progresses, however, the local connections within the PFC will be slowly updated to arrive at a minimal and task-optimised representation.

This hypothesis was tested in a non-human primate experiment described in **Chapter 2**. Two macaques learnt to combine two input features to generate an XOR representation. In line with our predictions, at the beginning of learning, all task features were uniformly coded by the PFC – a representation akin to codes generated by a randomly connected reservoir. As learning progressed, the neural codes became low-dimensional, representing predominantly the XOR dimension – which was the only feature relevant for task performance at the decision time. Importantly, this transition from high-dimensional and task-agnostic to low-

dimensional and task-optimised was predicted by a metabolically constrained neural network. These findings support the notion that the PFC can house generalist and specialist representational strategies at different timescales.

One of the predictions of the complementary learning systems theory posits that information stored in neocortical regions, including the PFC, should capture the abstract structure of the environment. Neocortical memories are predicted to have a high potential for cross-generalisation and thus facilitate the learning of new but related problems. I tested this hypothesis in a non-human primate described in **Chapter 3**. Using the same animals as in the study described in **Chapter 2** allowed to test whether the previously acquired XOR representation could be utilised to facilitate the learning of a new sensory instance of the same XOR structure. We found that the same XOR readout was used by the PFC to classify the new and old stimuli sets early in learning. Moreover, the results revealed that the animals also cross-generalised the dynamical landscape used to construct the XOR signal. These findings support the notion that abstract information stored in the PFC supports generalisation.

In **Chapter 4,** I discussed a machine-learning view conceptualising working memory as a computational resource, rather than a static storage mechanism. I hypothesised that working memory can be used to temporally decompose a complex problem into several simple operations, to reduce the computational complexity at decision time. This idea was tested using EEG in humans that learnt to solve an XOR problem. We found that working memory processes were used to

filter out irrelevant stimulus dimensions and then to maintain only task-relevant information throughout the delay. Participants who exhibited more pronounced maintenance processes used representations of reduced dimensionality at decision time. Crucially, representations at decision time strongly resembled the low-dimensional and energy-efficient codes found in non-human primates after extensive training (**Chapter 2**).

In this thesis, the machine learning framework was employed to conceptualise the neural processes supporting adaptive and intelligent behaviours. This perspective was invaluable, as it not only allowed to directly model the learning dynamics observed in primates, but also served as a conceptual lens illuminating the fundamental principles of efficient computation. It is crucial, however, to recognise the inherent limitations of this framework, as reflected in some of the presented empirical findings.

Limitations of the machine-learning perspective

A core concept in the machine learning perspective on brain function is the view that experience-induced changes to neural structures can be captured by optimisation processes (Hennig et al., 2021; Richards et al., 2019; A. Zador et al., 2023). While providing a cogent paradigm for exploring learning mechanisms, optimisation is not devoid of its challenges and criticisms (Hennig et al., 2021). For instance, several empirical reports demonstrate that neural activity has a covariance structure that persists over long periods of time (Hennig et al., 2021). Crucially, this covariance might be tangentially linked, or even unrelated, to the

task at hand (Hennig et al., 2021). Such a pattern suggests that the inherent stability in neural responses may limit the system's adaptability, potentially constraining the range of learnable representational states. A study using brain-computer interface in non-human primates confirmed this intuition (Sadtler et al., 2014). The experimental design of this study allowed the researchers to casually test whether aligning the task structure to the structure of neural covariance influences learning. In some conditions, the task was easily solved with representations that were based on the connectivity structures already embedded in the neural population. Conversely, when the task necessitated representational patterns misaligned with existing neural structures, learning faced substantial impediments.

These results suggest that the dynamics of actual neural populations are not fully amenable to experience-induced change and thus defy the characterisation offered by a pure optimisation view. In **Chapter 2** we found that PFC optimises neural representations for task performance as a function of learning. However, in line with the considerations above, we observed that this process has limits, as some task features remained resistant to elimination. More precisely, even though colour was irrelevant for task performance at decision time, colour coding persisted throughout the entire learning process. This might reflect the general importance of cue stimuli that the animals learnt prior to the experiment. The persistence of colour coding suggests that neural structures within the PFC might be less flexible than expected from a fully task-optimised system and that future learning can be strongly constrained by the already present connectivity.

A second aspect that argues against the prevailing optimisation narrative of learning is the notable variability inherent in biological systems. This variability seems at odds with what one would expect from systems predicated on fully task-optimized representations. For instance, humans are not immune to occasional errors, even after mastering a particular skill. Some posit that this variability is not merely incidental or a result of systemic noise but might be an evolved adaptive mechanism (Dhawale et al., 2017). Such a mechanism would be especially advantageous in unpredictable environments with changing reward landscapes. Notably, empirical research suggests that many species, including humans, can adjust their behavioural variability in response to the stability of their environment (as discussed in Hennig et al., 2021). The objective function discussed in **Chapter 1** – which balances performance maximisation against metabolic cost reduction – appears insufficient in replicating the variability seen in many electrophysiological studies or the explorative tendencies of primate behaviours.

The *Pure Learning Critique* (A. M. Zador, 2019) also presents a compelling counterargument to the optimisation-based view of learning. Traditionally, artificial neural networks are initialised with random connections, which are incrementally refined using numerous examples via supervised learning methods. This model bears only a superficial resemblance to the learning trajectories of young humans or animals. While guidance from caregivers might mirror supervised learning in the early stages of a young one's life, the emphasis on such supervision is far less pronounced in biological agents compared to artificial systems. To draw a parallel,

if a child's learning process was to mimic that of a standard convolutional network, they would need to continuously match a visual input with a corresponding label every second throughout their entire existence (A. M. Zador, 2019). Learning in biological agents is instead based on leveraging millions of years of experience encoded in the genetically predetermined neural connectivity (A. Zador et al., 2023; A. M. Zador, 2019). In small-brained species, information about the environment can be fully encoded in the genome and then transferred to new generations. The storage capacity of the genome is, however, not sufficient to fully determine the connectivity of large brains (A. M. Zador, 2019). Instead, in large-brained species, the genetic code is used to encode compressed and abstract information about connectivity through wiring principles and connectivity motifs. The storage limit imposing constrains on the format of stored information is called the *genomic bottleneck* (A. M. Zador, 2019). The *pure learning critique* thus posits that the supreme intelligence and adaptivity of biological compared to artificial agents stems from combining innate and learned mechanisms. Specifically, the genetically predetermined connectivity acts as a regulariser for future learning. A good example in this context are place cells. Primates are genetically predisposed to develop place cells, but it is only through learning that these types of neurons will start to encode specific locations.

Highlighting instances where the machine learning framework falls short in capturing neural processes found within biological agents does not imply that this perspective should be dismissed. Instead, such instances present opportunities to refine and expand machine learning models. By aligning them closer to biological

phenomena, we might render these models more adaptable and less brittle (Summerfield, 2022a).

Future directions

Most artificial neural networks used in neuroscience are initialised with a random connectivity matrix and thus make the implicit assumption of the modelled agent being a *tabula rasa* system. Networks constructed to perform transfer constitute an exception in this context (Pan & Yang, 2009). However, these models differ strongly from how biological networks utilise innate connectivity. Specifically, the learnt representation of the previous task is fully encoded in the entire or part of the connectivity matrix, whereas information transmitted genetically from one generation to the next generation of complex organisms needs to pass through the previously mentioned *genetical bottleneck* (Barabási et al., 2023; A. M. Zador, 2019). The capacity limit of DNA imposes constraints on the amount and thus the format of the encoded information (A. M. Zador, 2019). It is likely that this leads to developing priors for representations that are both more generalisable and robust to perturbations. Future research combining neuroscience and artificial intelligence could explore how to model the transformation of learnt information through the genomic bottleneck using connectivity compression methods (Barabási et al., 2023; A. M. Zador, 2019).

A promising research direction in this context is exemplified by studies exploring the influence of canonical connectivity motifs on neural computation (Bojanek et

al., 2020; Braganza & Beck, 2018). Detailed empirical research efforts examining the connectivity within neural populations have found that many of these populations contain small connectivity motifs in numbers surpassing what might be expected from a randomly connected network (S. Song et al., 2005). For example, several types of neural populations display a high prevalence of 3-cell connectivity motifs with varied directionality (S. Song et al., 2005). Some research categorizes these as cycles, middleman, fan-in, or fan-out motifs (Bojanek et al., 2020), underscoring their unique computational roles at a local level (Braganza & Beck, 2018). Notably, these connectivity motifs are simple enough to be encoded via a compressed format that can pass through the genomic bottleneck. Future research could probe how the distribution of these motifs influences the computations performed by neural systems. Innate connectivity represented via simple motifs could serve as an inductive bias, not only enhancing model's performance but also conserving training resources. Moreover, this line of research could furnish insights into the synergic relationship between innate and acquired information, elucidating the advanced learning capabilities evident in biological agents. This exploration also holds value for the broader machine learning community, potentially offering a method to compress the parameters of large models, which typically encompass millions of parameters and require significant resources to train.

Recent advances exploring inductive biases in the context of predetermined connectivity also focus on low-rank networks. These models try to reproduce the observation that biological neural connectivity never appears to be fully structured or entirely random, but rather oscillates between different mixtures of both types of

connectivity (Mastrogiuseppe & Ostojic, 2018). Low-rank models not only reproduce the low-dimensional neural dynamics observed in the primate brain (Gao & Ganguli, 2015) but also capture the heterogeneity of mixed selectivity neurons observed in areas such as the PFC (Rigotti et al., 2013). Importantly, models that exhibit a mixture of random and structured connectivity also exhibit an increased potential for cross-task generalisation (Mastrogiuseppe & Ostojic, 2018). Low-rank networks, due to their tractability and interpretability, are primary used as a theoretical and analysis framework, rather than a model of an inductive bias (Langdon et al., 2023). This is because the biological mechanism that leads to what appears to be low-rank connectivity in the brain is yet to be identified (Langdon et al., 2023). Future research could focus on exploring these mechanisms.

The integration of neuroscience and machine learning

Although the introduction of machine learning into neuroscience provides us with a flurry of exciting theoretical and empirical findings, the convergence of these fields is a topic of intense debate. This becomes particularly evident when discussing the impact that neuroscience and psychology should have on the pursuit of artificial general intelligence (AGI; Summerfield, 2022). Many experts with a neuroscience background highlight the value of conceptual tools developed within their field, particularly those addressing phenomena that span various levels of abstraction. They also emphasize the significance of model organisms with general intelligence capabilities, such as humans and certain primates. They argue that by understanding the workings of intelligence in these biological systems, we

can better replicate them in artificial models (Richards et al., 2019; A. Zador et al., 2023).

Conversely, critics of this integration often point to the recent successes of deep learning models that have no close links to biology (Summerfield, 2022a). Several of these models demonstrate analogues to higher cognitive skills, like language generation or complex decision-making (Bommasani et al., 2021). Yet, these skills often manifest in ways that do not mimic biological processes. Such experts posit that the path to true artificial general intelligence may lie more in enhancing the computational power of artificial agents and veering away from biological inspirations.

Regardless of one's stance on the importance of neuroscience in solving AGI, the integration of machine learning and neuroscience might be influenced by factors beyond the theoretical debate. Specifically, advancements in neuro-technologies, such as brain-computer interfaces (BCIs) and advanced recording devices (Steinmetz et al., 2021), might usher in a market-driven emphasis on neuroscientific principles within machine learning.

Conclusions

Emerging research on intelligent and adaptive behaviour, as elucidated in this thesis, accentuates its profound complexity and dynamical character. The machine

learning perspective, although forward-thinking, still faces challenges in reproducing the riches of neural processes observed in intelligent organisms like primates. However, casting aside the machine learning perspective due to these shortcomings would be rushed. By weaving together insights from neuroscience and the capabilities of artificial neural networks, there lies an opportunity to create a holistic model of intelligence, one that combines the robustness of biology with the systematic clarity offered by machine learning methods.

References

Acharya, J. N., Hani, A., Cheek, J., Thirumala, P., & Tsuchida, T. N. (2016). American Clinical Neurophysiology Society Guideline 2: Guidelines for Standard Electrode Position Nomenclature. *Journal of Clinical Neurophysiology*, *33*(4), 308–311. https://doi.org/10.1097/WNP.0000000000000316

Allen, R. J., & Waclaw, B. (2018). Bacterial growth: a statistical physicist's guide. *Reports on Progress in Physics*, *82*(1), 016601.

Asaad, W. F., Rainer, G., & Miller, E. K. (1998). Neural Activity in the Primate Prefrontal Cortex during Associative Learning). Medial temporal structures critical for long-term memories are also important: dam-age to the hippocampus and/or subjacent cortex (Mur. In *Neuron* (Vol. 21).

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. In *Annual Review of Neuroscience* (Vol. 28, pp. 403–450). https://doi.org/10.1146/annurev.neuro.28.061604.135709

Badre, D., Bhandari, A., Keglovits, H., & Kikumoto, A. (2021). The dimensionality of neural representations for control. In *Current Opinion in Behavioral Sciences* (Vol. 38, pp. 20–28). Elsevier Ltd. https://doi.org/10.1016/j.cobeha.2020.07.002

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., Pritzel, A., Chadwick, M. J., Degris, T., & Modayil, J. (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature*, *557*(7705), 429–433.

Barabási, D. L., Beynon, T., Katona, Á., & Perez-Nieves, N. (2023). Complex computation from developmental priors. *Nature Communications*, *14*(1). https://doi.org/10.1038/s41467-023-37980-1

Barak, O., Rigotti, M., & Fusi, S. (2013). The sparseness of mixed selectivity neurons controls the generalization-discrimination trade-off. *Journal of Neuroscience*, *33*(9), 3844–3856. https://doi.org/10.1523/JNEUROSCI.2753-12.2013

Bar-On, Y. M., Phillips, R., & Milo, R. (2018). The biomass distribution on Earth. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(25), 6506–6511. https://doi.org/10.1073/pnas.1711842115

Bartolo, R., Saunders, R. C., Mitz, A. R., & Averbeck, B. B. (2020). Dimensionality, information and learning in prefrontal cortex. *PLoS Computational Biology*, *16*(4). https://doi.org/10.1371/journal.pcbi.1007514

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., & Salzman, C. D. (2020). The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell*, *183*(4), 954-967.e21. https://doi.org/10.1016/j.cell.2020.09.031

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K. M., & Robbins, K. A. (2015). The PREP pipeline: Standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics*, *9*(JUNE), 1–19. https://doi.org/10.3389/fninf.2015.00016

Birney, D. P., & Beckmann, J. F. (2022). Intelligence IS Cognitive Flexibility: Why Multilevel Models of Within-Individual Processes Are Needed to Realise

This. *Journal of Intelligence*, *10*(3). https://doi.org/10.3390/jintelligence10030049

Bocincova, A., & Johnson, J. S. (2019). The time course of encoding and maintenance of task-relevant versus irrelevant object features in working memory. *Cortex*, *111*, 196–209. https://doi.org/10.1016/j.cortex.2018.10.013

Bojanek, K., Zhu, Y., & MacLean, J. (2020). Cyclic transitions between higher order motifs underlie sustained asynchronous spiking in sparse recurrent networks. *PLoS Computational Biology*, *16*(9). https://doi.org/10.1371/journal.pcbi.1007409

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). *On the Opportunities and Risks of Foundation Models*. http://arxiv.org/abs/2108.07258

Bostan, A. C., & Strick, P. L. (2018). The basal ganglia and the cerebellum: nodes in an integrated network. *Nature Reviews Neuroscience*, *19*(6), 338–350.

Braganza, O., & Beck, H. (2018). The Circuit Motif as a Conceptual Tool for Multilevel Neuroscience. In *Trends in Neurosciences* (Vol. 41, Issue 3, pp. 128–136). Elsevier Ltd. https://doi.org/10.1016/j.tins.2018.01.002

Braver, T. S. (2012). The variable nature of cognitive control: A dual mechanisms framework. In *Trends in Cognitive Sciences* (Vol. 16, Issue 2, pp. 106–113). https://doi.org/10.1016/j.tics.2011.12.010

Buckner, R. L., & Krienen, F. M. (2013). The evolution of distributed association networks in the human brain. In *Trends in Cognitive Sciences* (Vol. 17, Issue 12, pp. 648–665). https://doi.org/10.1016/j.tics.2013.09.017

Buonomano, D. V., & Maass, W. (2009). State-dependent computations: Spatiotemporal processing in cortical networks. In *Nature Reviews Neuroscience* (Vol. 10, Issue 2, pp. 113–125). https://doi.org/10.1038/nrn2558

Castro-Rodrigues, P., Akam, T., Snorasson, I., Camacho, M., Paixão, V., Maia, A., Barahona-Corrêa, J. B., Dayan, P., Simpson, H. B., Costa, R. M., & Oliveira-Maia, A. J. (2022). Explicit knowledge of task structure is a primary determinant of human model-based action. *Nature Human Behaviour*, *6*(8), 1126–1141. https://doi.org/10.1038/s41562-022-01346-2

Chaisangmongkon, W., Swaminathan, S. K., Freedman, D. J., & Wang, X. J. (2017). Computing by Robust Transience: How the Fronto-Parietal Network Performs Sequential, Category-Based Decisions. *Neuron*, *93*(6), 1504-1517.e4. https://doi.org/10.1016/j.neuron.2017.03.002

Chaudhuri, R., Gerçek, B., Pandey, B., Peyrache, A., & Fiete, I. (2019). The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep. *Nature Neuroscience*, *22*(9), 1512–1520. https://doi.org/10.1038/s41593-019-0460-x

Collins, A. G. E., & Frank, M. J. (2016). Neural signature of hierarchically structured expectations predicts clustering and transfer of rule sets in reinforcement learning. *Cognition*, *152*, 160–169.

Constantinidis, C., Qi, X.-L., & Meyer, T. (2016). Single-neuron spike train recordings from macaque prefrontal cortex during a visual working memory task before and after training. *CRCNS. Org*.

Corrigan, B. W., Gulli, R. A., Doucet, G., Roussy, M., Luna, R., Pradeepan, K. S., Sachs, A. J., & Martinez-Trujillo, J. C. (2022). Distinct neural codes in primate hippocampus and lateral prefrontal cortex during associative learning in virtual environments. *Neuron*, *110*(13), 2155-2169.e4. https://doi.org/10.1016/j.neuron.2022.04.016

Cromer, J. A., Roy, J. E., & Miller, E. K. (2010). Representation of Multiple, Independent Categories in the Primate Prefrontal Cortex. *Neuron*, *66*(5), 796–807. https://doi.org/10.1016/j.neuron.2010.05.005

Cueva, C. J., Saez, A., Marcos, E., Genovesio, A., Jazayeri, M., Romo, R., Salzman, C. D., Shadlen, M. N., & Fusi, S. (2020). Low-dimensional dynamics for working memory and time encoding. *PNAS*, *117*(37), 23021–23032. https://doi.org/10.1073/pnas.1915984117/-/DCSupplemental.y

Dasgupta, I., & Gershman, S. J. (2021). Memory as a Computational Resource. In *Trends in Cognitive Sciences* (Vol. 25, Issue 3, pp. 240–251). Elsevier Ltd. https://doi.org/10.1016/j.tics.2020.12.008

Dekker, R. B., Otto, F., & Summerfield, C. (2022). *Curriculum learning for human compositional generalization*. https://doi.org/10.1073/pnas

Dhawale, A. K., Smith, M. A., & Ölveczky, B. P. (2017). The role of variability in motor learning. *Annual Review of Neuroscience*, *40*, 479–498.

di Pellegrino, G., & Wise, S. P. (1993). Visuospatial versus Visuomotor Activity in the Premotor and Prefrontal Cortex of a Primate. In *The Journal of Neuroscience* (Vol. 13, Issue 3).

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature Neuroscience*, *18*(5), 767–772. https://doi.org/10.1038/nn.3981

Driscoll, L., Shenoy, K., & Sussillo, D. (2022). Flexible multitask computation in recurrent networks utilizes shared dynamical motifs. *BioRxiv*, 2022–2028.

Dubreuil, A., Valente, A., Beiran, M., Mastrogiuseppe, F., & Ostojic, S. (2021). The role of population structure in computations through neural dynamics. *BioRxiv*, 2020.07.03.185942. https://doi.org/10.1101/2020.07.03.185942

Dubreuil, A., Valente, A., Beiran, M., Mastrogiuseppe, F., & Ostojic, S. (2022). The role of population structure in computations through neural dynamics. *Nature Neuroscience*, *25*(6), 783–794. https://doi.org/10.1038/s41593-022-01088-4

Dunbar, R. I. M. (2003). The social brain: Mind, language, and society in evolutionary perspective. *Annual Review of Anthropology*, *32*, 163–181. https://doi.org/10.1146/annurev.anthro.32.061002.093158

Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. *Nature Reviews Neuroscience*, *2*(11), 820–829. https://doi.org/10.1038/35097575

Ehrlich, D. B., & Murray, J. D. (n.d.). *Geometry of neural computation unifies working memory and planning*. https://doi.org/10.1101/2021.02.01.429156

Eichenbaum, H. (2017). Prefrontal–hippocampal interactions in episodic memory. *Nature Reviews Neuroscience*, *18*(9), 547–558.

Enel, P., Procyk, E., Quilodran, R., & Dominey, P. F. (2016). Reservoir Computing Properties of Neural Dynamics in Prefrontal Cortex. *PLOS Computational Biology*, *12*(6), e1004967. https://doi.org/10.1371/JOURNAL.PCBI.1004967

Engel, T. A., Chaisangmongkon, W., Freedman, D. J., & Wang, X.-J. (2015). Choice-correlated activity fluctuations underlie learning of neuronal category representation. *Nature Communications*, *6*(1), 6454.

Feher da Silva, C., Lombardi, G., Edelson, M., & Hare, T. A. (2023). Rethinking model-based and model-free influences on mental effort and striatal prediction errors. *Nature Human Behaviour*, *7*(6), 956–969. https://doi.org/10.1038/s41562-023-01573-1

Fiebig, F., & Lansner, A. (2017). A spiking working memory model based on hebbian short-term potentiation. *Journal of Neuroscience*, *37*(1), 83–96. https://doi.org/10.1523/JNEUROSCI.1989-16.2016

Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., & Summerfield, C. (2022). Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, *110*(7), 1258-1270.e11. https://doi.org/10.1016/j.neuron.2022.01.005

Frankland, S. M., & Greene, J. D. (2019). *Language of Thought*. https://doi.org/10.1146/annurev-psych-122216

Frankland, S. M., & Greene, J. D. (2020). Concepts and compositionality: in search of the brain's language of thought. *Annual Review of Psychology*, *71*, 273–303.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, *3*(4), 128–135.

Frischkorn, G. T., Wilhelm, O., & Oberauer, K. (2022). Process-oriented intelligence research: A review from the cognitive perspective. *Intelligence*, *94*. https://doi.org/10.1016/j.intell.2022.101681

Fusi, S., Miller, E. K., Rigotti, M., Karpova, A., & Kiani, R. (2016). Why neurons mix: high dimensionality for higher cognition This review comes from a themed issue on Neurobiology of behavior. *Current Opinion in Neurobiology*, *37*, 66–74. https://doi.org/10.1016/j.conb.2016.01.010

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

Ganguli, S., Bisley, J. W., Roitman, J. D., Shadlen, M. N., Goldberg, M. E., & Miller, K. D. (2008). One-Dimensional Dynamics of Attention and Decision Making in LIP. *Neuron*, *58*(1), 15–25. https://doi.org/10.1016/j.neuron.2008.01.038

Gao, P., & Ganguli, S. (2015). On simplicity and complexity in the brave new world of large-scale neuroscience. In *Current Opinion in Neurobiology* (Vol. 32, pp. 148–155). Elsevier Ltd. https://doi.org/10.1016/j.conb.2015.04.003

Gardner, H. E. (2011). *Frames of mind: The theory of multiple intelligences*. Basic books.

Ghashghaei, H. T., Hilgetag, C. C., & Barbas, H. (2007). Sequence of information processing for emotions based on the anatomic dialogue between prefrontal cortex and amygdala. *Neuroimage*, *34*(3), 905–923.

Girardeau, G., Benchenane, K., Wiener, S. I., Buzsáki, G., & Zugaro, M. B. (2009). Selective suppression of hippocampal ripples impairs spatial memory. *Nature Neuroscience*, *12*(10), 1222–1223.

Goldman, M. S. (2009). Memory without feedback in a neural network. *Neuron*, *61*(4), 621–634.

Goldman-Rakic, P. S. (1995). Cellular Basis of Working Memory Review. In *Neuron* (Vol. 14).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7 DEC*. https://doi.org/10.3389/fnins.2013.00267

Greedy, W., Zhu, H. W., Pemberton, J., Mellor, J., & Ponte Costa, R. (2022). Single-phase deep learning in cortico-cortical networks. *Advances in Neural Information Processing Systems*, *35*, 24213–24225.

Griffiths, T. L., Callaway, F., Chang, M. B., Grant, E., Krueger, P. M., & Lieder, F. (2019). Doing more with less: meta-reasoning and meta-learning in humans and machines. In *Current Opinion in Behavioral Sciences* (Vol. 29, pp. 24–30). Elsevier Ltd. https://doi.org/10.1016/j.cobeha.2019.01.005

Haith, A. M., & Krakauer, J. W. (2013). Model-based and model-free mechanisms of human motor learning. In *Progress in motor control: Neural, computational and dynamic approaches* (pp. 1–21). Springer.

Hara, Y., Yamaguchi, K., Onimaru, K., Kadota, M., Koyanagi, M., Keeley, S. D., Tatsumi, K., Tanaka, K., Motone, F., Kageyama, Y., Nozu, R., Adachi, N., Nishimura, O., Nakagawa, R., Tanegashima, C., Kiyatake, I., Matsumoto, R., Murakumo, K., Nishida, K., … Kuraku, S. (2018). Shark genomes provide insights into elasmobranch evolution and the origin of vertebrates. *Nature Ecology and Evolution*, *2*(11), 1761–1771. https://doi.org/10.1038/s41559-018-0673-5

Hebb, D. O. (1942). The effect of early and late brain injury upon test scores, and the nature of normal adult intelligence. *Proceedings of the American Philosophical Society*, 275–292.

Hennig, J. A., Oby, E. R., Losey, D. M., Batista, A. P., Yu, B. M., & Chase, S. M. (2021). How learning unfolds in the brain: toward an optimization view. In *Neuron* (Vol. 109, Issue 23, pp. 3720–3735). Cell Press. https://doi.org/10.1016/j.neuron.2021.09.005

Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science*, *317*(5843), 1360–1366.

Hirokawa, J., Vaughan, A., Masset, P., Ott, T., & Kepecs, A. (2019). Frontal cortex neuron types categorically encode single decision variables. *Nature*, *576*(7787), 446–451. https://doi.org/10.1038/s41586-019-1816-9

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, *81*(10), 3088–3092.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*(3), 574.

Hunt, L. T., Daw, N. D., Kaanders, P., MacIver, M. A., Mugan, U., Procyk, E., Redish, A. D., Russo, E., Scholl, J., Stachenfeld, K., Wilson, C. R. E., & Kolling, N. (2021). Formalizing planning and information search in naturalistic decision-making. In *Nature Neuroscience* (Vol. 24, Issue 8, pp. 1051–1064). Nature Research. https://doi.org/10.1038/s41593-021-00866-w

Jacob, S. N., & Nieder, A. (2014). Complementary roles for primate frontal and parietal cortex in guarding working memory from distractor stimuli. *Neuron*, *83*(1), 226–237. https://doi.org/10.1016/j.neuron.2014.05.009

Jas, M., Engemann, D. A., Bekhti, Y., Raimondo, F., & Gramfort, A. (2017). Autoreject: Automated artifact rejection for MEG and EEG data. *NeuroImage*, *159*, 417–429. https://doi.org/10.1016/j.neuroimage.2017.06.030

Jonas, E., & Kording, K. P. (2017). Could a Neuroscientist Understand a Microprocessor? *PLoS Computational Biology*, *13*(1). https://doi.org/10.1371/journal.pcbi.1005268

Kao, T. C., & Hennequin, G. (2019). Neuroscience out of control: control-theoretic perspectives on neural circuit dynamics. In *Current Opinion in Neurobiology* (Vol. 58, pp. 122–129). Elsevier Ltd. https://doi.org/10.1016/j.conb.2019.09.001

Kao, T. C., Sadabadi, M. S., & Hennequin, G. (2021). Optimal anticipatory control as a theory of motor preparation: A thalamo-cortical circuit model. *Neuron*, *109*(9), 1567-1581.e12. https://doi.org/10.1016/j.neuron.2021.03.009

Kasai, H., Ucar, H., Morimoto, Y., Eto, F., & Okazaki, H. (2023). Mechanical transmission at spine synapses: Short-term potentiation and working memory. In *Current Opinion in Neurobiology* (Vol. 80). Elsevier Ltd. https://doi.org/10.1016/j.conb.2023.102706

Kaufman, M. T., Benna, M. K., Rigotti, M., Stefanini, F., Fusi, S., & Churchland, A. K. (2022). The implications of categorical and category-free mixed selectivity on representational geometries. In *Current Opinion in Neurobiology* (Vol. 77). Elsevier Ltd. https://doi.org/10.1016/j.conb.2022.102644

Khona, M., & Fiete, I. R. (2022). Attractor and integrator networks in the brain. *Nature Reviews Neuroscience*, *23*(12), 744–766.

Kim, J. S., Greene, M. J., Zlateski, A., Lee, K., Richardson, M., Turaga, S. C., Purcaro, M., Balkam, M., Robinson, A., & Behabadi, B. F. (2014). Space–time wiring specificity supports direction selectivity in the retina. *Nature*, *509*(7500), 331–336.

Klyubin, A. S., Polani, D., & Nehaniv, C. L. (2005). Empowerment: A universal agent-centric measure of control. *2005 Ieee Congress on Evolutionary Computation*, *1*, 128–135.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*. https://doi.org/10.1017/S0140525X16001837

Langdon, C., Genkin, M., & Engel, T. A. (2023). A unifying perspective on neural manifolds and circuits for cognition. *Nature Reviews Neuroscience*. https://doi.org/10.1038/s41583-023-00693-x

Laughlin, S. B. (2001). Energy as a constraint on the coding and processing of sensory information. *Current Opinion in Neurobiology*, *11*(4), 475–480.

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*(2), 365–411. https://doi.org/10.1016/S0047-259X(03)00096-4

Leggio, M. G., Mandolesi, L., Federico, F., Spirito, F., Ricci, B., Gelfo, F., & Petrosini, L. (2005). Environmental enrichment promotes improved spatial abilities and enhanced dendritic growth in the rat. *Behavioural Brain Research*, *163*(1), 78–90.

Logie, R., Camos, V., & Cowan, N. (2021). *Working Memory: State of the Science*. Oxford University Press. https://books.google.co.uk/books?id=l8KRzQEACAAJ

Losos, J. B. (2017). *The Princeton guide to evolution*. Princeton University Press.

Maass, W., Natschläger, T., & Markram, H. (2002). Real-Time Computing Without Stable States: A New Framework for Neural Computation Based on Perturbations. In *Neural Computation* (Vol. 14).

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*(7474), 78–84. https://doi.org/10.1038/nature12742

Marder, E., & Bucher, D. (2001). Central pattern generators and the control of rhythmic movements. *Current Biology*, *11*(23), R986–R996.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Marra, N. J., Stanhope, M. J., Jue, N. K., Wang, M., Sun, Q., Pavinski Bitar, P., Richards, V. P., Komissarov, A., Rayko, M., Kliver, S., Stanhope, B. J., Winkler, C., O'brien, S. J., Antunes, A., Jorgensen, S., Shivji, M. S., Bowen, B. W., Douady, C. J., Koop, B. F., & Lyon, B. (2018). *White shark genome reveals ancient elasmobranch adaptations associated with wound healing and the maintenance of genome stability*. https://doi.org/10.5061/dryad.9r2p3ks

Mastrogiuseppe, F., & Ostojic, S. (2018). Linking Connectivity, Dynamics, and Computations in Low-Rank Recurrent Neural Networks. *Neuron*, *99*(3), 609-623.e29. https://doi.org/10.1016/j.neuron.2018.07.003

Mayr, U., & Kliegl, R. (2003). Differential Effects of Cue Changes and Task Changes on Task-Set Selection Costs. *Journal of Experimental Psychology: Learning Memory and Cognition*, *29*(3), 362–372. https://doi.org/10.1037/0278-7393.29.3.362

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv Preprint ArXiv:1802.03426*.

Meyer, T., Qi, X. L., Stanford, T. R., & Constantinidis, C. (2011). Stimulus selectivity in dorsal and ventral prefrontal cortex after training in working memory tasks. *Journal of Neuroscience*, *31*(17), 6266–6276. https://doi.org/10.1523/JNEUROSCI.6798-10.2011

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*(1), 167–202.
Miller, E. K., Lundqvist, M., & Bastos, A. M. (2018). Working Memory 2.0. In *Neuron* (Vol. 100, Issue 2, pp. 463–475). Cell Press. https://doi.org/10.1016/j.neuron.2018.09.023
Moll, H., & Tomasello, M. (2007). Cooperation and human cognition: the Vygotskian intelligence hypothesis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 639–648.
Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic theory of working memory. *Science*, *319*(5869), 1543–1546.
Musslick, S., & Cohen, J. D. (2021). Rationalizing constraints on the capacity for cognitive control. In *Trends in Cognitive Sciences* (Vol. 25, Issue 9, pp. 757–775). Elsevier Ltd. https://doi.org/10.1016/j.tics.2021.06.001
Musslick, S., Saxe, A., Novick, A., Reichman, D., & Cohen, J. D. (n.d.). *Running head: RATIONAL BOUNDEDNESS OF COGNITIVE CONTROL 1 On the Rational Boundedness of Cognitive Control: Shared Versus Separated Representations*.
Nielsen, J., Hedeholm, R. B., Heinemeier, J., Bushnell, P. G., Christiansen, J. S., Olsen, J., Ramsey, C. B., Brill, R. W., Simon, M., & Steffensen, K. F. (2016). Eye lens radiocarbon reveals centuries of longevity in the Greenland shark (Somniosus microcephalus). *Science*, *353*(6300), 702–704.
Nithianantharajah, J., & Hannan, A. J. (2006). Enriched environments, experience-dependent plasticity and disorders of the nervous system. In *Nature Reviews Neuroscience* (Vol. 7, Issue 9, pp. 697–709). https://doi.org/10.1038/nrn1970
Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154.
Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013). The Curse of Planning: Dissecting Multiple Reinforcement-Learning Systems by Taxing the Central Executive. *Psychological Science*, *24*(5), 751–761. https://doi.org/10.1177/0956797612463080
Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359.
Park, I. M., Ságodi, Á., & Sokół, P. A. (2023). *Persistent learning signals and working memory without continuous attractors*. http://arxiv.org/abs/2308.12585
Pemberton, J., Chadderton, P., & Costa, R. P. (2022). Cerebellar-driven cortical dynamics enable task acquisition, switching and consolidation. *BioRxiv*, 2011–2022.
Piwek, E. P., Stokes, M. G., & Summerfield, C. P. (2022). A recurrent neural network model of prefrontal brain activity during a working memory task. *BioRxiv*, 2022–2029.
Pospisil, D. A., Pasupathy, A., & Bair, W. (2018). 'Artiphysiology'reveals V4-like shape tuning in a deep network trained for image classification. *Elife*, *7*, e38242.
Qi, X. L., Meyer, T., Stanford, T. R., & Constantinidis, C. (2011). Changes in prefrontal neuronal activity after learning to perform a spatial working

memory task. *Cerebral Cortex*, *21*(12), 2722–2732. https://doi.org/10.1093/cercor/bhr058
Raposo, D., Kaufman, M. T., & Churchland, A. K. (2014). A category-free neural population supports evolving demands during decision-making. *Nature Neuroscience*, *17*(12), 1784–1792. https://doi.org/10.1038/nn.3865
Reader, S. M., & Laland, K. N. (2002). Social intelligence, innovation, and enhanced brain size in primates. *Proceedings of the National Academy of Sciences*, *99*(7), 4436–4441.
Reinert, S., Hübener, M., Bonhoeffer, T., & Goltstein, P. M. (2021). Mouse prefrontal cortex represents learned rules for categorization. *Nature*, *593*(7859), 411–417. https://doi.org/10.1038/s41586-021-03452-z
Reverberi, C., Görgen, K., & Haynes, J. D. (2012). Compositionality of rule representations in human prefrontal cortex. *Cerebral Cortex*, *22*(6), 1237–1246. https://doi.org/10.1093/cercor/bhr200
Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., … Kording, K. P. (2019). A deep learning framework for neuroscience. In *Nature Neuroscience* (Vol. 22, Issue 11, pp. 1761–1770). Nature Publishing Group. https://doi.org/10.1038/s41593-019-0520-2
Rigotti, M., Barak, O., Warden, M. R., Wang, X. J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, *497*(7451), 585–590. https://doi.org/10.1038/nature12160
Rios, A., & Itti, L. (2020). Lifelong Learning without a Task Oracle. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, *2020-November*, 255–263. https://doi.org/10.1109/ICTAI50040.2020.00049
Roberts, P., & Stewart, B. A. (2018). Defining the 'generalist specialist' niche for Pleistocene Homo sapiens. In *Nature Human Behaviour* (Vol. 2, Issue 8, pp. 542–550). Nature Publishing Group. https://doi.org/10.1038/s41562-018-0394-4
Roelfsema, P. R., & Ooyen, A. van. (2005). Attention-gated reinforcement learning of internal representations for classification. *Neural Computation*, *17*(10), 2176–2214.
Rose, N. S., Larocque, J. J., Riggall, A. C., Gosseries, O., Starrett, M. J., Meyering, E. E., & Postle, B. R. (n.d.). *Reactivation of latent working memories with transcranial magnetic stimulation*. https://www.science.org
Rourke, O. L. C., & Butts, D. A. (2017). Cortical computations via transient attractors. *PLoS ONE*, *12*(12). https://doi.org/10.1371/journal.pone.0188562
Roy, J. E., Riesenhuber, M., Poggio, T., & Miller, E. K. (2010). Prefrontal Cortex Activity during Flexible Categorization. *Journal of Neuroscience*, *30*(25), 8519–8528. https://doi.org/10.1523/JNEUROSCI.4837-09.2010
Ruesseler, M., Weber, L. A., Marshall, T. R., O, J. X., & Hunt, L. T. (n.d.). *Decision-making in dynamic, continuously evolving environments: quantifying the flexibility of human choice*. https://doi.org/10.1101/2022.08.18.504278
Ruff, C. C., Bestmann, S., Blankenburg, F., Bjoertomt, O., Josephs, O., Weiskopf, N., Deichmann, R., & Driver, J. (2008). Distinct causal influences

of parietal versus frontal areas on human visual cortex: Evidence from concurrent TMS-fMRI. *Cerebral Cortex*, *18*(4), 817–827. https://doi.org/10.1093/cercor/bhm128

Sadtler, P. T., Quick, K. M., Golub, M. D., Chase, S. M., Ryu, S. I., Tyler-Kabara, E. C., Yu, B. M., & Batista, A. P. (2014). Neural constraints on learning. *Nature*, *512*(7515), 423–426. https://doi.org/10.1038/nature13665

Samborska, V., Butler, J. L., Walton, M. E., Behrens, T. E. J., & Akam, T. (2022). Complementary task representations in hippocampus and prefrontal cortex for generalizing the structure of problems. *Nature Neuroscience*, *25*(10), 1314–1326.

Saxe, A., Nelli, S., & Summerfield, C. (2021). If deep learning is the answer, what is the question? In *Nature Reviews Neuroscience* (Vol. 22, Issue 1, pp. 55–67). Nature Research. https://doi.org/10.1038/s41583-020-00395-8

Semedo, J. D., Zandvakili, A., Machens, C. K., Byron, M. Y., & Kohn, A. (2019). Cortical areas interact through a communication subspace. *Neuron*, *102*(1), 249–259.

Seung, H. S. (1996). How the brain keeps the eyes still. *Proceedings of the National Academy of Sciences*, *93*(23), 13339–13344.

Shallice, T., & Evans, M. E. (1978). The involvement of the frontal lobes in cognitive estimation. *Cortex*, *14*(2), 294–303.

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

Sherman, S. M. (2016). Thalamus plays a central role in ongoing cortical functioning. *Nature Neuroscience*, *19*(4), 533–541.

Smittenaar, P., FitzGerald, T. H. B., Romei, V., Wright, N. D., & Dolan, R. J. (2013). Disruption of Dorsolateral Prefrontal Cortex Decreases Model-Based in Favor of Model-free Control in Humans. *Neuron*, *80*(4), 914–919. https://doi.org/10.1016/j.neuron.2013.08.009

Song, H. F., Yang, G. R., & Wang, X.-J. (2017). Reward-based training of recurrent neural networks for cognitive and value-based tasks. *Elife*, *6*, e21492.

Song, S., Sjöström, P. J., Reigl, M., Nelson, S., & Chklovskii, D. B. (2005). Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS Biology*, *3*(3), 0507–0519. https://doi.org/10.1371/journal.pbio.0030068

Souza, A. S., & Oberauer, K. (2015). Time-based forgetting in visual working memory reflects temporal distinctiveness, not decay. *Psychonomic Bulletin & Review*, *22*, 156–162.

Spaak, E., Watanabe, K., Funahashi, S., & Stokes, M. G. (2017). Stable and dynamic coding for working memory in primate prefrontal cortex. *Journal of Neuroscience*, *37*(27), 6503–6516. https://doi.org/10.1523/JNEUROSCI.3364-16.2017

Spearman, C. (1923). *The nature of" intelligence" and the principles of cognition*. Macmillan.

Steinmetz, N. A., Aydin, C., Lebedeva, A., Okun, M., Pachitariu, M., Bauza, M., Beau, M., Bhagat, J., Böhm, C., Broux, M., Chen, S., Colonell, J., Gardner, R. J., Karsh, B., Kloosterman, F., Kostadinov, D., Mora-Lopez, C., O'Callaghan, J., Park, J., … Harris, T. D. (2021). Neuropixels 2.0: A

miniaturized high-density probe for stable, long-term brain recordings. *Science*, *372*(6539). https://doi.org/10.1126/science.abf4588
Stenhouse, D. (1974). *The Evolution of Intelligence: A General Theory and Some of Its Implications*. Barnes & Noble Books. https://books.google.co.uk/books?id=6O9SvwEACAAJ
Sternberg, R. J. (2022). The search for the elusive basic processes underlying human intelligence: Historical and contemporary perspectives. *Journal of Intelligence*, *10*(2), 28.
Sternberg, R. J., & Kaufman, S. B. (2011). *The Cambridge handbook of intelligence*. Cambridge University Press.
Stokes, M. G. (2015). "Activity-silent" working memory in prefrontal cortex: A dynamic coding framework. In *Trends in Cognitive Sciences* (Vol. 19, Issue 7, pp. 394–405). Elsevier Ltd. https://doi.org/10.1016/j.tics.2015.05.004
Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., & Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. *Neuron*, *78*(2), 364–375. https://doi.org/10.1016/j.neuron.2013.01.039
Stroud, J. P., Watanabe, K., Suzuki, T., Stokes, M. G., & Lengyel, M. (n.d.). *Optimal information loading into working memory in prefrontal cortex explains dynamic coding*. https://doi.org/10.1101/2021.11.16.468360
Stroud, J. P., Wójcik, M., Jensen, K. T., Kusunoki, M., Kadohisa, M., Duncan, J., Stokes, M. G., & Lengyel, M. (n.d.). *Ignorance is bliss: effects of noise and metabolic cost on cortical task representations*. https://doi.org/10.1101/2023.07.11.548492
Summerfield, C. (2022a). *Natural General Intelligence: How understanding the brain can help us build AI*. Oxford University Press.
Summerfield, C. (2022b). *Natural General Intelligence: How understanding the brain can help us build AI*. Oxford University Press.
Sun, W., Advani, M., Spruston, N., Saxe, A., & Fitzgerald, J. E. (2023). Organizing memories for generalization in complementary learning systems. *Nature Neuroscience*, *26*(8), 1438–1448. https://doi.org/10.1038/s41593-023-01382-9
Sussillo, D., Churchland, M. M., Kaufman, M. T., & Shenoy, K. V. (2015). A neural network that finds a naturalistic solution for the production of muscle activity. *Nature Neuroscience*, *18*(7), 1025–1033. https://doi.org/10.1038/nn.4042
Thurstone, L. L. (1999). *The Nature of Intelligence*. Routledge. https://books.google.co.uk/books?id=7naPxER92YwC
Tkalčič, M., & Tasič, J. F. (n.d.). *Colour spaces-perceptual, historical and applicational background*. http://www.mathworks.com
Tolias, A. S., Ecker, A. S., Siapas, A. G., Hoenselaar, A., Keliris, G. A., & Logothetis, N. K. (2007). Recording chronically from the same neurons in awake, behaving primates. *Journal of Neurophysiology*, *98*(6), 3780–3790. https://doi.org/10.1152/jn.00260.2007
van Schaik, C. P., Isler, K., & Burkart, J. M. (2012). Explaining brain size variation: From social to cultural brain. In *Trends in Cognitive Sciences* (Vol. 16, Issue 5, pp. 277–284). https://doi.org/10.1016/j.tics.2012.04.004

Vandierendonck, A., Liefooghe, B., & Verbruggen, F. (2010). Task Switching: Interplay of Reconfiguration and Interference Control. *Psychological Bulletin*, *136*(4), 601–626. https://doi.org/10.1037/a0019791

Varki, A., Geschwind, D. H., & Eichler, E. E. (2008). Human uniqueness: Genome interactions with environment, behaviour and culture. In *Nature Reviews Genetics* (Vol. 9, Issue 10, pp. 749–763). https://doi.org/10.1038/nrg2428

Vygotsky, L. S. (2012). *Thought and language*. MIT press.

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., Hassabis, D., & Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, *21*(6), 860–868. https://doi.org/10.1038/s41593-018-0147-8

Wang, X.-J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, *36*(5), 955–968.

Watanabe, E., Kitaoka, A., Sakamoto, K., Yasugi, M., & Tanaka, K. (2018). Illusory motion reproduced by deep neural networks trained for prediction. *Frontiers in Psychology*, 345.

Watanabe, M. (1986). Prefrontal Unit Activity During Delayed Conditional Go/No-Go Discrimination in the Monkey. II. Relation to Go and No-Go Responses. In *Brain Research* (Vol. 382).

Whittington, J. C. R., & Bogacz, R. (2017). An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural Computation*, *29*(5), 1229–1262.

Whittington, J. C. R., Dorrell, W., Ganguli, S., & Behrens, T. E. J. (2022). *Disentanglement with Biological Constraints: A Theory of Functional Cell Types*. http://arxiv.org/abs/2210.01768

Wójcik, M. J., Stroud, J. P., Wasmuht, D., Kusunoki, M., Kadohisa, M., Myers, N. E., Hunt, L. T., Duncan, J., & Stokes, M. G. (n.d.). *Learning shapes neural geometry in the prefrontal cortex*. https://doi.org/10.1101/2023.04.24.538054

Wolff, M. J., Jochim, J., Akyürek, E. G., & Stokes, M. G. (2017a). Dynamic hidden states underlying working-memory-guided behavior. *Nature Neuroscience*, *20*(6), 864–871.

Wolff, M. J., Jochim, J., Akyürek, E. G., & Stokes, M. G. (2017b). Dynamic hidden states underlying working-memory-guided behavior. *Nature Neuroscience*, *20*(6), 864–871. https://doi.org/10.1038/nn.4546

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82.

Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., & Wang, X. J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, *22*(2), 297–306. https://doi.org/10.1038/s41593-018-0310-2

Yoon, Y. B., Shin, W.-G., Lee, T. Y., Hur, J.-W., Cho, K. I. K., Sohn, W. S., Kim, S.-G., Lee, K.-H., & Kwon, J. S. (2017). Brain structural networks associated with intelligence and visuomotor ability. *Scientific Reports*, *7*(1), 2177.

Zador, A., Escola, S., Richards, B., Ölveczky, B., Bengio, Y., Boahen, K., Botvinick, M., Chklovskii, D., Churchland, A., Clopath, C., DiCarlo, J., Ganguli, S., Hawkins, J., Körding, K., Koulakov, A., LeCun, Y., Lillicrap, T., Marblestone, A., Olshausen, B., … Tsao, D. (2023). Catalyzing next-

generation Artificial Intelligence through NeuroAI. *Nature Communications*, *14*(1), 1597. https://doi.org/10.1038/s41467-023-37180-x

Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. In *Nature Communications* (Vol. 10, Issue 1). Nature Publishing Group. https://doi.org/10.1038/s41467-019-11786-6

Zanto, T. P., Rubens, M. T., Bollinger, J., & Gazzaley, A. (2010). Top-down modulation of visual feature processing: The role of the inferior frontal junction. *NeuroImage*, *53*(2), 736–745. https://doi.org/10.1016/j.neuroimage.2010.06.012

Zanto, T. P., Rubens, M. T., Thangavel, A., & Gazzaley, A. (2011). Causal role of the prefrontal cortex in top-down modulation of visual processing and working memory. *Nature Neuroscience*, *14*(5), 656–663. https://doi.org/10.1038/nn.2773

Zeller, E., Timmermann, A., Yun, K.-S., Raia, P., Stein, K., & Ruan, J. (2023). Human adaptation to diverse biomes over the past 3 million years. *Science*, *380*(6645), 604–608.

Zhao, S., Tang, X., Tian, W., Partarrieu, S., Liu, R., Shen, H., Lee, J., Guo, S., Lin, Z., & Liu, J. (2023). Tracking neural activity from the same cells during the entire adult life of mice. *Nature Neuroscience*. https://doi.org/10.1038/s41593-023-01267-x

Supplementary information

Section 1: How to set selectivity parameters for optimal XOR decoding

We assume that neural activities $\boldsymbol{x}$ of $\boldsymbol{N}$ neurons are given by the following regression model

$$\boldsymbol{x} = \boldsymbol{\beta}_0 + \delta_{c,c_1}\boldsymbol{\beta}_C + \delta_{s,s_1}\boldsymbol{\beta}_S + \delta_{c,c_1}\delta_{s,s_1}\boldsymbol{\beta}_I + \boldsymbol{\eta} \tag{1}$$

Where $\boldsymbol{\beta}_C, \boldsymbol{\beta}_S, \boldsymbol{\beta}_I$ are the regression 'coefficients' for colour, shape, and the interaction term, respectively, $\delta_{c,c_1} = 1$ if the colour c is colour 1 and -1 otherwise (same for δ_{s,s_1} for shapes s), and $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_1)$. Therefore,

$$\boldsymbol{x}|r_1 \sim \mathcal{N}(\boldsymbol{\mu}_{r_1}, \boldsymbol{\Sigma}_1) \tag{2}$$

$$\boldsymbol{x}|r_2 \sim \mathcal{N}(\boldsymbol{\mu}_{r_2}, \boldsymbol{\Sigma}_2). \tag{3}$$

Where r_1 is one XOR condition (i.e., $c = 1, s = 1$ or $c = 2, s = 2$) and r_2 is the other XOR condition (i.e., $c = 1, s = 2$ or $c = 2, s = 1$) and $\boldsymbol{\Sigma}_2$ is some noise covariance matrix. Note that with the inclusion of the interaction term, it is sufficient to separate the two XOR conditions. We now calculate $\boldsymbol{\mu}_{r_1}$ and $\boldsymbol{\mu}_{r_2}$:

$$\boldsymbol{\mu}_{r_2} = \frac{\boldsymbol{\beta}_0 + \boldsymbol{\beta}_C + \boldsymbol{\beta}_S + \boldsymbol{\beta}_I}{2} + \frac{\boldsymbol{\beta}_0 - \boldsymbol{\beta}_C - \boldsymbol{\beta}_S + \boldsymbol{\beta}_I}{2}$$

$$= \boldsymbol{\beta}_0 + \boldsymbol{\beta}_I \tag{4}$$

and,

$$\boldsymbol{\mu}_{r_2} = \frac{\boldsymbol{\beta}_0 - \boldsymbol{\beta}_C + \boldsymbol{\beta}_S - \boldsymbol{\beta}_I}{2} + \frac{\boldsymbol{\beta}_0 + \boldsymbol{\beta}_C - \boldsymbol{\beta}_S - \boldsymbol{\beta}_I}{2}$$

$$= \boldsymbol{\beta}_0 - \boldsymbol{\beta}_I \tag{5}$$

Therefore,

$$\boldsymbol{x}|r_1 \sim \mathcal{N}(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_I, \boldsymbol{\Sigma}_2) \tag{6}$$

$$\boldsymbol{x}|r_2 \sim \mathcal{N}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_I, \boldsymbol{\Sigma}_2). \tag{7}$$

Therefore, we need $\boldsymbol{\beta}_I > 0$ to be able to separate the two XOR conditions.

Section 2: An energy cost on unnecessary neural activity

If we also consider minimising the squared norm of neural activity for each condition (i.e., an energy cost), we have

$$\mathbb{E}\left[\left\|\boldsymbol{x}_{c,s}\right\|^2\right] = \mathbb{E}\left[\boldsymbol{x}_{c,s}\right]^{\mathrm{T}}\,\mathbb{E}\left[\boldsymbol{x}_{c,s}\right] + \mathrm{Tr}(\boldsymbol{\Sigma}_1)$$

(8)

where the subscripts c and s correspond to colour and shape indices, respectively. Therefore,

$$\mathbb{E}\left[\left\|\boldsymbol{x}_{1,2}\right\|^2\right] = (\boldsymbol{\beta}_0 - \boldsymbol{\beta}_C + \boldsymbol{\beta}_S - \boldsymbol{\beta}_I)^{\mathrm{T}}\,(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_C + \boldsymbol{\beta}_S - \boldsymbol{\beta}_I) + \mathrm{Tr}(\boldsymbol{\Sigma}_1)$$

(9)

$$\mathbb{E}\left[\left\|\boldsymbol{x}_{2,1}\right\|^2\right] = (\boldsymbol{\beta}_0 + \boldsymbol{\beta}_C - \boldsymbol{\beta}_S - \boldsymbol{\beta}_I)^{\mathrm{T}}\,(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_C - \boldsymbol{\beta}_S - \boldsymbol{\beta}_I) + \mathrm{Tr}(\boldsymbol{\Sigma}_1)$$

(10)

$$\mathbb{E}\left[\left\|\boldsymbol{x}_{1,1}\right\|^2\right] = (\boldsymbol{\beta}_0 - \boldsymbol{\beta}_C - \boldsymbol{\beta}_S + \boldsymbol{\beta}_I)^{\mathrm{T}}\,(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_C - \boldsymbol{\beta}_S + \boldsymbol{\beta}_I) + \mathrm{Tr}(\boldsymbol{\Sigma}_1)$$

(11)

$$\mathbb{E}\left[\left\|\boldsymbol{x}_{2,2}\right\|^2\right] = (\boldsymbol{\beta}_0 + \boldsymbol{\beta}_C + \boldsymbol{\beta}_S + \boldsymbol{\beta}_I)^{\mathrm{T}}\,(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_C + \boldsymbol{\beta}_S + \boldsymbol{\beta}_I) + \mathrm{Tr}(\boldsymbol{\Sigma}_1)$$

(12)

Therefore, the total mean energy cost m is given by

$$m = \frac{1}{4}\sum_{c=1,2}\sum_{s=1,2}\mathbb{E}\left[\left\|\boldsymbol{x}_{c,s}\right\|^2\right]$$

(13)

$$= \boldsymbol{\beta}_0{}^{\mathrm{T}}\boldsymbol{\beta}_0 + \boldsymbol{\beta}_C{}^{\mathrm{T}}\boldsymbol{\beta}_C + \boldsymbol{\beta}_S{}^{\mathrm{T}}\boldsymbol{\beta}_S + \boldsymbol{\beta}_I{}^{\mathrm{T}}\boldsymbol{\beta}_I + \mathrm{Tr}(\boldsymbol{\Sigma}_1)$$

(14)

$$= \|\boldsymbol{\beta}_0\|^2 + \|\boldsymbol{\beta}_C\|^2 + \|\boldsymbol{\beta}_S\|^2 + \|\boldsymbol{\beta}_I\|^2 + \mathrm{Tr}(\boldsymbol{\Sigma}_1)$$

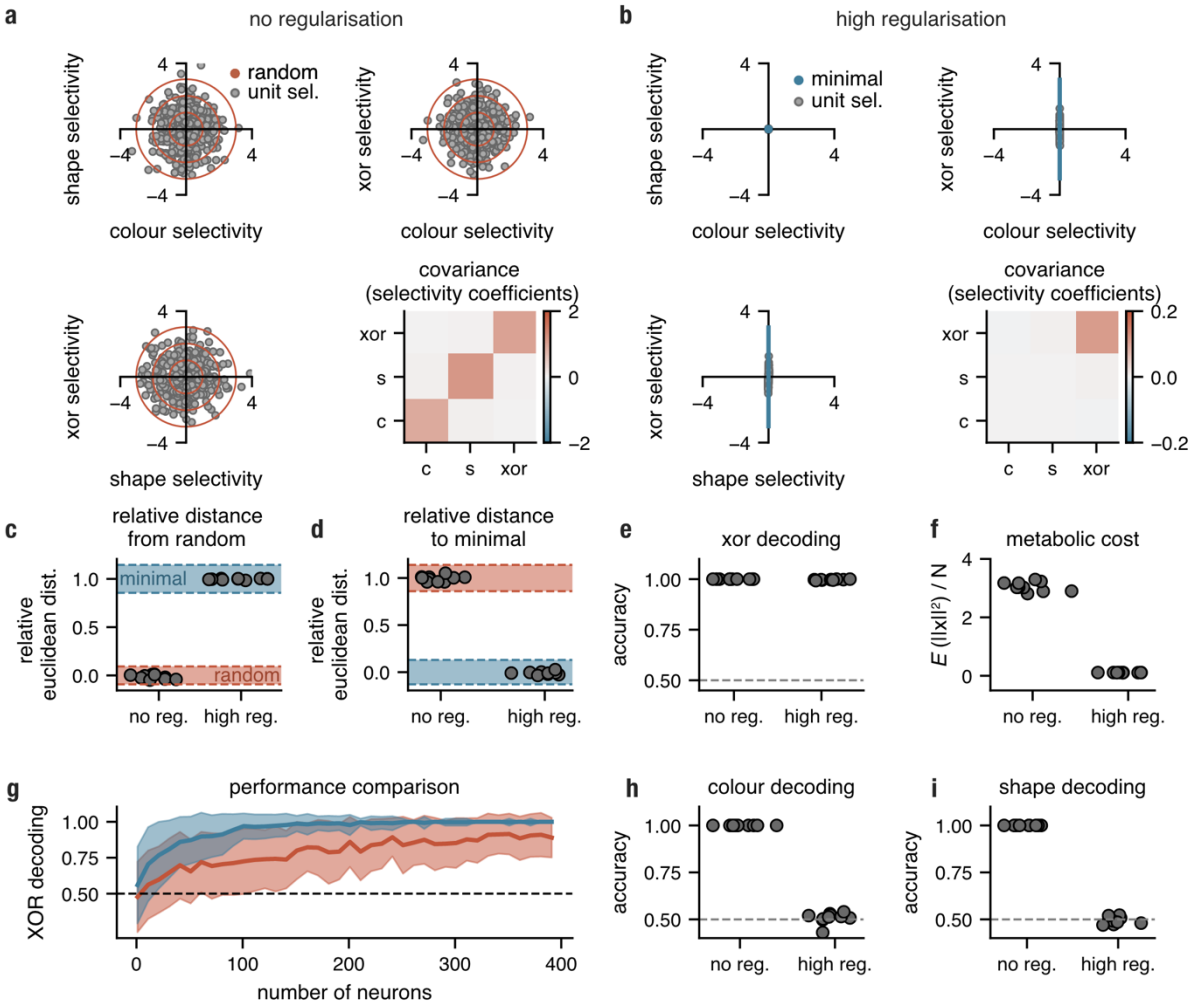
(15)

To minimise m while keeping $\boldsymbol{\beta}_I > 0$, which we need for performance, we can set $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_C = \boldsymbol{\beta}_S = \mathbf{0}$ which gives

$$m = \|\boldsymbol{\beta}_I\|^2 + \mathrm{Tr}(\boldsymbol{\Sigma}_1)$$

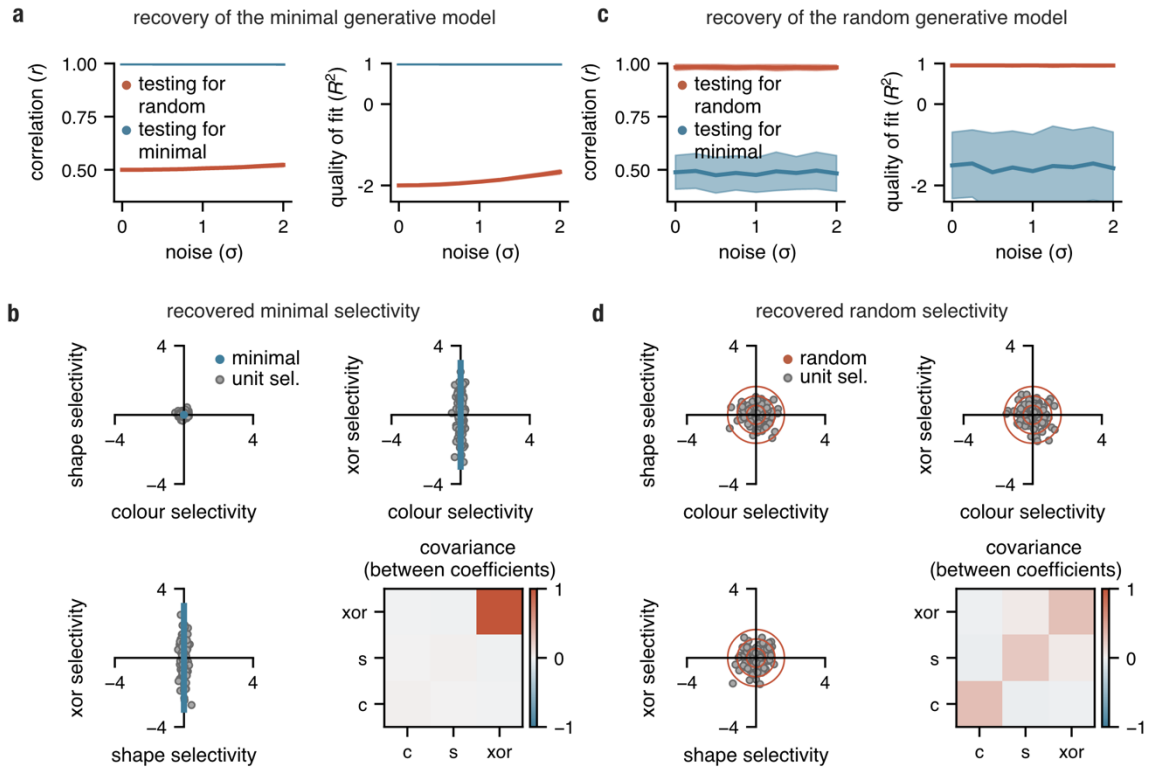
(16)

Section 3: Supplementary figures

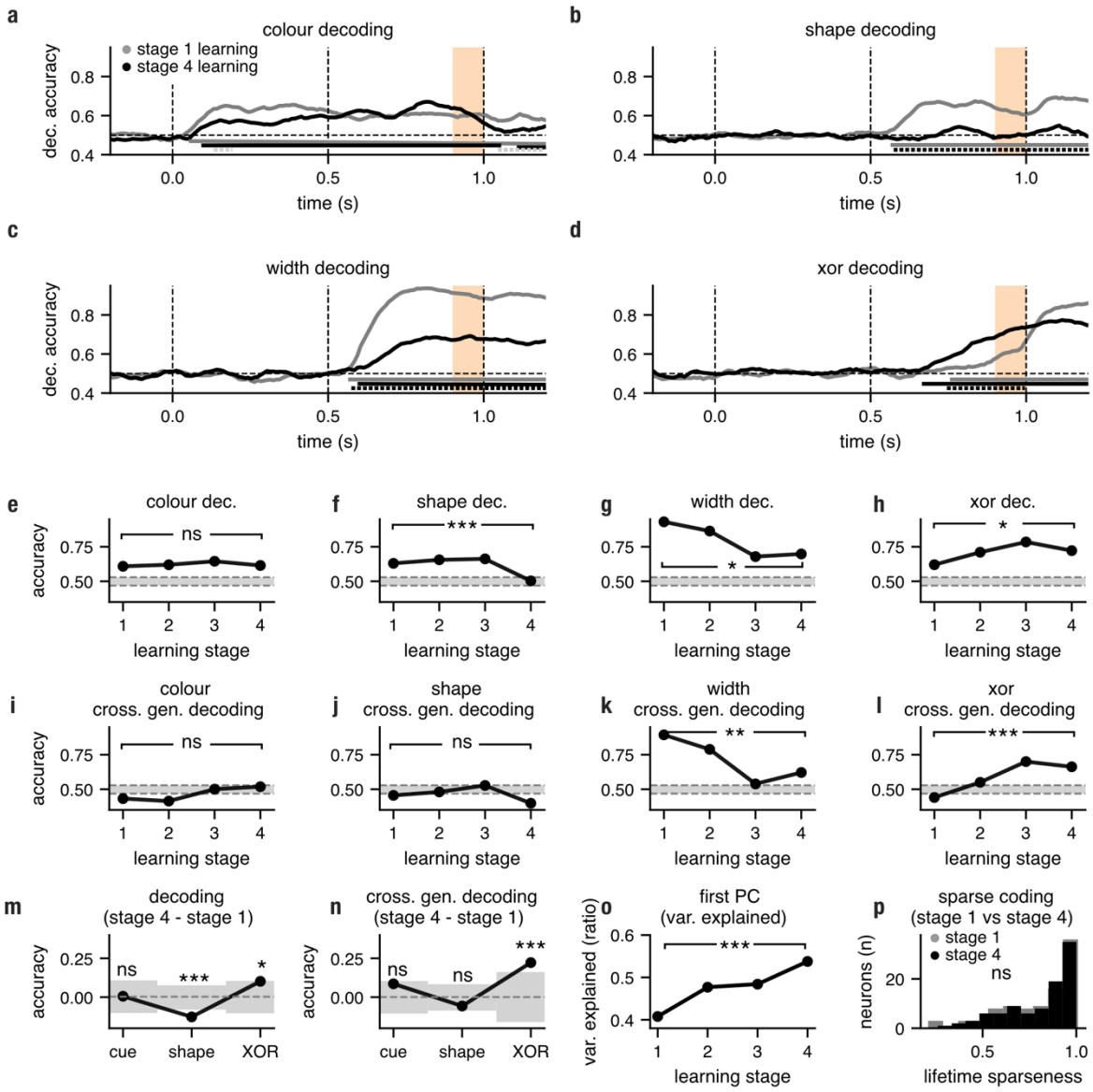


Supplementary figure 1. Optimised feedforward networks converge to the minimal XOR selectivity model. Twenty feedforward networks were trained (10 with high levels of regularisation and 10 with no regularisation) to perform the XOR task. **a, b,** Selectivity observed in no and high regularisation networks, respectively; models of random (red ellipses) and minimal selectivity (blue ellipses) well approximated the observed selectivity. **c, d**, After training, low regularisation networks converged on a random mixed selectivity regime and high regularisation networks on a minimal XOR regime. **e**, XOR decoding (linear SVM) for both no and high regularisation models. **f**, No regularisation models exhibited substantially lower metabolic cost (cf. **Supp. materials** eq. 8). **g**,

Comparison of XOR decoding obtained from minimal and random generative models as a function of population size. **h**, **i**, Colour and shape decoding (linear SVM) for no and high regularisation models, respectively.

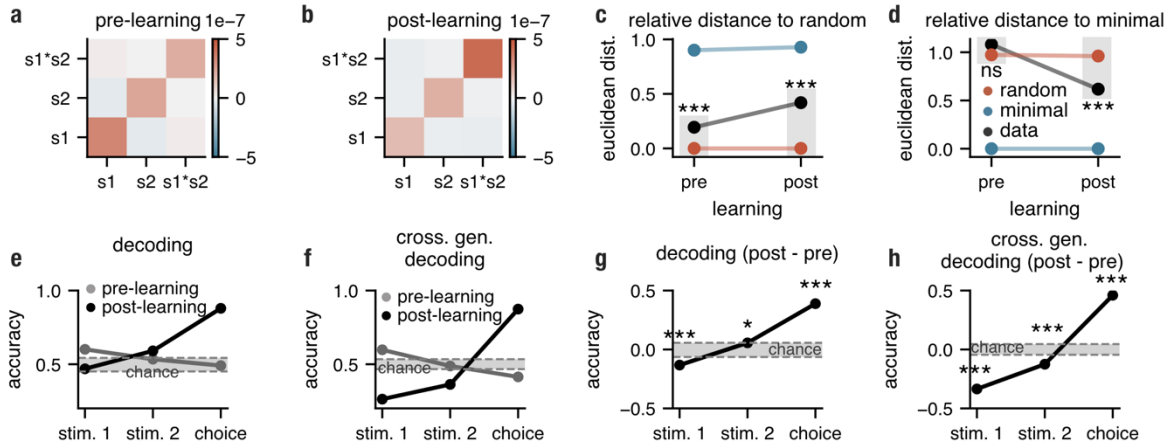


Supplementary figure 2. Linear regression recovers underlying generative models. a, Pearson correlation and R^2 computed between the covariance matrix of recovered selectivity and true underlying selectivity (when minimal generative model was used) for different levels of noise. **b**, Selectivity coefficients obtained after running a linear regression plotted for each neuron in selectivity space (for $\sigma = 2$); minimal model overlaid in blue; covariance matrix computed between the recovered selectivity coefficients. **c**,**d**, Analogous to a,b but when the random model was used to generate data; random model overlaid in red. Shaded areas in a and c indicate the mean $\pm 1\ s.d.$ computed over 100 different initialisations.

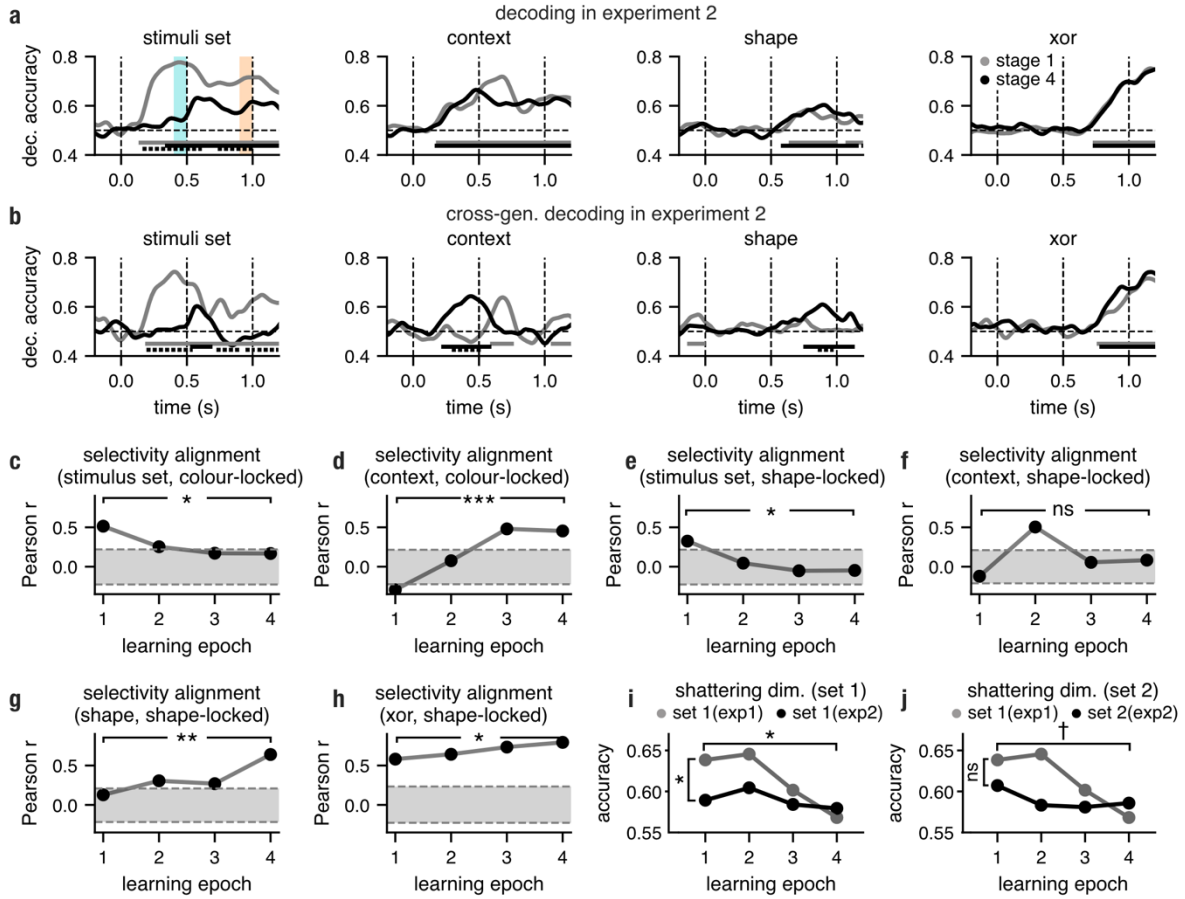


Supplementary figure 3. Decoding of task variables as a function of learning in experiment 1. **a-d**, Time resolved decoding in stage 1 and stage 4 of colour, shape, width and XOR, respectively; horizontal bars indicate statistical significance; dashed black bars indicate significant differences in coding between stage 1 and stage 4; dashed grey bars indicate trend-level significance in stage 1 vs stage 4 decoding. The pale orange area indicates the time window for which all subsequent decoding analyses were run. Vertical three dashed lines show the onset of the colour, shape and the outcome, respectively. **e-h**, The learning dynamics of the decoding of the main four variables. **i-l**, The learning dynamics of cross-generalised decoding of the main four variables. **m, n**, Statistical test for the difference in decoding and cross-generalised decoding, respectively, between stage 1 and stage 4 for three main task variables. **o**, variance explained

(ratio) by the first principal component plotted as a function of learning (see *Methods, principal component analysis* for details). **p**, Distributions of the lifetime sparseness index observed in stage 1 (grey) and stage 4 (black). A Kolmogorov–Smirnov test revealed no differences between these distributions, $D = .06$, $p = .99$. Shaded areas in panels **e-n** indicate chance levels**.** All p-values were calculated from permutation tests (***, $p < 0.01$; **, $p < 0.01$; *, $p < 0.05$; †, $p < 0.01$; $n.s.$, not significant).

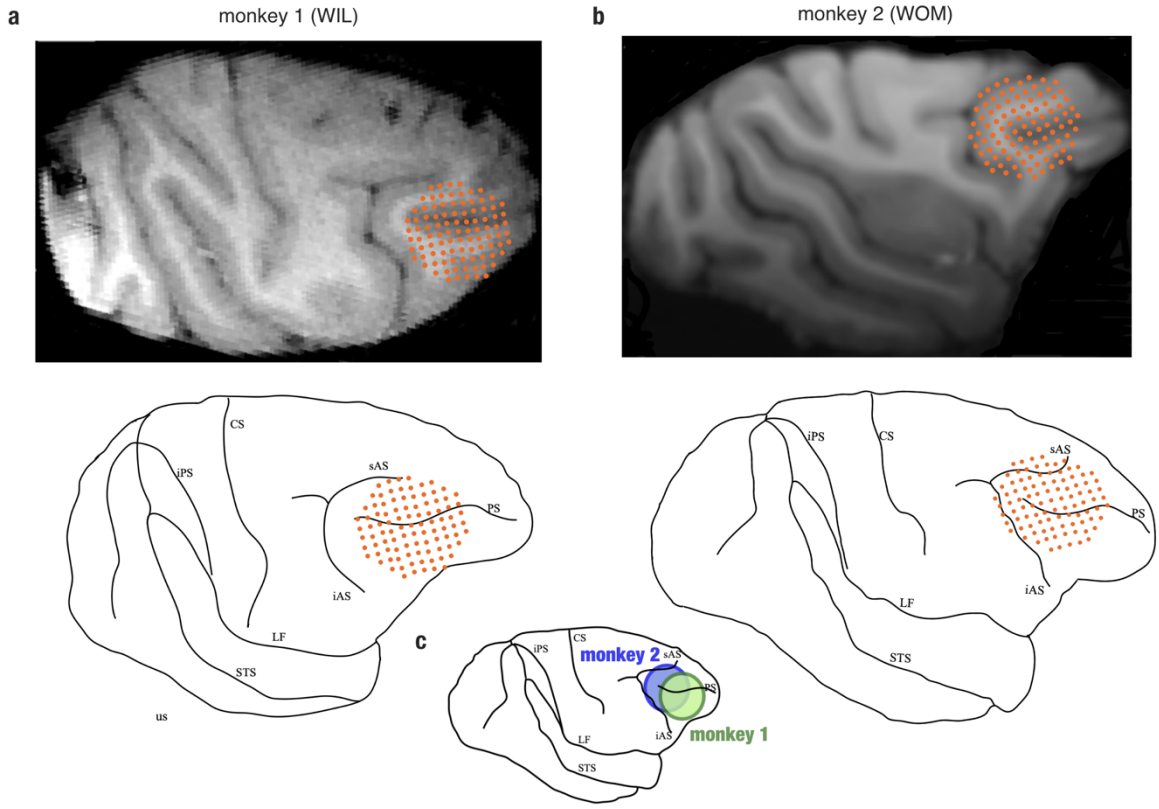


Supplementary figure 4. The re-analysis of Constantinidis et al.(Constantinidis et al., 2016; Meyer et al., 2011; Qi et al., 2011) **dataset.** We employed the same analysis methods as in **Fig 2,** and **Fig 3** to test whether the PFC activity reported in Constantinidis et al. converged on a minimal XOR model. **a**, the covariance matrix describing relations between the selectivity for stimulus 1, stimulus 2 and their interaction (XOR) in the pre-learning phase of the experiment. **b**, Same as panel a but post learning. **c**, Relative Euclidean distance between the covariance matrix of selectivity coefficients and the covariance matrix expected from random selectivity (with matched total variance) plotted as a function of learning (*Methods*, *measuring similarity between selectivity distributions*). **d**, Same as panel **c** but we show the relative distance from the covariance matrix expected from minimal selectivity (with matched total variance). **e**, Decoding of task variables for pre- and post-learning stages. **f**, Cross-generalised decoding of task variables plotted as a function of learning. **g, h**, Learning-induced accuracy differences in decoding and cross-generalised decoding, respectively. Shaded areas in **e-h** illustrate chance-level decoding obtained by shuffling trial labels (for details see *Methods, statistical testing*). All p-values were calculated from permutation tests (***, $p < 0.01$; **, $p < 0.01$; *, $p < 0.05$; n.s., not significant).

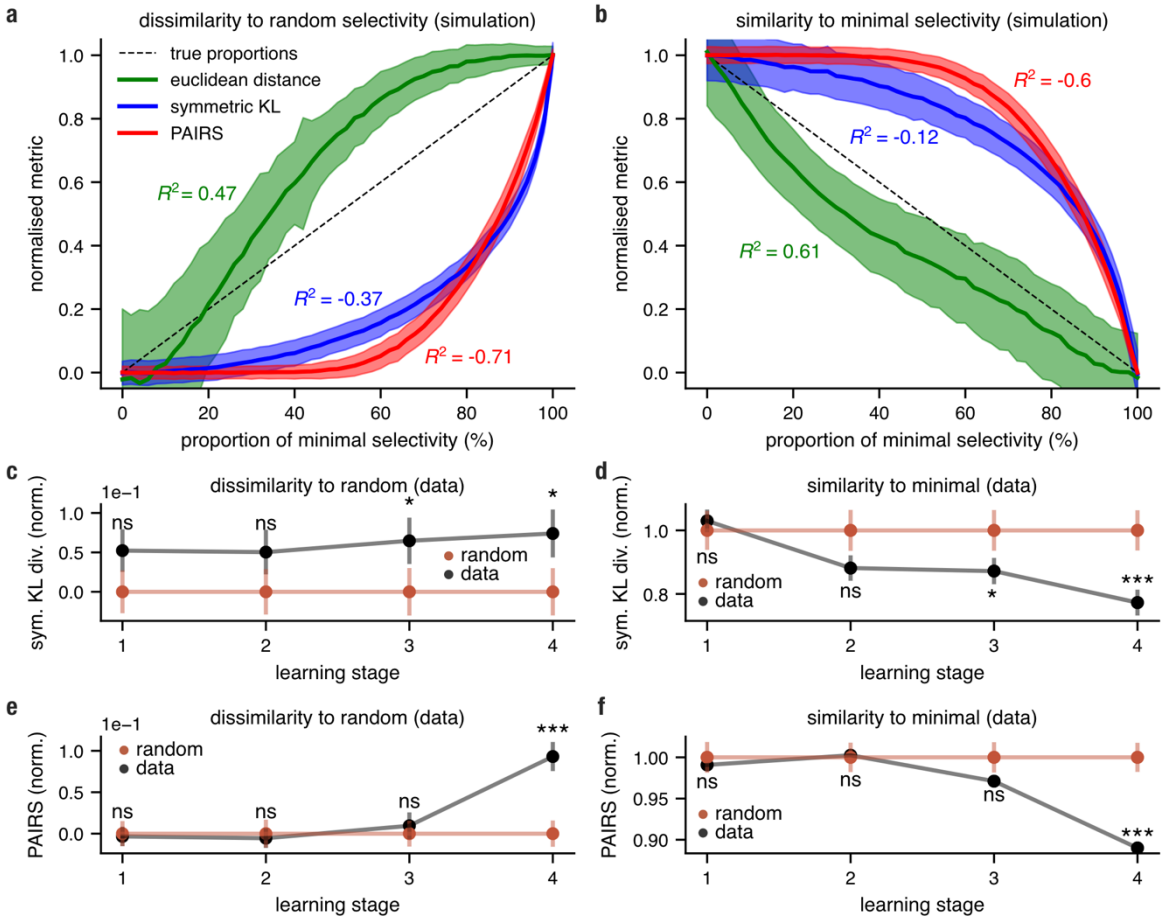


Supplementary figure 5. **a**, Temporally resolved linear SVM decoding for main task variables in experiment 2. Turquoise shading in the far left panel indicates the time window in which subsequent selectivity analyses were performed for context and stimulus set (cf. panels c-d); pale orange shading indicates when the remaining selectivity analyses were performed (cf. panels e-h). Horizontal dotted lines represent chance-level decoding whereas vertical dotted lines indicate the onset of the colour, shape and the trial outcome. The grey, black and dotted black bars indicate the statistical significance of stage 1 decoding, stage 4 decoding and significant differences in decoding between stage 1 and stage 4, respectively. **b**, Temporally resolved cross-generalised decoding of main task variables (trained on set 1 and tested on set 2 (and vice-versa) for context, shape and XOR); cross-generalised stimulus set decoding was obtained by training on set1 vs set2 in context 1 and testing on context 2 (and vice versa) **c**, Correlation between selectivity coefficients for stimulus set (set 1 vs set2 2) from context 1 and context 2 in late the colour-locked period (cf. turquoise in a) plotted as a function of learning; shaded areas indicate chance-level correlations between two random selectivity vectors obtained after shuffling trial labels. **d**, Correlation between

selectivity coefficients for context (context 1 vs context 2) in stimulus set 1 and stimulus set 2 in late the colour-locked period plotted as a function of learning. **e**, **f**, Analogous to c and f but computed for the late target-locked period (cf. pale orange in a). **g, h**, Correlations of selectivity coefficients obtained from stimulus set 1 and stimulus set 2 for shape and XOR, respectively. **i**, Comparison of learning dynamics of mean shattering dimensionality for set 1 in experiment 1 and set 1 in experiment 2. **j**, Comparison of learning dynamics of mean shattering dimensionality for set 1 in experiment 1 and set 2 in experiment 2. All p-values were calculated from permutation tests (***, $p < 0.01$; **, $p < 0.01$; *, $p < 0.05$; †, $p < 0.01$; n.s., not significant).



Supplementary figure 6. Electrode locations in monkey 1 (a) and monkey 2 (b) and their comparison (c).



Supplementary figure 7. The precision of Euclidean distance, symmetric KL estimate and PAIRS metrics in tracking learning-induced changes to neural selectivity. a, Relative Euclidean distance between the covariance matrix of selectivity coefficients obtained from a simulated mixed population (random-minimal) and the covariance matrix expected from pure random selectivity plotted as a function of minimal selectivity proportions (0-100%; for details see *methods*, *measuring similarity between selectivity distributions*). Coloured annotations indicate mean R^2 values computed between the true proportions (dotted lines) and estimated proportions (coloured bold lines). **b**, Same as panel **a** but we show the relative distance from the covariance matrix expected from minimal selectivity; shaded areas illustrate mean $\pm 1\, s.d.$ for each of the metrics computed from 1000 randomly drawn selectivity models. **c, d** Selectivity results from experiment 1(**Fig. 3d, e**) computed using symmetric KL divergence estimate. **e, f** Selectivity results from experiment 1(**Fig. 3d, e**) computed using the PAIRS metric.